Bioinformatics for Beginners

Genes, Genomes, Molecular Evolution, Databases and Analytical Tools

Supratim Choudhuri

BIOINFORMATICS FOR BEGINNERS

BIOINFORMATICS FOR BEGINNERS

Genes, Genomes, Molecular Evolution, Databases and Analytical Tools

Supratim Choudhuri

With contribution from Dr Michael Kotewicz on the Optical Mapping of DNA

Center for Food Safety and Applied Nutrition, FDA, College Park, Maryland

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

2014 Published by Elsevier Inc.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-410471-6

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

To my Family

Preface

As the title of the book suggests, this book is indeed for "beginners." It is not intended for advanced students of bioinformatics or practicing bioinformaticians. This book has been written from the perspective of an end-user who wants to use the freely available web-based databases and tools for bioinformatic analysis. The audience of this book could include any scientist or student who has a background in basic molecular biology but has not used web-based databases and tools for sequence analysis, or has not done bioinformatic analysis on a regular basis. The total number of chapters is only nine. This is because related sections have been combined into one chapter for coherence and understanding. These sections could have been easily split into separate stand-alone chapters to increase the number of chapters.

More than a decade into the first human genome sequencing, the use of bioinformatic analysis has been steadily increasing. There are more web-based freely available databases and analytical tools than ever before. Modern biology has pervaded even the social sciences. For example, sociologists and psychologists are now probing how the epigenomic effects of environmental factors (including social factors) might shape the personality and behavior of the offspring postnatally. The National Center for Biotechnology Information has established an epigenomics database, which will be immensely useful to scientists in the near future. Thus, bioinformatics has been slowly but steadily pervading all branches of biology and beyond. In keeping with this, more and more bioinformatics books are being written for experts, which do not necessarily cater to the needs of the non-experts.

Because this book is about bioinformatic analysis using web-based databases and tools, the emphasis is on sequence analysis. Global gene-expression profiling has not been emphasized other than a short discussion. The makers of gene-expression analysis platforms provide necessary software for analysis. Lastly, it is not possible to show every type of analysis in a book with a defined word count; nor is it possible to discuss all the links and all the functions associated with a database or analysis. Therefore, this book should serve as an initial guide, and it is expected that the reader will take it upon himself/herself to explore further using the databases and tools. Terms such as program, tool, algorithm, and web server have been used interchangeably throughout the book. These terms essentially mean the same thing in the context of this book. However, the term web server could be used to mean both the hardware and the software.

Because the principal audience of the book is supposed to be non-specialists, it was felt necessary to introduce the science and some core concepts of genomics as well as some important genomic techniques before embarking on the bioinformatic analysis. By the same token, some fundamental aspects of molecular evolution have been discussed in this book because the goal of many applications of bioinformatics is to trace the signatures of molecular evolution, as well as study the relatedness of taxa. In order to minimize the number of references in the text, reviews are cited wherever possible.

Supratim Choudhuri

Acknowledgment

The author would like to acknowledge the invaluable contributions of all scientists and engineers who developed databases and online tools for analysis, and made them freely available. The author would also like to acknowledge the contributions of the groups/institutions/organizations for hosting and maintaining these resources on web servers. A number of links for freely available databases and web-based tools for analysis have been provided throughout the book. Wherever possible, the latest relevant publications (which usually include the previous publications as well) describing these resources have been cited to acknowledge the contribution. The scientific community is truly grateful to the developers of these tools and databases and for making them freely available to facilitate bioinformatic analysis and learning.

The author would like to thank Dr Steve Gendel for his careful reading of the allergenicity prediction section in Chapter 8, and providing helpful suggestions.

The author would also like to thank many colleagues for their encouragement, enthusiasm, and support for the project.

Last but not the least, the author is grateful to Mr Graham Nisbet and Ms Catherine Mullane of Elsevier for making this project a reality, helping to bring it to successful completion, and being available whenever help and advice were needed.

CHAPTER

1

Fundamentals of Genes and Genomes*

OUTLINE

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

1.1 BIOLOGICAL MACROMOLECULES, GENOMICS, AND BIOINFORMATICS

Genetic information is stored in the cell in the form of biological macromolecules, such as nucleic acids and proteins. The genetic information not only drives the functioning of the whole organism, but also drives the evolutionary engine. Thus, an understanding of the molecular basis of life is fundamental to understanding how genetic information shapes life and drives its evolution. The following discussion captures some fundamental aspects of the structure and function of genes and genomes with special notes (in boxes) on the applications of this information.

1.2 DNA AS THE UNIVERSAL GENETIC MATERIAL

With some exceptions, deoxyribonucleic acid (DNA) is the universal genetic material. In some viruses, termed RNA viruses, RNA is the genetic material. The term **ribovirus** is used for viruses with single- and double-stranded RNA genomes, including retroviruses, which are RNA-based for a portion of their life cycle.[1]

Among the RNA viruses, **retroviruses** are well known; they include the notorious AIDS virus. Retroviruses are unique because in their life cycle they have both RNA and DNA versions of their genome. A complete retrovirus contains an RNA genome. The RNA genome encodes some protein products that are necessary for converting the single-stranded RNA genome into a double-stranded DNA genome and then its subsequent integration into the host genome. One such protein product of the retroviral genome is the reverse transcriptase (RT) enzyme. Upon entry into the cell, the reverse transcriptase is produced from the viral RNA genome using the host cellular machinery. The RT then copies the single-stranded RNA genome into a single-stranded DNA, which then produces a double-stranded viral DNA genome. The double-stranded viral DNA genome is referred to as the **provirus**, which gets incorporated into the host genome from where it keeps producing more retrovirus particles with single-stranded RNA genomes.

1.3 DNA DOUBLE HELIX

The structure of the DNA double helix and its building blocks are described in all biology textbooks. Here, some other aspects are also highlighted, including the information in Box 1.1. DNA is a **double-stranded right-handed** helix; the two strands are **complementary** because of complementary base pairing, and **antiparallel** because the two strands have opposite 5′−3′ orientation (Figure 1.1A). The diameter of the helical DNA molecule is 20 Å (=2 nm). The helical conformation of DNA creates the alternate **major groove** and **minor groove** (Figure 1.1B).

1.3.1 Structural Units of DNA

DNA is composed of structural units called **nucleotides** (deoxyribonucleotides). Each nucleotide is composed of a pentose sugar (2′-deoxy-D-ribose); one of the four nitrogenous bases—adenine (A), thymine (T), guanine (G), or cytosine (C); and a phosphate. The pentose sugar has five carbon atoms and they are numbered 1′ (1-prime) through 5′ (5-prime). The base is attached to the 1′ carbon atom of the sugar, and the phosphate is attached to the 5′ carbon atom (Figure 1.1A). The sugar and base form a **nucleoside**, whereas nucleoside plus phosphate makes a nucleotide. Hence, nucleoside = sugar + base, whereas nucleotide = sugar + base + phosphate. Table 1.1 shows the naming of nucleosides and nucleotides.

BOX 1.1

1. The major grooves in DNA can bind proteins. This is an important property of DNA structure because the major grooves in the upstream regulatory regions of a gene bind transcription-regulatory proteins. For example, for Zn-finger transcription factors, each Zn finger recognizes and binds to a specific trinucleotide sequence in the major groove of DNA.[2]
2. Any double-stranded nucleic acid (whether DNA double strand, DNA−RNA hybrid double strand, or RNA−RNA double strand) is antiparallel in nature. The complementary and antiparallel nature of double-stranded nucleic acids is an important property to remember while designing synthetic oligonucleotides for hybridization (probes or primers).
3. By convention, nucleic acid (DNA or RNA) sequence is written 5′→3′ from left to right, such as 5′-ATGTAAGCAC-3′. If the 5′→3′ designation is not mentioned, it is assumed that the sequence has been written in a 5′→3′ direction, following convention.

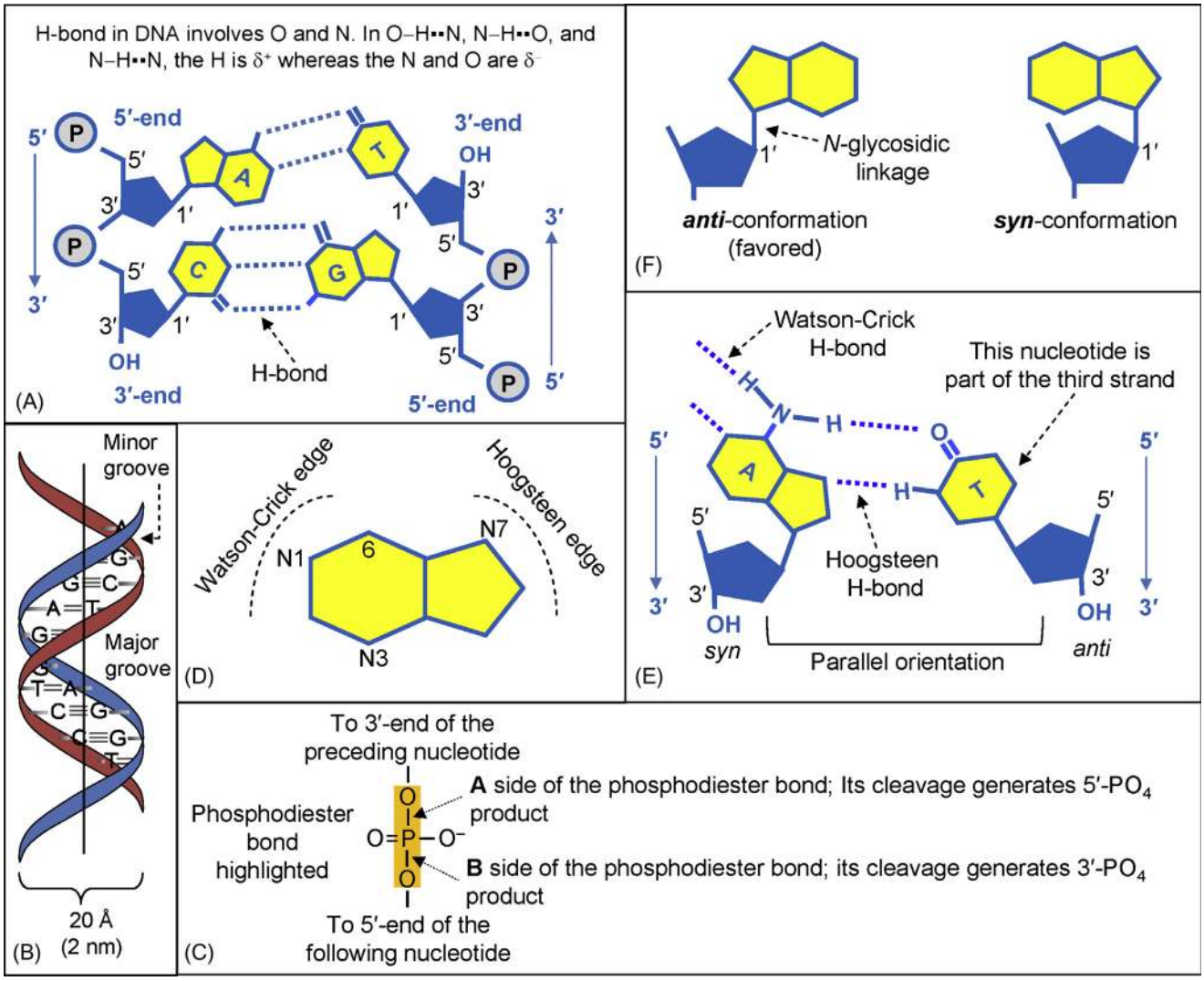


FIGURE 1.1 **DNA structure.** (A) Two nucleotides of the DNA double helix, showing their antiparallel orientation, two H-bonds between A and T and three H-bonds between G and C; (B) the DNA double helix showing the major and minor grooves as well as the diameter of the molecule; (C) the convention of classifying the two sides of the phosphodiester bond and the products generated from their cleavage; (D) the front side (Watson–Crick edge) and the back side (Hoogsteen edge) of a purine; (E) how Hoogsteen H-bonding aids in the formation of the triple helix (see Section 1.3.3); (F) the *anti* and the *syn* conformations of bases around the *N*-glycosidic bond.

TABLE 1.1 Naming of Nucleosides and Nucleotides

Base	Nucleoside (base + sugar)	Nucleotide (base + sugar + phosphate)
Adenine	Deoxyadenosine (sugar = deoxyribose)	Deoxyadenylic acid OR deoxyadenosine monophosphate
Guanine	Deoxyguanosine (sugar = deoxyribose)	Deoxyguanylic acid OR deoxyguanosine monophosphate
Cytosine	Deoxycytidine (sugar = deoxyribose)	Deoxycytidylic acid OR deoxycytidine monophosphate
Thymine	Deoxythymidine (sugar = deoxyribose)	Deoxythymidylic acid OR deoxythymidine monophosphate
Uracil (in RNA)	Uridine (in RNA) (sugar = ribose)	Uridylic acid OR uridine monophosphate

Each nucleotide in DNA (as well as in RNA) has one replaceable hydrogen, which is what makes the DNA (and RNA) acidic.

1.3.2 Linkage between Nucleotides

The nucleotides are joined by 5′–3′ phosphodiester linkage; that is, the 5′-phosphate of a nucleotide is linked to the 3′-OH of the preceding nucleotide by a phosphodiester linkage. In a linear DNA molecule, the 5′-end has a free phosphate and the 3′-end has a free OH group (Figure 1.1A). Each phosphodiester bond has two sides: a 3′-side that is linked to the 3′-end of the preceding nucleotide, and a 5′-side that is linked to 5′-end of the following nucleotide. The 3′-side is called

the **A side** by convention and its cleavage generates a 5′-PO_4 product. The 5′-side is called the **B side** by convention and its cleavage generates a 3′-PO_4 product (Figure 1.1C).

1.3.3 Base-Pairing Rules, Double Helix, and Triple Helix

In the double-stranded DNA, A pairs with T by two hydrogen bonds and G pairs with C by three hydrogen bonds (Figures 1.1A and 1.1B); thus GC-rich regions of DNA have more hydrogen bonds and consequently are more resistant to thermal denaturation. Each **nucleotide pair** (A—T and G—C) has a molecular weight of approximately 660 Da (sodium salt; 610 without sodium). In the helical double-stranded DNA molecule, the sugar—phosphate backbone lies outside and the bases are inside. Base pairs are stacked and horizontal; hence they are perpendicular to the axis of DNA. Because of the stacked nature of the base pairs in DNA, spatially flat molecules can intercalate between them. Of the four bases, A and G are **purines** whereas T and C are **pyrimidines**. In double-stranded DNA, a purine pairs with a pyrimidine (A with T and G with C). Therefore, total amount of purine should equal total amount of pyrimidine; in other words, the purine/pyrimidine ratio should be 1.0 or close to 1.0. This purine—pyrimidine equivalence in double-stranded DNA is called **Chargaff's rule**.

In the bases, the side with the N1 position of the heterocyclic ring is the "front," also called the **Watson—Crick edge** (Figure 1.1D); the opposite side is the "back," also called the **Hoogsteen edge**. Purines have an imidazole ring, which forms the "back"; so in purines, the N7 position of the imidazole ring is part of the Hoogsteen edge (Figure 1.1D). The Hoogsteen edge of the bases is located towards the edge (outside) of the DNA double helix, whereas the Watson—Crick edge is internal. In normal base pairing in DNA and RNA (Watson—Crick base pairing), the Watson—Crick edge (i.e. the front) of the two complementary bases is involved. However, the Hoogsteen edge provides an additional hydrogen bonding site. Therefore, the A—T and G—C base pairs in the normal double helix can form additional hydrogen bonds (**Hoogsteen hydrogen bonds**) to give rise to a triple helix involving the Hoogsteen edge of the purines, i.e. N7 of A and G for the third strand (Figure 1.1E). Hoogsteen hydrogen bonds can also form in RNA. In nucleic acids, the presence of a stretch of homopurine allows a stretch of homopyrimidine to hybridize through Hoogsteen hydrogen bonding to form a section of **DNA triple helix**. *The homopyrimidine-containing third strand is oriented parallel to the oligopurine strand* (Figure 1.1E), *whereas the homopurine-containing third strand is oriented antiparallel to the oligopurine strand* (see Box 1.2).[3—5]

For bases, two conformational variations are possible. The bond joining the 1′-carbon of the deoxyribose sugar to the base is the ***N*-glycosidic bond**. Rotation about this base-to-sugar glycosidic bond gives rise to ***syn*** and ***anti*** conformations. The *anti* conformation is the most common one (Figure 1.1F); however, the *syn* conformation can trigger the formation of triple helix (Figure 1.1E) and also play a role in transversion mutation (see Molecular basis of mutation, Section 2.3.1 in Chapter 2).

1.3.4 Single-Stranded DNA

Many DNA viruses have single-stranded DNA (for example, φX-174, parvoviruses). RNA viruses have RNA as the genetic material, and the RNA genome can be single or double stranded. Single-stranded DNA does not have base equivalence and hence does not follow Chargaff's base equivalence rule.

BOX 1.2

1. Each phosphate has three replaceable H^+; phosphodiester-bond formation between two nucleotides leaves one replaceable H^+. These replaceable H^+ make the DNA (and RNA) acidic (Figures 1.1 and 1.3).
2. The intercalation property of spatially flat molecules is utilized to visualize DNA (and RNA) in a gel using flat aromatic molecules that fluoresce under UV, such as ethidium bromide and acridine orange. The intercalation of these molecules can also cause frameshift mutation during DNA replication.
3. The purine—pyrimidine equivalence can be utilized to determine if a DNA molecule from an unknown source is double stranded or single stranded. In a double-stranded DNA molecule, the purine/pyrimidine ratio should be 1.0 (or close to 1.0); in contrast, in a single-stranded DNA molecule this equivalence is lacking.
4. The differential thermal stability of AT-rich versus GC-rich regions in double-stranded nucleic acids is taken into consideration while designing oligonucleotides for hybridization for different

BOX 1.2 (*cont'd*)

purposes, such as high-stringency hybridization, primers for polymerase chain reaction (PCR), or for sequencing. For example, an oligoprobe that will be used for high-stringency hybridization can have ≥ 55% G + C content.

5. If the molecular weight of an unknown double-stranded DNA is determined, the total base-pair content of the DNA can be calculated based on the fact that each **nucleotide pair** has an approximate molecular weight of 660 Da. By the same token, if the total number of base pairs in a DNA molecule is known, its molecular weight can be determined as well.
6. Hoogsteen hydrogen bonding can create short transient stretches of triple helix in vivo; triple helix formation can also be induced under experimental conditions. Synthetic oligodeoxynucleotides that can form triple helix have been used in vitro to inhibit gene expression in cells. Triple-helix-forming oligonucleotides coupled to DNA-modifying agents can be introduced into cells to modify the DNA target in a highly sequence-specific manner. This tool can be used to introduce genome modification, modulate specific gene expression, or even repair DNA.[6,7]

1.3.5 Base Sequence and the Genetic Code

The genetic information—that is, the genetic code with information for the amino acid sequence of the protein—lies in the sequence of bases in DNA. Genetic code exists in the form of a sequence of three bases; each three-base sequence is called a **codon**, which codes for an amino acid. Transcription of mRNA copies the codons from DNA to mRNA, which is translated to yield the protein (polypeptide) product. ATG in DNA (corresponding to AUG in RNA) is the start codon that codes for methionine. Translation begins by recognizing the start codon and incorporating methionine as the first amino acid. Similarly, TAG (**amber**), TGA (**opal**), and TAA (**ochre**) (corresponding to UAG, UGA, and UAA, respectively, in mRNA) are the three stop codons that do not code for any amino acids (exceptions to this rule are discussed below). In addition to being triplet (read as three-nucleotide codons), genetic code is (almost) **universal**, **non-overlapping** (adjacent codons do not share nucleotides), and **degenerate** (most amino acids can be coded by more than one codon). There are 64 (4^3) possible codons (61 coding and 3 noncoding). Genetic code normally codes for 20 standard amino acids. The two known cases of direct incorporation of non-standard amino acids are that of **selenocysteine** (the 21st amino acid) and **pyrrolysine** (22nd amino acid). Selenocysteine has been found in lower as well as higher organisms, including mammals, while pyrrolysine has so far been found in certain archaebacteria. Both these amino acids are encoded by stop codons; selenocysteine is encoded by UGA and pyrrolysine is encoded by UAG in mRNA.

1.4 CONFORMATIONS OF DNA

There are three major conformations of DNA: **B-DNA**, **A-DNA**, and **Z-DNA**. The DNA structure that Watson and Crick proposed was the B form of DNA (B-DNA), and this is the physiological form of DNA. In B-DNA, the diameter of the helix is 2 nm (=20 Å). Each pitch—that is, one complete turn (360°)—is 3.4 nm (=34 Å) long and contains 10 base pairs. A-DNA has been identified in vitro under different salt concentrations, as well as in DNA−RNA hybrids. It is also a right-handed helix. The diameter of the helix is 2.3 nm (=23 Å). Each pitch is 2.6 nm (=26 Å) and contains 11 base pairs. So, for a given length, the A-form is wider and shorter than the B-form. Z-DNA is a **left-handed helix** (Z = zigzag). This form has been identified both in vitro and within the cell. Small, localized regions within the physiological B-form of DNA can attain a left-handed conformation. Formation of the left-handed Z-DNA conformation is dictated by regions of alternating purines and pyrimidines residues, such as 5′-GCGCGCGCGCGCGCGC-3′. In Z-DNA, the diameter of the helix is 1.8 nm (=18 Å). Each pitch is 3.7 nm (=37 Å) long and contains 12 base pairs. Thus, the Z-form is narrower and longer than the B-form. It is thought that local Z-DNA conformations may play important roles in gene transcription.

1.5 TYPICAL EUKARYOTIC GENE STRUCTURE

According to the classical view of transcription, for any given gene, one of the two strands of DNA is

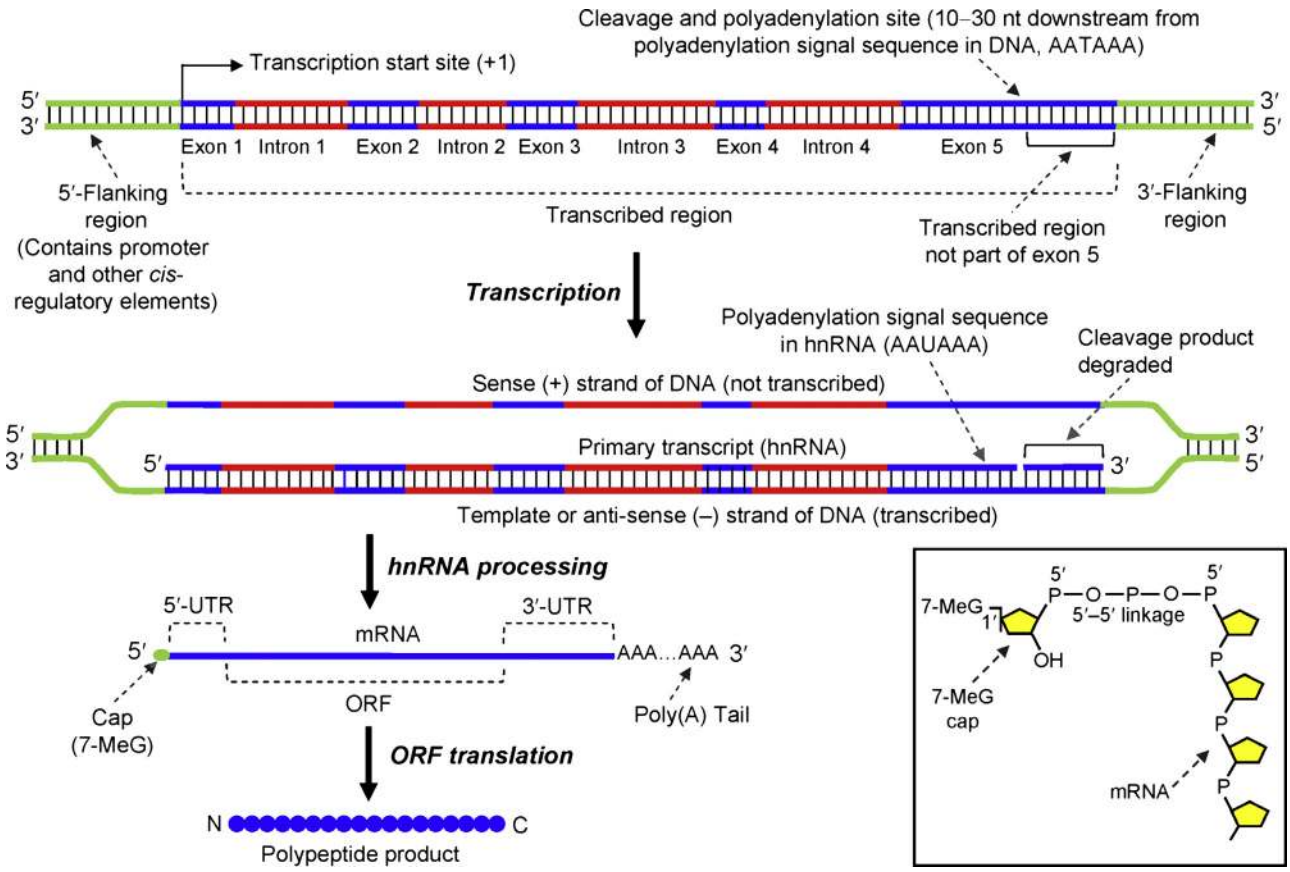


FIGURE 1.2 **Gene–hnRNA–mRNA–protein relationship.** Exon 1 is noncoding. Thus, the 5′-untranslated region (5′-UTR) is derived from exon 1, and the 3′-UTR is derived from the noncoding part of exon 5, which is the last and the longest exon. The sense strand of DNA has a "T" where the mRNA has "U"—for example, the poly(A) signal sequence in the sense strand is AATAAA, but in RNA it is AAUAAA. The transcription initiation site is +1 and the base to the left (upstream) of it is −1; there is no 0 position. Also, note that RNA polymerase transcribes well beyond the poly(A) site; this extra part of the transcript is degraded and does not form part of the last exon. Inset shows the mRNA cap (7-MeG) and its 5′–5′ linkage with the first base of mRNA. nt, nucleotide; ORF, open reading frame.

transcribed, the other is not[a]. The DNA strand that is NOT transcribed is called the **sense** or **plus** (+) or **coding** strand because it has the same sequence as that of the mRNA (except for U in RNA and T in DNA)—that is, the same sequence of codons in the same 5′→3′ direction, so that the polypeptide sequence can be predicted from the sense strand sequence (see Box 1.3). In contrast, the strand that is transcribed is called the **template** or **antisense** or **minus** (−) or **noncoding** strand because its sequence is complementary to the coding sequence; hence, the polypeptide sequence cannot be predicted from the template strand sequence. A typical mRNA-coding eukaryotic gene has three major parts: a transcribed region, a 5′-flanking region, and a 3′-flankng region (Figure 1.2). In eukaryotes, different types of RNAs are transcribed from the DNA by different RNA polymerases: RNA polymerase I (pol I) transcribes ribosomal RNA (rRNA), RNA polymerase II (pol II) transcribes messenger RNA (mRNA), RNA polymerase III (pol III) transcribes transfer RNA (tRNA). For mRNA, the primary transcript that contains both exons and introns is called the **heterogeneous nuclear RNA** (**hnRNA**) or **pre-mRNA**. The hnRNA is processed to remove the introns (**splicing**), add a 7-methyl guanine **cap** at the 5′-end by 5′–5′ linkage (Figure 1.2 inset), and add a **poly(A) tail** at the 3′-end, which is about 200 bp long in mammals.

[a]The classical view of transcription is an oversimplification. Deep sequencing and global transcriptome analysis have demonstrated that a significant proportion of the genome can produce both sense and antisense transcripts. When the sense and antisense transcripts are produced from the opposite strands of DNA in the same genomic locus, the antisense transcript is called a ***cis*-antisense** transcript because its target is the sense transcript. In contrast, ***trans*-antisense** transcripts are transcribed from a different location than their targets (e.g. microRNAs).

1.5.1 Transcribed Region

The nucleotide sequence of a gene that is transcribed into mRNA is composed of discrete sequences called **exons** and **introns**. Introns are also known as intervening sequences (abbreviated as **IS**) (Figure 1.2). After transcription of the gene, a longer primary transcript (the hnRNA or pre-mRNA) is produced. The hnRNA has the same exon−intron organization as the gene: exons are interrupted by introns. The hnRNA is processed to produce the mature mRNA. Exons are maintained in the mature mRNA, while introns are spliced out (in most cases). The structural unit of mRNA is the ribonucleotide (Figure 1.3). Introns do not contain information for the coding of the polypeptide. However, some introns, usually at the 5′-end of the gene, contain signals for transcriptional regulation. Introns of many genes also contain **nested genes** that have distinct expression profiles.[8] In mRNAs, a few terminal exons are noncoding, whereas the internal exons code for amino acids. These terminal noncoding exons form the 5′- and 3′-untranslated regions (UTRs) of the mRNA. In most mRNAs, the last exon (at the 3′-end) is usually the longest of all exons, and is partially coding (see Box 1.4).

1.5.1.1 Intron-Splicing Signals

Most introns in genes have GT at the 5′-splice site (in the DNA sense strand; hence GU in the hnRNA), called the **splice donor** site, and AG at the 3′-splice site, called the **splice acceptor** site. These introns are referred to as GT−AG introns. However, introns may also contain GC or AT as the splice donor sites, and AC as the splice acceptor site (hence, GC−AG introns, AT−AC introns).

In most eukaryotic genes, the nucleotides surrounding the splice donor and acceptor sites show a great degree of conservation. The usual nucleotide distribution around the splice sites is as follows:

5′-splice site: 5′-. . .NNNAG**gt**annn. . .3′ (**gt** = splice donor site in the intron; N = any nucleotide in the exon; n = any nucleotide in the intron; bases underlined are usually conserved; AG are the last two bases of the preceding exon, and a is the base that immediately follows the splice donor site).

3′-splice site: 5′-. . .nnnc**ag**NNN. . .3′ (**ag** = splice acceptor site in the intron; N = any nucleotide in the following exon; n = any nucleotide in the intron; the base underlined is usually conserved; c is the base immediately preceding the splice acceptor site).

Two other important sequence elements are the **branch point** and the **polypyrimidine tract** in the introns. The branch point is located 20−50 nucleotides upstream from the splice acceptor site. The consensus sequence of the branch point site is (C/T)(T/C)(A/G)**A**(C/T), in which the **A**-residue is conserved in all genes. This **A**-residue is called the branch point and it plays a crucial role in splicing. The polypyrimidine tract is located downstream from the branch point.

BOX 1.3

1. An easy way to remember the sense and antisense designations is to remember just one fact: that the sequence of mRNA is sense. This is because the codons can be found in the coding sequence of mRNA; as a result the amino acid sequence of the polypeptide can be predicted from the mRNA coding sequence. Hence, any sequence that is same as the mRNA sequence along with the same 5′→3′ polarity is also sense. That is why the DNA strand that has the same sequence and polarity as the mRNA is also sense. Likewise, any sequence that is complementary to the mRNA sequence, along with the opposite 5′→3′ polarity, is antisense. Hence, the template DNA strand is antisense (Figure 1.4A).
2. By the same token, the probe used to detect mRNA in northern blot or in situ hybridization is antisense because it is complementary and has an opposite polarity to the mRNA. When designing antisense DNA oligoprobes for RNA or DNA hybridization, the complementary and antiparallel sequence of the sense strand of DNA is used. For example, in Figure 1.4, the mRNA partial sequence shown is 5′-AUG UGU AGA UCG AUG A-3′. That region of the antisense DNA probe will have the sequence 3′-TAC ACA TCT AGC TAC T-5′. Following convention, the DNA probe sequence has to be rewritten in a 5′→3′ direction from left to right. Hence, this DNA probe partial sequence will be rewritten (for reporting the sequence) as 5′-TCA TCG ATC TAC ACA T-3′ (Figure 1.4B).
3. In the nucleotide databases, such as in National Center for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), or The European Molecular Biology Laboratory (EMBL), the reported mRNA sequences do not contain U but instead contain T. This is because the mRNA sequence is reported as the sense strand of the cloned complementary DNA (cDNA).

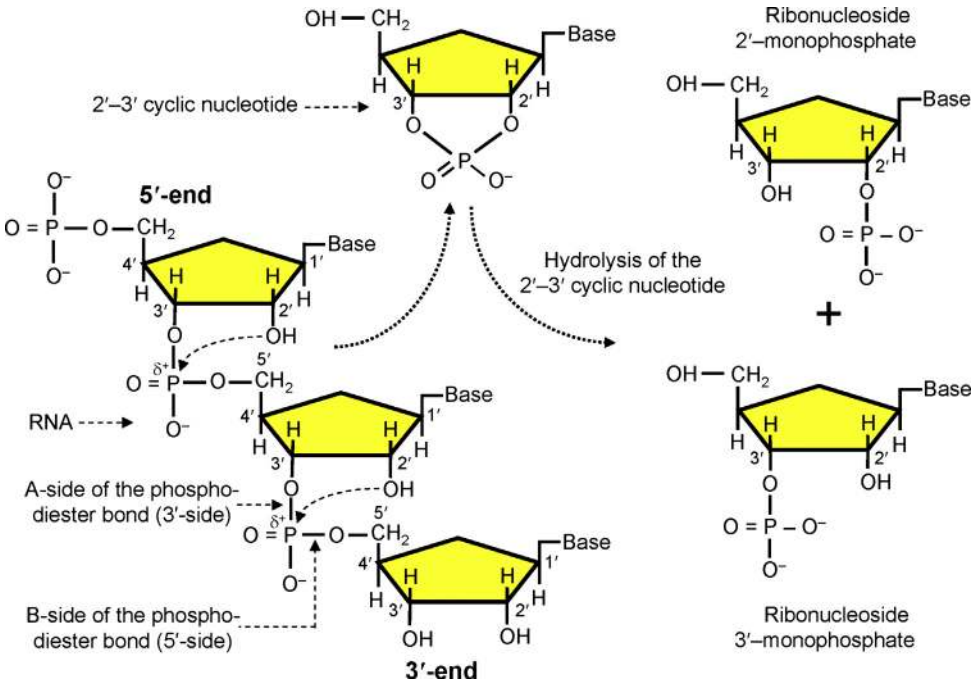


FIGURE 1.3 **Alkaline hydrolysis of RNA.** In an alkaline pH, the OH^- can abstract the H from the 2′-OH of ribose, generating the nucleophile 2′$-O^-$, which carries out a nucleophilic attack on the δ^+ P of the phosphate. This results in the cleavage of the phosphodiester bond and the formation of 2′−3′ cyclic nucleotide; the cyclic nucleotide hydrolyzes into ribonucleoside 2′- and 3′-monophosphate end products.

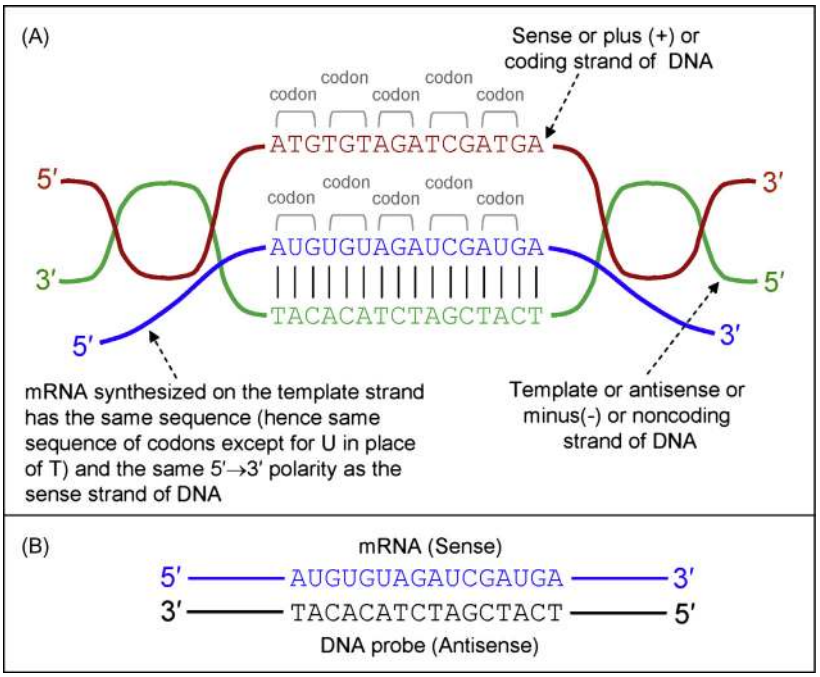


FIGURE 1.4 **Sense and antisense strands of DNA.** (A) The two strands of DNA have been drawn in different colors so that their respective 5′- and 3′-ends could be easily distinguished. The figure shows that mRNA and the sense strand have the same sequence (except for "U" in RNA and "T" in DNA) and the same 5′−3′ polarity. (B) The mRNA and antisense probe relationship.

BOX 1.4

1. Sometimes an intron may be retained in the mature mRNA and perform specific regulatory functions. For example, migration stimulatory factor (MSF) is a truncated oncofetal isoform of fibronectin. Two types of MSF mRNAs have been detected: a shorter 2.1-kb[b] transcript and a longer 5.9-kb transcript, which differ only in the length of their 3′-UTRs. In the smaller transcript, the intron-derived 30-nucleotide (nt) coding sequence is followed by a 165-nt intron-derived 3′-UTR. This makes a total of 195-nt intron-derived sequence in the smaller transcript.[9] This intron-derived 3′-UTR also provides the polyadenylation signal. The smaller transcript is transported to the cytoplasm and eventually secreted, while the larger transcript is retained in the nucleus.
2. After a gene is cloned and sequenced, the exon–intron boundaries are identified by comparing the gene sequence with its complementary DNA (cDNA) (mRNA) sequence. Identification of the exon–intron boundaries of a gene is essential when attempting to manipulate the DNA, such as making a gene-targeting construct.
3. The majority of internal exons in vertebrate genes are less than 300 bp; the average length being 135 bp; exons larger than 800 bp are rare.[10]
4. For most genes, the last exon (at the 3′-end) is the longest exon (could be well over 1 kb) and partially coding.
5. For most genes, the 5′-UTR is derived from more than one exon. Of these 5′ noncoding exons, the most downstream one is usually partially noncoding because the open reading frame (ORF) begins at some place in this exon, making it partially noncoding and partially coding.
6. For most genes, the 3′-UTR is three to five times longer than the 5′-UTR, particularly in vertebrates.
7. In vertebrates, exons are small and introns are large. In contrast, in lower eukaryotes, the opposite is true.[11]
8. The transcription start site (+1) in most genes begins with a purine (mostly an "A").

[b]kb, kilobase = 1000 bases; Mb, megabase = 1000 kb; Gb, gigabase = 1000 Mb. In the context of DNA, these mean base pairs (hence, kbp, Mbp, and Gbp).

BOX 1.5

Knowledge of the intron phases helps predict which exon(s) can or cannot be targeted for alternative splicing. Exceptions to this rule have also been reported in the literature. For example, the alternative splicing of *rat liver-specific organic anion transporter* pre-mRNA, generating a functional mRNA, involves the removal of exon 10, which is an asymmetrical exon flanked by a phase 1 and a phase 2 intron. The creation of a frameshift mutation in this unusually spliced mRNA is averted by retaining 91 bp from the 5′-end of exon 10 in the mature mRNA.[12]

1.5.1.2 Effect of Intron Phase on Alternative Splicing

Introns can be divided into three types based on phases: **phase 0**, **phase 1**, and **phase 2**. A phase 0 intron does not disrupt a codon, a phase 1 intron disrupts a codon between the first and second bases, whereas a phase 2 intron disrupts a codon between the second and third bases. An exon flanked by two introns of the same phase is called a **symmetrical exon**, whereas an exon flanked by two introns of different phases is called an **asymmetrical exon**. Intron phase determines which exons may or may not be targeted for alternative splicing. With a few rare exceptions, exons that are subjected to alternative splicing are always symmetrical exons—that is, exons flanked by same-phase introns. In contrast, asymmetrical exons—that is, exons flanked by different-phase introns—cannot be alternatively spliced because such alternative splicing will throw the normal open reading frame (ORF) out of frame beyond the 3′-splice site (Figure 1.5). Such frameshift results in the creation of premature stop codon and truncation of the ORF. Intron phase determines exon shuffling potential, which determines protein domain shuffling during protein evolution and the evolution of organismal complexity (discussed in Chapter 2; see Box 1.5).

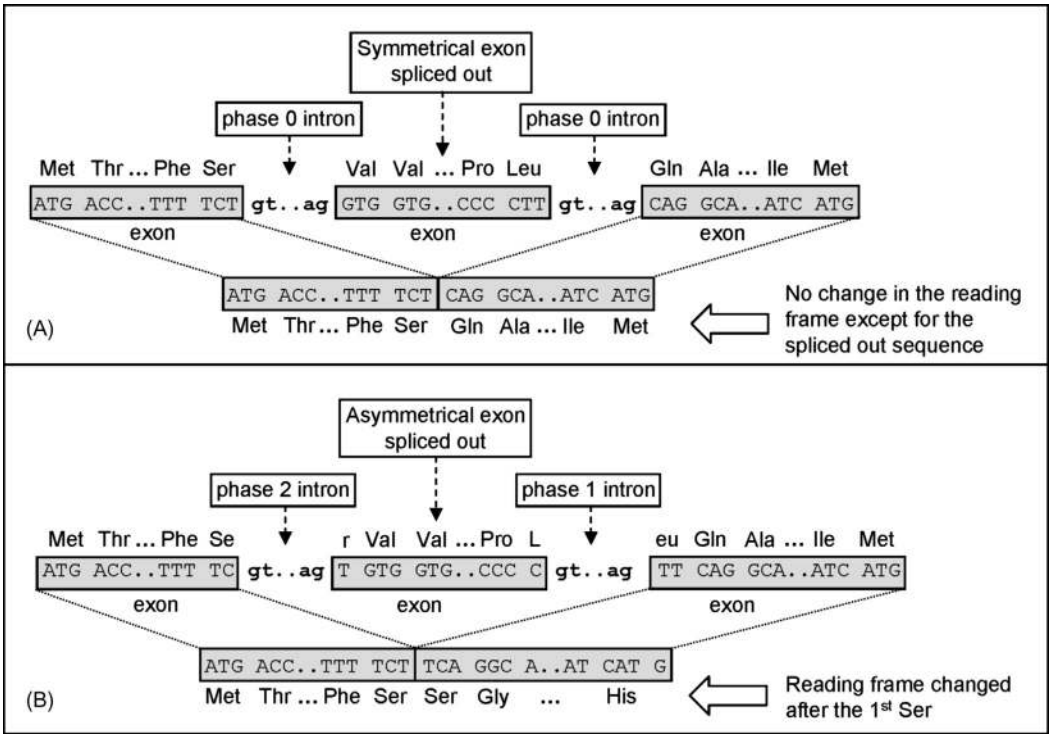


FIGURE 1.5 **The effect of intron phase on alternative splicing.** (A) Alternative splicing involving the removal of a symmetrical exon (flanked by introns of the same phase; 0–0) does not cause a frameshift in the ORF except for the deletion of the amino acids encoded by the removed exon; (B) alternative splicing involving the removal of an asymmetrical exon (flanked by introns of different phase; 2–1) causes a frameshift in the ORF downstream from the 3′-splice site. Such frameshift results in the creation of a premature stop codon and truncation of the ORF.

1.5.1.3 Evolution of Introns

After the initial discovery of introns in 1977, the **introns-early theory** was proposed to explain the origin and evolution of introns. According to the introns-early theory, introns were present as intergenic regions in the genome of the common ancestor of prokaryotes and eukaryotes. These intergenic genomic regions were subsequently lost in all prokaryote lineages; in contrast they were maintained in eukaryotes as introns owing to the appearance of the spliceosomal machinery. Walter Gilbert suggested that the presence of introns allowed exon shuffling, which resulted in genomes being more complex and diversified. The accumulation of genomic data has helped reconstruct the evolutionary history of introns and replace the introns early theory with the **introns-late theory**. According to the introns-late theory, self-splicing introns (also known as **retrointrons**) first invaded eukaryotic genomes, and spliceosomal introns were subsequently derived from self-splicing introns. Hence, spliceosomal introns only appeared in eukaryotes. Spliceosomal machinery evolved as a means of removing spliceosomal introns. Therefore, the last common ancestor of eukaryotes had a spliceosomal-intron-rich genome. The intron-containing genomes probably spread due to **population bottlenecks**[c]. Further massive intron invasion of the genome was likely limited only to those genomes that underwent significant evolutionary innovations. Intron loss in many lineages also occurred, resulting in the present-day intron-poor species.[13,14]

Introns-late theory envisages that early introns had no functions; hence their presence was deleterious for the genomes. However, early introns were transcribed and were free from selective constraints; hence, at some point during evolution, they might have gained

[c]Population bottleneck is a phenomenon in which the population size is drastically reduced through events like environmental disaster, habitat destruction, or massive predation and hunting. As a result, only a small fraction of the genetic diversity of the original population survives. When the population multiplies, the surviving genetic diversity spreads in the population. Thus, if the intron-containing genome survived through a population bottleneck, it subsequently spread in the resulting population. In general, population bottleneck results in a drastic reduction of the gene pool and genetic diversity in the resulting population. Owing to the loss of genetic variation, the new population could be genetically distinct from the original population. Loss of genetic diversity, particularly in a small population, can cause genetic drift and rare alleles face increased chance of being lost.

some functions. One of the best known functions of introns is their ability to increase transcription and ultimately protein expression of intron-bearing genes compared to intronless genes. In making transgenic organisms, particularly transgenic plants, specific introns are frequently included in the construct to increase the expression of the transgene.

Introns are now known to mediate their function by modulating every possible step of transcription: initiation, elongation, termination, mRNA maturation, nuclear export, and mRNA stabilization. The mechanism of action of many introns is not known. However, the functions of introns can be sequence-dependent, length-dependent, position-dependent, and splicing-dependent.[15]

1.5.2 5′-Flanking Region of Transcribed Genes

The 5′-flanking region of transcribed genes contains the **promoter**. The promoter contains specific sequences for binding the proteins necessary for transcription by RNA polymerase. The specific sequence in the promoter that positions the pol II is called the **TATA box** (consensus 5′-TATAAA-3′; some variants exist). Typically, the TATA box is located 25−30 bp upstream of the transcription start site (that is, −25 to −30 bp position), and for any given gene the position of the TATA box is fixed. However, many gene promoters lack the TATA box (TATA-less promoters). Accurate positioning of pol II in TATA-less promoters is thought to be mediated by two other *cis*-acting sequence elements, the **initiator** element (**Inr**) and the **downstream promoter element** (**DPE**). Inr has a consensus sequence of Y-+1-N-T/A-Y-Y (where Y is a pyrimidine, +1 is the transcription initiation site, N is any nucleotide), and DPE has a consensus sequence of $(A/G)_{+28}G(A/T)(C/T)(G/A/C)_{+32}$. Therefore, Inr occurs around the transcription start site and DPE occurs between 28 and 32 bases downstream from the transcription start site. Many variants of the Inr sequence have been reported. DPE has been most extensively studied in *Drosophila*. Some other sequences in the promoter that are found in most genes are the CAAT-box (around −75 to −80 bp position) and the GC-box (around −90 bp position).

Various regions of the promoter have been termed the core (or basal), proximal, and distal promotor depending on their distance from the transcription start site. The **core promoter** is about 35 bp long and extends 35 bp upstream or downstream from the transcription site (−35 to +35), the **proximal promoter** is around 250 bp long, whereas the **distal promoter** is located further upstream. Therefore, the TATA box, Inr, and DPE are all contained within the core promoter, whereas the CAAT-box and the GC-box are contained within the proximal promoter. Core, proximal, and distal promoter elements cooperate to regulate transcription.

The proximal promoter contains additional *cis*-acting sequences that are necessary for the regulation of gene expression in response to specific stimuli. These sequences are called **response elements** or **regulatory elements** (**RE**). For example, genes that are induced by glucocorticoids have a glucocorticoid response element (GRE) in their promoters. Many such response elements have been identified so far in a number of animal and plant gene promoters. These response elements bind specific transcription regulatory proteins called transcription factors that control gene expression. Regulatory elements can also be found far upstream of the TATA box, far downstream in the 3′-flanking sequence, and even within introns. These elements typically act as enhancers because they significantly upregulate the expression of genes (see Box 1.6).

1.5.3 3′-Flanking Region of Transcribed Genes

Although it is often said that the 3′-flanking region contains the transcription termination signal, eukaryotic pol II does not terminate transcription at any definitive

BOX 1.6

Promoter-bashing experiments help identify the importance of specific promoter sequences in regulating gene expression. These experiments make use of deletion mutations to narrow down the region of interest; then individual bases are mutated to define the core functional sequence involved in regulating transcription. Bioinformatic software uses the available information on various identified transcriptional activator- or repressor-binding sequences, and scans the 5′-flanking sequences of a gene to predict putative binding sites in the promoter. However, many of the putative binding sites predicted through bioinformatic analysis may turn out to have no effect on transcription when verified through promoter-bashing experiments. Thus, predicted regulatory sequences are only a rough guide and need functional verification through experimentation.

termination signals in the DNA. For most eukaryotic protein-coding genes, pol II transcribes the template strand 500−2000 nucleotides beyond the polyadenylation site (Figure 1.2). Transcription termination is facilitated by a number of protein factors (such as Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Stimulation Factor (CStF), etc.) that become associated with the pol II as soon as the enzyme leaves the promoter. These factors, along with capping and splicing factors, ride on the C-terminal domain (CTD) tail of pol II. Transcription of the poly(A) signal sequence triggers the endonucleolytic cleavage of the nascent transcript, degradation of the downstream cleavage product, and termination of transcription. The pausing of pol II downstream from the poly(A) site appears to be an obligatory step leading to termination, which involves the displacement of pol II from the template. *The 5′- and 3′-ends of a gene are the same as the 5′- and 3′-ends of the sense strand*.

1.6 MUTATIONS IN THE DNA SEQUENCE

The sequence that codes for a polypeptide is referred to as the coding region or **open reading frame** (**ORF**). Various mutations in the ORF may or may not lead to changes in the amino acid sequence in the polypeptide product. If a mutation in DNA leads to an amino acid change in the polypeptide, it is called a **missense** or **nonsynonymous** mutation; if a mutation does not lead to an amino acid change in the polypeptide, it is called a **silent** or **synonymous** mutation. Traditional wisdom assumes that a synonymous mutation does not alter the protein function because there is no change in the amino acid. However, recent findings indicate that, in many proteins, synonymous mutations may also alter protein function because they result in an altered conformation of the protein. Because protein folding is a co-translational process, proper protein folding is tightly linked to the speed of translation. Synonymous mutations that affect codon usage may disrupt this process resulting in a wrongly folded polypeptide. In fact, some human diseases could be linked to such synonymous mutations.[16]

1.7 SOME FEATURES OF RNA

In traditional molecular biology, a discussion on RNA focused on three types of RNA associated with protein synthesis: ribosomal RNA (rRNA), messenger RNA (mRNA), and transfer RNA (tRNA), of which rRNA and tRNA are noncoding, whereas mRNA is protein coding. The world of functional noncoding RNA molecules has since been greatly expanded (discussed later). As mentioned above, RNA is the genetic material in retroviruses. An RNA molecule is single stranded, except in regions where base complementarity makes the molecule fold back on itself forming double-stranded segments. Like DNA, RNA is also composed of nucleotides (ribonucleotides). However, there are two differences from DNA: the sugar is ribose and the base uracil ("U") is present instead of "T"; thus the base pairing is between "A" and "U." Of the three RNAs associated with translation (rRNA, mRNA, and tRNA), the following discussion focuses on mRNA.

1.7.1 Instability of mRNA

Apart from the ubiquitous presence of the enzyme RNAse that can easily degrade mRNA, the structure of mRNA itself also contributes to its instability. The ribose sugar makes RNA less stable than DNA, especially at alkaline pH. At alkaline pH, the 2′-OH of the ribose sugar undergoes alkaline hydrolysis, which results in the breakage of the phosphate bond between adjacent nucleotides, and formation of the 2′−3′ cyclic nucleotide (Figure 1.3). Hydrolysis of this 2′−3′ cyclic nucleotide gives rise to a mixture of ribonucleoside 2′- and 3′-monophosphate products. In contrast, in DNA the 2′ carbon has an H instead of an OH, which prevents the formation of the 2′−3′ cyclic nucleotide; this prevents alkaline hydrolysis and makes DNA stable at alkaline pH. At acidic pH, however, phosphodiester bond hydrolysis occurs in both DNA and RNA. Because RNA undergoes rapid alkaline hydrolysis, particularly around 37°C, use of NaOH (even ice-cold) to denature RNA is not recommended.

1.7.2 5′- and 3′-Untranslated Regions of mRNA

A typical eukaryotic mRNA has three regions: a 5′-untranslated region (5′-UTR), a coding region or ORF, and a 3′-untranslated region (3′-UTR). The translational start codon is AUG, and there is one of the three translational stop codons, UAA, UGA, and UAG. The 5′-end of mRNA has the cap (7-methyl GTP) attached to the first base through a **5′−5′ linkage**. The 5′- and 3′-UTRs are composed of noncoding exons or noncoding parts of partially coding exons, whereas the ORF is composed of coding exons. The last exon at the 3′-end is usually the longest. The 3′-UTR of mRNAs contains the **poly(A) signal** sequence 5′-AAUAAA-3′, which is located 10−30 nucleotides upstream of the polyadenylation site (see Box 1.7). The poly(A) tail is around 200 bp long in mammals. The cap at the 5′-end and the poly

(A) tail at the 3′-end help in translation and also aid in the stability of the mRNA. If the 3′-UTR of an mRNA contains multiple poly(A) signal sequence, the mRNA may undergo **alternative polyadenylation**, producing transcripts with very different stability. Alternatively polyadenylated mRNAs also differ in the length of their 3′-UTRs; they can be observed in different tissues or at different developmental stages where the half-life of the same mRNA may markedly vary.[17] Many mRNAs with more than one poly(A) signal sequence have been reported in the database, but not all of them have been experimentally tested to confirm the generation of alternatively polyadenylated transcripts.

The 5′-UTR of mRNA controls the initiation of translation. An important sequence relevant for translation initiation and identification of the correct AUG codon (translation start codon) is called the **Kozak sequence**, after its discoverer, Marilyn Kozak. The original Kozak sequence described was 5′-CCRCC**AUG**G-3′ where **AUG** is the translation start codon, and R is a purine. Later on, a shorter yet highly effective version of the Kozak sequence was described as 5′-ACC**AUG**G-3′. Although many mRNAs contain the consensus Kozak sequence or some variant of it, there are many other mRNAs that do not contain any Kozak sequence at all.

The 5′-and 3′-UTRs of mRNAs can also regulate gene expression and mRNA stability by interacting with proteins or nonprotein ligands. For example, the expression of feritin mRNA is regulated by the binding of specific regulatory proteins to its 5′-UTR, whereas the stability of transferrin receptor mRNA is regulated by the binding of specific regulatory proteins to its 3′-UTR. In contrast to protein ligands, in bacteria certain mRNAs can regulate gene expression by binding specific nonprotein ligands. The part of the mRNA that binds to the small molecule and acts as the genetic switch is called a **riboswitch**. Some examples include coenzyme-B12-binding riboswitch, flavin mononucleotide (FMN)-binding riboswitch, thiamine or thiamine pyrophosphate (TPP)-binding riboswitch—all located in the 5′-UTR of the relevant mRNAs.[18]

1.7.3 Secondary Structures in RNA

RNA crystallography has revealed the existence of a rich variety of base pairing, giving rise to a multitude of complex tertiary structural motifs. Leontis and Westhof[19] proposed that the planar edge-to-edge hydrogen-bonding interactions between RNA bases involve one of three distinct edges: the **Watson−Crick edge**, the **Hoogsteen edge**, and the **sugar edge** (which includes the 2′-OH). About 60% of the bases participate in canonical Watson−Crick base pairs. The original geometric nomenclature and classification has been recently revisited by Abu Almakarem et al.,[20] who developed a classification scheme that is predicted to help identify recurrent base triplets (referred to as "base triples" in the publication) that can substitute for each other while conserving RNA three-dimensional structure. Hence, the system has applications in RNA three-dimensional structure prediction and analysis of RNA sequence evolution. Taking into consideration the spatial orientations in which bases can interact, Leontis and Westhof identified 12 basic geometric types with at least two H-bonds connecting the bases. In other words, Leontis and Westhof defined 12 base-pair families. Using the combinatorial enumeration of these 12 base-pair families, Abu Almakarem and coworkers predicted the existence of 108 potential geometric base-triple (triplet) families. Searching representative atomic-resolution RNA three-dimensional structures revealed instances of 68 of the 108 predicted base-triple families. Further model building suggested that some of the remaining 40 families may be unlikely to form for steric reasons.

BOX 1.7

1. Bioinformatic analysis of any sequence that might code for a polypeptide will produce a total of six reading frames: three in sense, three in antisense. Of these, one reading frame is always the longest, providing the legitimate ORF. Some software produces only three sense-frame output.
2. The polyadenylation (poly(A)) signal sequence is highly conserved. The canonical poly(A) signal sequence identified in cloned complementary DNA (cDNA)/gene sequence is AATAAA (AAUAAA in the mRNA). The only other known functional variant of the poly(A) signal sequence is ATTAAA (AUUAAA in the mRNA).

1.8 CODING VERSUS NONCODING RNA

In addition to rRNA and tRNA, a few other classes of ncRNAs have been known for some time, such as snRNA (small nuclear RNA), snoRNA (small nucleolar RNA), gRNA (guide RNA), *Xist* (X inactive-specific transcript) and *Tsix* (an antisense regulator of *Xist*), *H19*, *Air*, and *Kcnq1ot1* (potassium channel Q1 overlapping transcript 1). These ncRNAs are very different in length (e.g. 50–70 nucleotides (nt), such as gRNA, to more than 100 kb, such as *Air* ncRNA), and they serve diverse functions. For example, snRNAs are essential for mRNA splicing, snoRNAs are important in methylation of rRNAs, gRNAs are essential in RNA editing, whereas *Xist*, *Tsix*, *H19*, *Air*, and *Kcnq1ot1* are all involved in the epigenetic regulation of gene and genome expression; for example, *Xist* and *Tsix* are involved in X-chromosome inactivation in mammals whereas *H19*, *Air*, and *Kcnq1ot1* are associated with imprinted loci and genomic imprinting. Since the 1990s, the RNA universe has been producing regular surprises that have enriched our idea about RNA's role in gene regulation, and the breadth of the cellular gene regulatory network itself.

1.8.1 Small Noncoding RNA, Long Noncoding RNA, Competing Endogenous RNA, and Circular RNA

In recent years, a new class of ncRNAs, the **small ncRNAs** (~20–30nt long), has been identified as very powerful regulators of gene expression. Examples include **microRNA** (**miRNA**, abbreviated as **miR**), **small interfering RNA** (**siRNA**), and **Piwi-interacting RNA** (**piRNA**).[21,22]

These small ncRNAs are generated through the processing of double-stranded segments of long precursor RNAs. Accordingly, software has been developed to identify putative genomic sequences that may give rise to small ncRNAs, as well as potential target sequences of these putative ncRNAs. These theoretical predictions have to be experimentally confirmed. An ever-increasing number of studies have implicated miRNAs and siRNAs in human health and disease, ranging from metabolic disorders to diseases of various organ systems, including various forms of cancer. More than 30% of all human genes have been predicted to be miRNA targets. Consequently, a number of freely accessible web-based miRNA databases have been developed that contain both predicted and experimentally verified miRNA sequences. One such database is the miRBase (http://microrna.sanger.ac.uk/), which is one of the most comprehensive miRNA databases. Release 19.0 (August 2012) of the miRBase reports a total of 21,264 identified miRNAs in different species, of which 2214 are identified in humans. Examples of some other miRNA databases are:

miRNAviewer (http://cbio.mskcc.org/mirnaviewer/)
miRWalk (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/)
MicroRNA.org (http://www.microrna.org/microrna/home.do)
miRGator (http://genome.ewha.ac.kr/miRGator/).

Long noncoding RNAs (**lncRNAs**) are >200 nucleotides in length and do not code for protein. The lncRNAs are the least understood among the ncRNAs, but evidence suggests that they play important roles in a broad range of biological processes.[23] The *Air*, *Xist*, *Tsix*, and *Kcnq1ot1* RNAs discussed above are all lncRNAs. A good lncRNA database can be accessed at http://www.lncrnadb.org/.[24]

Just as an efficient regulatory network should have multiple control points, the regulation of gene expression by miRNAs is further regulated by other RNAs. Two such recently discovered miRNA-regulatory RNAs are competing endogenous RNA (ceRNA) and the most recently reported circular RNA (circRNA). Functionally, both these RNAs antagonize the effects of miRNA. The discovery of these anti-miR RNA molecules will trigger a reevaluation of the model of the RNA regulatory network, and the gene regulatory potential of miRNAs.

As the name implies, **competing endogenous RNAs** (**ceRNAs**) are noncoding RNA molecules that contain binding sites for miRNAs, referred to as miRNA response elements (MREs), and thus compete with the miRNA targets to bind the miRNAs. In sequestering the miRNAs, the ceRNAs allow the miRNA target RNAs to be expressed. According to this definition of ceRNA, the RNA products of expressed pseudogenes containing miRNA binding sites will qualify as ceRNAs. Likewise, lncRNA can act as ceRNA as well. For example, *linc-MD1* is a validated cytoplasmic lncRNA expressed during myoblast differentiation; it acts as a ceRNA for miR-133 and miR-135 targets. Phosphatase and tensin homolog (*PTEN*) is a tumor suppressor gene whose expression is frequently altered in many human cancers. The regulation of *PTEN* expression by a whole plethora of miRNAs is further modulated by ceRNAs, such as VAPA and CNOT6L.[25]

The **circular RNAs** (**circRNAs**) with a functional role are the latest addition to the RNA universe. The existence of RNAs in circular form at a low level had been reported earlier; these were treated as unique, sporadic observations. The extensiveness of circRNA expression was reported in 2012.[26] The authors concluded that a non-canonical mode of RNA splicing, resulting in a circular RNA isoform, is a general

feature of the gene-expression program in human cells, and that the expression of circRNAs is more prevalent and widespread than once thought. However, the regulatory role of circular RNAs was highlighted by two recent publications.[27,28] Both these publications described highly stable circular RNAs in human and mouse brain (termed CDR1as, for antisense (as) to the cerebellar-degeneration-related protein 1 transcript CDR1, by Memczak et al., and ciRS-7 for circular RNA sponge for miR-7, by Hansen et al.). These circRNAs bind many copies of miR-7 and terminate miR-7-mediated suppression of target mRNAs. These circular RNAs contain approximately 70 conserved binding sequences for miR-7. Overexpression of this circRNA reversed the miR-7-mediated suppression of the target mRNAs; hence, expressing this circRNA or deleting the miR-7 had the same phenotypic outcome. Hansen et al. also reported that the testis-specific circRNA *Sry* (sex-determining region Y) serves as a miR-138 sponge.

The existence of the different forms of noncoding regulatory RNAs makes sense from the standpoint of building robustness in the regulatory network. However, it is tempting to speculate that the coexistence of various forms of noncoding RNAs may also determine the degree of titration needed to reach the threshold of effects in a cell-specific manner.

1.9 PROTEIN STRUCTURE AND FUNCTION

Proteins (polypeptides) are translated from the mRNA, which carries the amino acid sequence information for the polypeptide. Translation proceeds from the N-terminal to C-terminal direction of the polypeptide being synthesized. Proteins are made up of structural units called amino acids. All amino acids are **α-amino acids**. They are called α-amino acids because the amino group ($-NH_2$) is attached to the α-carbon atom—that is, the carbon atom linked to the carbonyl carbon of the carboxyl group (−COOH). The basic formula of an amino acid is shown in Figure 1.6A.

1.9.1 Configuration and Chirality of Amino Acids

All amino acids except glycine (R = H) are **chiral** because the α-carbon is chiral or asymmetric. So, **except for glycine** all amino acids can have two mirror-image stereoisomers (enantiomers). According to the DL system of Fischer, all natural amino acids are in L-configuration (as opposed to monosaccharides, which exist in D-configuration) (Figure 1.6B); according to the RS system of Cahn−Ingold−Prelog, all natural amino

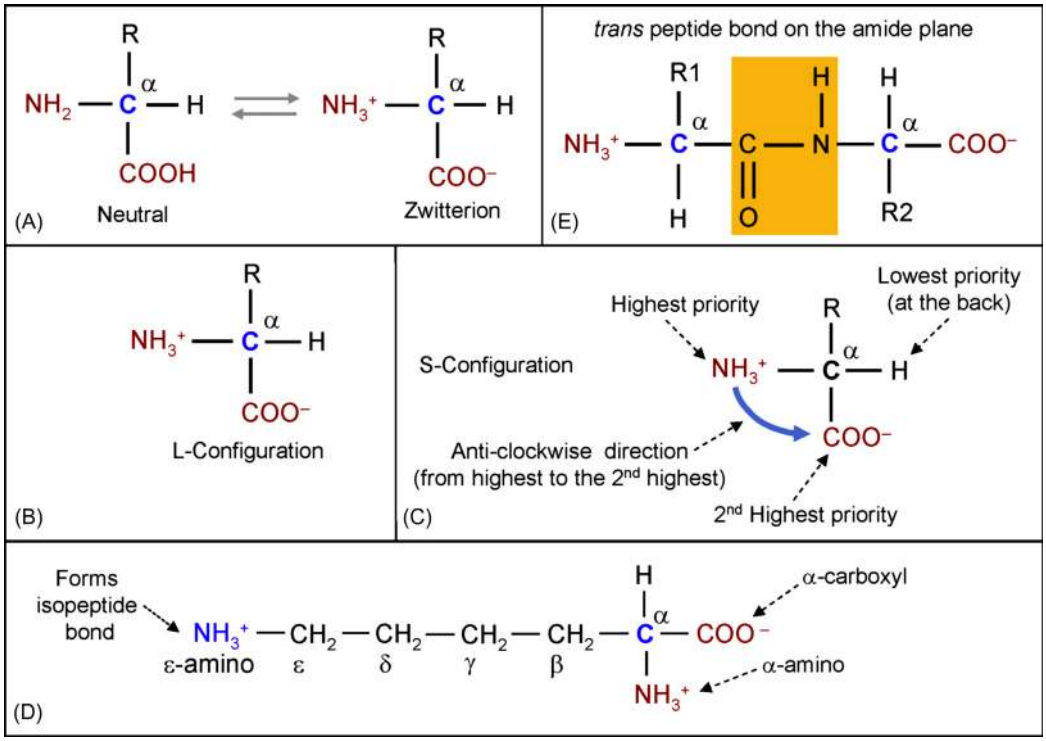


FIGURE 1.6 **Amino acid structure and peptide bond.** All amino acids except glycine (in which R = H) are chiral because the α-carbon is asymmetric. (A) Basic formula of amino acids; (B) L-configuration of amino acid per Fischer's system; (C) S-configuration of amino acid per Cahn−Ingold−Prelog rules; (D) the numbering of carbon atoms for lysine; (E) the peptide bond is a *trans* bond on the amide plane (in color).

BOX 1.8

1. The DL system of denoting enantiomers, originally introduced by Emil Fischer, is an old way of denoting the chirality of biological macromolecules. A more recent system is the RS system introduced by Robert Cahn, Christopher Ingold, and Vladimir Prelog. Naturally occurring amino acids have L-configuration according to the DL system, and S-configuration according to the RS system. In the RS system, first the priority of the groups attached to the chiral center is established. Then the order from the highest priority group to the second highest priority group, and so on, is established. If the order is clockwise, the molecule is said to have the R- (rectus) configuration; if the order is anticlockwise, the molecule is said to have S- (sinistrus) configuration. In Figure 1.6, NH_3^+ has the highest priority (because the atomic number of N is 7), followed by COO^- (because the atomic number of C is 6). If the first atom of two groups has the same atomic number, then the priority of the group is determined by the second atom and so on. Thus, COOH will have higher priority than CH_2OH.
2. The presence of two H atoms makes the α-carbon of glycine achiral (not chiral) or symmetric. As a result, glycine does not have any enantiomer (D/R or L/S isomer) and has no optical activity (dextro or levo).

acids are in the S-configuration (Figure 1.6C). So, the S-form is analogous to the L-form (see Box 1.8). Located on the alpha carbon is the "R" group, called the **side chain**. The nature of this side chain determines the identity of a particular amino acid. Glycine is the simplest amino acid because R = H. Amino acid side chains can be polar or nonpolar. Polar side chains may be charged or neutral. For example, two negatively charged amino acids are aspartic acid and glutamic acid. Two positively charged (i.e. protonated) amino acids are lysine and arginine. Figure 1.6D shows the numbering of carbon atoms of lysine. A small fraction of histidine is also positively charged at physiological pH. Proline is the only amino acid that has an imino group rather than an amino group. Although there are many more amino acids known so far, only 20 of them are standard amino acids used by all organisms during translation to synthesize proteins because they are encoded by the genetic code.

1.9.2 Ionic Character of Amino Acids

In solution at physiological pH (7.4), amino acids exist as dipole ions or **zwitterions**, where the amino group (NH_2) exists as an ammonium ion (NH_3^+) and the carboxyl group (COOH) exists as a carboxylate ion (COO^-) (Figure 1.6A). An amino acid can therefore act as a base as well as an acid, and hence is an ampholyte (having amphoteric properties). In a zwitterion, the + and − charges cancel each other to give the molecule a net charge of zero. However, at pH that is significantly higher or lower than physiological pH, amino acids undergo ionization. At acidic pH that is significantly lower than 7.4, the amino group has a positive charge while the carboxyl is neutral. At alkaline pH that is significantly higher than 7.4, the amino group is neutral while the carboxyl has a negative charge.

Amino acids of proteins in solution accept or lose protons depending on the nature of the side chains. The pK_a values of amino acids (i.e. the tendency of amino acids to lose protons) play an important role in determining the pH-dependent properties of a protein in solution. Internal ionizable groups in proteins are essential for catalysis. During a cycle of function, these internal ionizable groups can experience different microenvironments, and their pK_a values and charged states adjust accordingly.[29]

1.9.3 Relationship between Protein Function and the Location of Amino Acids in the Polypeptide Chain

The location of amino acids in the folded conformation of a protein is relevant for the protein's function and its interaction with the environment. For example, proteins located in a hydrophobic environment, such as membrane, have nonpolar (hydrophobic) side chains on the surface interacting with the membrane lipids. In contrast, proteins located in an aqueous environment, such as cytosol, have polar side chains (hydrophilic) on the surface interacting with the aqueous environment.

Arginine and lysine carry positive charges, and are often located on the interacting surface of proteins that interact with negatively charged molecules. Predictably, arginine and lysine are found on the surface of DNA-binding proteins that interact with the negatively charged phosphate group of DNA.

Similarly, aspartic acid and glutamic acid carry negative charges, and are often located on the interacting surface of proteins that interact with positively charged molecules. Aspartic acid and glutamic acid in calmodulin bind Ca^{++} ions, which carry a complementary positive charge. Many proteins in halophilic archaebacteria, which live in an extremely salty environment, have high localized concentrations (high charge density) of acidic amino acids on the surface. Such high charge density of acidic amino acids very effectively sequesters sodium ions, thus preventing denaturation and precipitation of cellular proteins. In fact, these proteins are denatured if placed in low salt concentration because the removal of sodium ions leaves many closely placed negative charges exposed, which strongly repel each other.

Serine, threonine, and tyrosine have hydroxyl groups (−OH) in their side chains. These OH groups can serve as phosphate attachment sites during phosphorylation. Many receptors that are involved in signal transduction are phosphorylated for activation, and consequently have these amino acid residues in their active sites. Phosphorylation causes conformational change in these receptors.

The sulfhydryl (-SH) group in cysteine is ideal for binding metals through metal—thiolate bonds. Naturally, cysteines are prevalent in many storage proteins that bind heavy metals. For example, in metallothionein, the intracellular metal-binding protein, one third of the amino acid residues are cysteines. The -SH group is also ideal for forming strong covalent disulfide linkages that stabilize the conformation of proteins. Expectedly, cysteines are found in many enzymes that function in harsh conditions of salt and pH, such as digestive enzymes like pepsin and chymotrypsin. The structure of many small proteins, such as insulin and ribonuclease, is stabilized by cysteine disulfide linkages. Cysteine disulfide linkages also confer rigidity to protein tertiary structure and are found in proteins like keratin in hair.

Proline occurs near the bend of polypeptide chains, and its ring forms a useful kink in the protein chain. Therefore, proline helps redirect the protein chain back inwards or around a tight corner.

Glycine and alanine, being very small, are flexible and can easily fit into tight spots. For example, glycine is the most abundant amino acid in the tight triple helix of collagen (about one-third of all amino acids). Alanine, being small and chemically inconspicuous, can be accommodated on the inside as well as outside of proteins. Alanine residues are very common in proteins. Attempts to confirm the functional role of specific amino acid residues in proteins involve mutagenesis experiments, and oftentimes the target amino acid is replaced by alanine.

1.9.4 Linkage between Amino Acids—The Peptide Bond

Amino acids are linked together by peptide bonds (**alpha peptide bonds**), which are simply **amide linkages** between the NH_2 and COOH groups of neighboring amino acids. The peptide bond has unique characteristics, which contribute to the overall structure of proteins. The peptide bond has a partial double-bond character. Thus, it is rigid and planar and not free to rotate. The plane on which it lies is called the **amide plane**. Peptide bonds are generally ***trans*** bonds—that is, the carbonyl oxygen and amide hydrogen are in *trans* position (Figure 1.6E). The Cα−C bonds are not rigid and they can freely rotate, being only limited by the size and character of the R groups. In lysine, the ε-amino group (Figure 1.6D) also participates in the formation of a peptide bond, which is called an **isopeptide bond** because it does not involve the usual α-amino group.

1.9.5 Four Levels of Protein Structure

Proteins have four levels of structure: primary, secondary, tertiary, and quaternary. **Primary structure** refers to the amino acid sequence of a protein. **Secondary structure** refers to the conformation of the polypeptide backbone. Examples of secondary structures are helices (**α-helix**), pleated sheets (**β-pleated sheet**), and bends or turns (**β-bend**). **Tertiary structure** of a protein refers to its three-dimensional structure—that is, further folding of the secondary structure in the three-dimensional space. **Quaternary structure** refers to a structure achieved by proteins composed of more than one polypeptide chain. Each polypeptide chain, called a subunit, has its own primary, secondary, and tertiary structure. In quaternary structure, protein chains (subunits) can associate with one another to form dimers, trimers, and other higher orders of oligomers. Recent studies have shown that despite having definitive structure, many proteins have specific regions that are intrinsically disordered (see Box 1.9).

1.9.6 Acidic and Basic Proteins

At physiological pH (7.4), acidic proteins tend to be negatively charged and have a higher proportion of acidic amino acids (e.g. aspartic acid, glutamic acid), whereas basic proteins tend to be positively charged and have a higher proportion of basic amino acids (e.g. arginine, lysine).

Hydrophilic and charged amino acids are frequently associated with antigenic determinants (**epitopes**),

BOX 1.9

INTRINSICALLY DISORDERED PROTEINS: THE "UNSTRUCTURAL" ASPECT OF STRUCTURAL BIOLOGY[30]

It has long been known that structural flexibility exists in proteins and aids in ligand binding. Nevertheless, the "structure–function paradigm"—that is, that proteins possess definitive three-dimensional structures in order to perform their function—has been the standard paradigm in protein biochemistry. Experimental evidence accumulating since the turn of the millennium has brought to light a unique aspect of protein structure that challenges this traditional structure–function paradigm once thought to be a universal theme applicable to all proteins. These findings demonstrate that under native functional conditions, many proteins or specific regions of some proteins are intrinsically disordered, existing as molten globules, collapsed or extended random coils, transiently structured forms, etc. These proteins are called **intrinsically disordered proteins** (**IDP**s). IDPs lack a unique three dimensional structure, either entirely or in part, when alone in solution. About 10–35% of prokaryotic and about 15–45% of eukaryotic proteins are estimated to contain disordered regions that are at least 30 amino acid residues in length. A significant number of IDPs are involved in regulatory and signaling functions; hence, IDPs are more prevalent in eukaryotes than in prokaryotes. IDPs and IDP databases are discussed in section 8.11 (Chapter 8).

such as arginine, lysine, aspartic acid, glutamic acid, asparagine, glutamine, serine, and threonine.

1.9.7 Nonstandard Amino Acids in Polypeptide Chains

As indicated earlier, selenocysteine and pyrrolysine are the two nonstandard amino acids that are incorporated directly into the polypeptide chain during translation. Selenocysteine has been found in lower as well as higher organisms (including mammals), while pyrrolysine has so far been found in certain archaebacteria. However, their occurrence in proteins is not nearly as universal as the 20 standard amino acids.

1.10 GENOME STRUCTURE AND ORGANIZATION

The genomic DNA in the nucleus exists in combination with **histone** proteins; the DNA–protein complex is known as **chromatin**. The unit of chromatin is the **nucleosome**; thus, chromatin can be envisioned as a repeat of regularly spaced nucleosomes. A nucleosome core particle is composed of a **histone** octamer and the DNA that wraps around the octamer (Figure 1.7). Histones are globular basic proteins with a flexible N-terminal end (the so-called "tail") that is subject to various covalent modifications (epigenetic modifications). The histone octamer is composed of two molecules each of histones H2A, H2B, H3, and H4. DNA wraps around the octamer in a left-handed supercoil of about 1.75 turns that each contain approximately 150 bp. Histone H1 is the **linker histone** that, along with **linker DNA**, physically connects the adjacent nucleosome core particles. Each nucleosome has a diameter of 10 nm, and the nucleosomes are compacted into a solenoid fiber structure of 30 nm (see Box 1.10). The 30-nm solenoid fibers undergo further progressive compaction into 300-nm filament, and ultimately into a 700-nm chromosome. During cell division, when the chromosomes duplicate, a 1400-nm metaphase chromosome is produced, containing two chromatids, each chromatid being 700 nm (Figure 1.7).

The major non-histone proteins associated with chromatin are the **high mobility group (HMG**) proteins. Whereas histones increase the compactness of the chromatin, HMG proteins decrease its compactness. By decreasing the compactness of the chromatin, HMG proteins facilitate the accessibility of various regulatory factors to DNA. HMG proteins can also bind to DNA and cause significant bending of the DNA. DNA bending is important for the interaction between transcription factors and **coregulators (coactivators/corepressors**[d]) in regulating transcription.

Various protein–DNA interactions can make the chromatin undergo changes in its conformation in response to various cellular metabolic demands. Altered chromatin conformation, in turn, can limit or enhance the accessibility and binding of the transcription machinery, thereby regulating transcription. Some of these regulatory effects could be mediated epigenetically.

[d]Coactivators and corepressors are proteins that do not bind DNA themselves, but interact with DNA-binding proteins, to either upregulate or downregulate transcription.

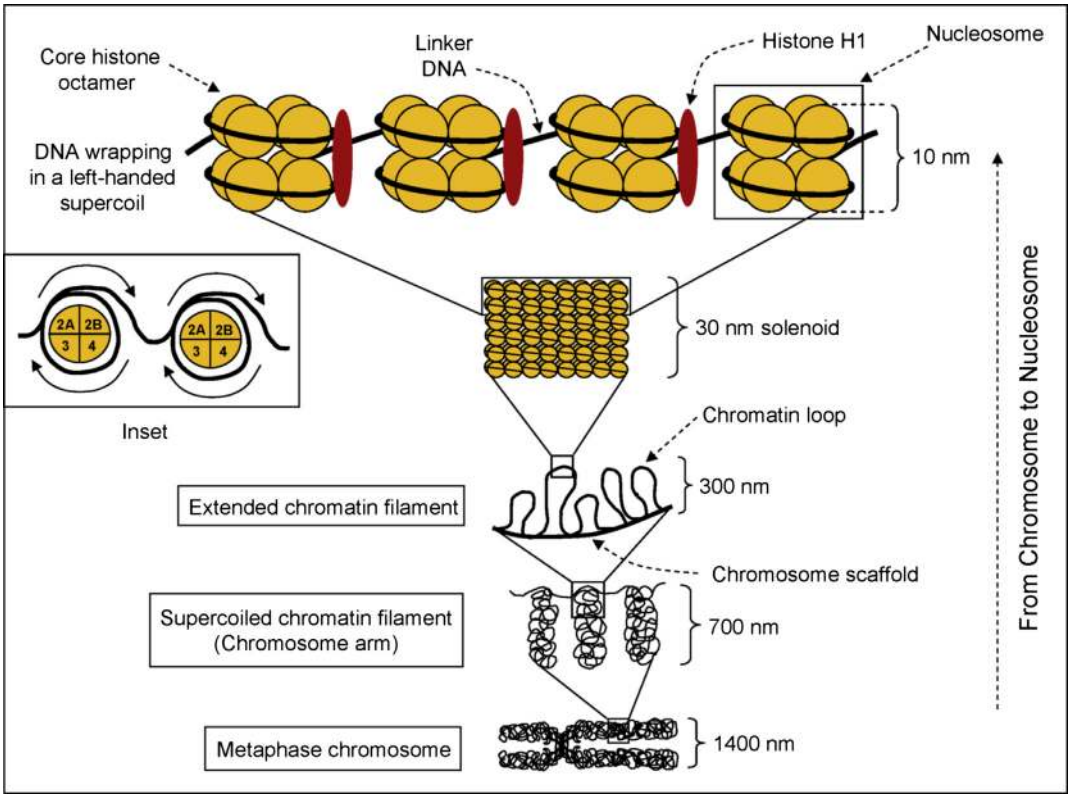


FIGURE 1.7 **The hierarchy of organization from chromosome to nucleosome.** Inset shows the relative position of histone monomers with respect to one another and the direction of wrapping of DNA around nucleosomes. *(Figure reproduced from Choudhuri et al. (2010) Toxicol. Appl. Pharmacol. 245: 378–393, with some modifications.)*

BOX 1.10

CHROMATIN FIBERS: 30 NM OR 10 NM?

Figure 1.7 shows the prevailing model of genome organization, which is the subject of textbooks. This model has been in existence since the mid-1970s, and it describes chromatin as a 30-nm fiber, which is formed by the coiling of the basic 10-nm fiber. Recent experimental evidence has challenged this traditional concept of chromatin organization.[31] By combining electron spectroscopic imaging with tomography, the authors generated a three-dimensional image that revealed that both open and closed chromatin domains in mouse somatic cells comprise 10-nm fibers. This indicates that the 30-nm chromatin model does not reflect the true regulatory structure in vivo. So, why was chromatin fiber reported to be 30 nm? This puzzle remains to be solved to the satisfaction of chromatin biologists. It has been suggested that it could be a combination of methodological artifact associated with chromatin isolation, as well as the inability to detect and distinguish the existence of the 10-nm fibers in the background of 30-nm fibers. Additional studies are expected to resolve this issue in the near future.

1.10.1 The Structure of a Representative Genome—The Human Genome

The human genome is discussed here as the representative genome.[32–34] The human genome consists of 3.2 billion (3.2×10^9) base pairs (=3.2 Gbp), distributed in 23 pairs of chromosomes (22 pairs of autosomes + XX or XY sex chromosomes). There are ~21,000 protein-coding genes, and the protein-coding fraction of the DNA constitutes ~1.5–2% of the entire genomic DNA. About two-thirds of the protein-coding genes have 1:1 **orthologs**[e] across placental mammals. Regulatory

[e]Genes in different species but related by speciation events are called **orthologous genes** or **orthologs**. Depending on the number of genes found in each species, the relationship of orthologs could be 1:1, 1:many, and many:many.

sequences constitute ~3–3.5% of the genome. The genome also codes for a significant number of noncoding regulatory RNAs. Initial studies suggested that more than 10% of the genome is represented in mature transcripts, and ~20% of the genome may be functionally important. These estimates have been revised and significantly expanded based on the findings of the Encyclopedia of the DNA Elements (ENCODE) project, discussed later. *The genomes of two humans are about 99.9% identical*.

Repeat sequences account for ~50% of the human genome; hence repeat sequences constitute a significant source of genetic diversity. Repeat sequences are of various types: **simple repeats** (e.g. $(A)_n$, $(CA)_n$, $(CGG)_n$), **tandem repeat blocks** (e.g. centromeric repeats, telomeric repeats, ribosomal gene clusters), **segmental duplications** (e.g. blocks of 1–200 kb or longer repeats copied from one region of the genome and integrated into another region of the genome), **interspersed repeats** (transposable-element-derived), and **processed pseudogenes**. In addition to the repeat content, further functional genetic diversity is imparted by **single nucleotide polymorphism** (**SNP**) and **copy number variation** (**CNV**), also called **copy number polymorphism** (**CNP**). According to older definition, a point mutation has to occur in at least 1% of the population in order to qualify as an SNP, but this is no longer strictly followed; all point mutations are called SNPs. In the human genome, >65% of all SNPs are C→T transition mutations.

Recent evidence suggests that the human genome is extensively transcribed. However, the fraction of the genome that is transcribed into functional noncoding transcripts is yet to be estimated precisely. The findings from the **Encyclopedia of the DNA Elements** (**ENCODE**) project suggest that the noncoding yet functional fraction of the genome may vary significantly from chromosome to chromosome. There is also evidence for both sense and antisense transcription in the human genome. There is extensive alternative splicing of transcripts so that there are well above 100,000 proteins encoded by the human genome.

The G + C-rich regions of the genome are gene-dense, and the genes in these regions are smaller and more compact due to smaller intron size. Conversely, A + T-rich regions are gene-poor and the genes in these regions are longer because of longer intron size. Average G + C content of the entire human genome is 41%, but local G + C contents may vary significantly. An important component of the G + C-rich genomic regions is the CpG sequence, which may or may not occur in clusters. CpG clusters are called **CpG islands**. The human genome contains about 0.8% CpG islands. However, based on the G + C content (~41%), the CpG island frequency should be ~4%. The difference is due to the fact that the cytosine of the CpG island is methylated, and over evolutionary time the methyl cytosine (^{me}C) tends to spontaneously deaminate to thymine, hence converting CpG to TpG. The $^{me}C \rightarrow T$ mutation creates a T–G mismatch in the DNA double strand and is normally repaired; however, it sometimes escapes the repair machinery (e.g. if it happens before replication and strand separation). The CpG islands are associated with the 5′-ends of many genes. Identification of CpG islands thus helps define the 5′-ends of genes. Methylation of the C of CpG is associated with transcriptional silencing, and the absence of methylation is associated with active transcription. Thus, unmethylated CpG islands are associated with the promoters of transcriptionally active genes, such as housekeeping genes, and genes showing tissue-specific expression.

The birth of new genes and the death of existing genes in the genome are important events that contribute to genome evolution. New genes can be born or acquired by a genome. New genes can be born through one of multiple genomic events, such as **gene duplication**, **de novo gene origination**, and **transposable element (TE) domestication**. Duplicated genes can diverge and acquire new function. These genes are called **paralogous genes** or **paralogs**[f]. New genes can be born de novo by functionalization of a previously noncoding region of the DNA. Sometimes genomes can recruit TEs and use the TE-encoded protein as the cellular protein. New genes can also be acquired through **lateral gene transfer**. Genome evolution is discussed in more detail in Chapter 2.

Gene death occurs when genes acquire inactivating mutations and lose function. Pseudogenization is a common mechanism of gene death. Pseudogenes may be **non-processed pseudogenes** or **processed pseudogenes**. Non-processed pseudogenes are an inactivated form of a gene that has acquired inactivating mutations; hence they may have intact exon–intron organization but the ORF is disrupted. In contrast, processed pseudogenes result from the reverse transcription of mRNA into complementary DNA (cDNA), followed by the integration of the cDNA into the genome. Thus, processed pseudogenes may have a poly(A) tail but they lack a promoter and other 5′-regulatory elements. (see Box 1.11)

[f]**Paralogous genes** or **paralogs** are produced through gene duplication within a genome. Paralogs may evolve new functions or may become pseudogenes.

BOX 1.11

More than a decade after genome sequencing, we are still far from understanding many aspects of structural and functional genomics, such as the exact number of protein-coding and non-protein-coding genes and their genomic locations; the genome-wide distribution of functional regulatory elements; the regulation and coordination of gene expression at different levels and regulation of the regulators; chromatin dynamics; epigenetic editing of the language of DNA; gene and protein networks; protein−protein interactions; regulation of interaction specificity in biological systems and the specificity determinants, such as protein interaction specificity and signaling specificity; the correlation between genetic diversity and disease susceptibility; the molecular determinants of humanness, that is, what it means to be a human at the molecular level; and many such similar questions.

1.10.2 Functional Sequence Elements in the Genome

Functional sequence elements of the genome regulate genome expression. These are promoters, enhancers, silencers, locus control regions (LCRs), and insulators. Elements that aid in the termination of transcription (terminators) are not discussed here.

1.10.2.1 Promoters

The 5′-flanking region of the gene is the region upstream of the transcription start site (+1). It contains the promoter and other *cis*-acting transcription regulatory sequence elements. A **promoter** is a *cis*-acting transcription regulatory element that initiates the transcription of a gene. The various regions of the promoter are termed the **core** (or **basal**) promoter, **proximal** promoter, and **distal** promoter, based on their distance from the transcription start site. Typically, the core promoter is about 35 bp long, and can extend between the −35- and +35-nt position (with respect to the +1 site). The core promoter may contain two or more of the following sequence motifs: **TATA box**, **initiator** (**Inr**) **element**, and **downstream promoter element** (**DPE**). Upstream of core promoter is the proximal promoter, which is about 250-bp long and can extend between the −250 and +250-nt position. However, in the literature, sequences far upstream of −250 have also been referred to as proximal promoter sequences. Sequences that are further upstream of the proximal promoter elements are called the distal promoter. In general, the transcription start site is determined by the TATA box and the initiator element, or in the case of TATA-less promoters, by the initiator element and the downstream promoter element, all located within the core promoter.[18]

1.10.2.2 Enhancers

Enhancers bind specific transcriptional activators and enhance the rate of transcription. Enhancers can be located close to the transcription start site, upstream or downstream from the transcription start site, and even within introns. An enhancer can regulate more than one gene in a position- and orientation-independent manner. The mechanism of enhancer action is thought to involve looping of the DNA, thereby bringing the enhancer-bound transcriptional activators close to the promoter-bound transcription factors. In doing so, enhancers increase the concentration of activators near the promoter, which directly or indirectly interact with the promoter to initiate transcription. The interaction of enhancer-bound transcriptional activators and promoter-bound transcription factors is mediated by coactivators. **Coactivators** are proteins that do not bind DNA themselves but interact with DNA-bound transcriptional activator proteins, thereby facilitating protein−protein interaction. Some examples of coactivator proteins are CBP/p300, p160, p300/CBP-interacting protein (p/CIP), p300/CBP-associated factor (p/CAF), yeast transcriptional adaptor GCN5, steroid receptor coactivator-1 (SRC-1), and there are many others. The opposite of enhancers are **silencers**, which bind transcriptional suppressor proteins and suppress transcription, thereby acting as negative regulatory elements. Like enhancers, silencers can also function in an orientation-, position-, and distance-independent manner, and they can also be located within introns.

1.10.2.3 Locus Control Regions

A locus control region (LCR) enhances the transcription of a cluster of linked genes by inducing a more open conformation of the chromatin flanking the locus. The LCR of the human β-globin locus has been well studied. The transcription-enhancing activity of LCRs is mediated by the binding of specific transcriptional activator proteins. Because LCRs can induce conformational change of chromatin, they play important roles in regulating the transcriptional activity of the euchromatic regions of chromosomes.

1.10.2.4 Insulators

Insulators are gene-boundary elements; these are DNA sequence elements that, when bound to insulator-binding proteins, shield a promoter from the effects of nearby regulatory elements. There are two types of insulator functions: an **enhancer-blocking function** and a **heterochromatin barrier function**. When an insulator is located in between a promoter and an enhancer, the enhancer-blocking function of the insulator shields the promoter from the transcription-enhancing influence of the enhancer. The heterochromatin barrier function of an insulator prevents a transcriptionally active euchromatic region from turning into transcriptionally inactive heterochromatin by the inactivating effect of the invading adjacent heterochromatin[g]. An example of an enhancer-blocking insulator is the **gypsy** insulator in *Drosophila*. The chicken **β-globin insulator** (**cHS4**), which is highly rich in G + C and the most extensively studied vertebrate insulator, has both enhancer-blocking and heterochromatic barrier functions. The mechanism of the enhancer-blocking function may involve DNA looping, but it is yet to be established. However, the mechanism of heterochromatic barrier function understandably involves the maintenance of active chromatin configuration through histone modifications at the boundary. Various proteins that bind to these insulator sequences have been identified.[35]

1.10.3 Epigenetic Modifications of the Genome Can Edit the Language Written in the DNA Sequence and Add an Extra Layer of Complexity in Genome Expression

Epigenetics is the study of mitotically or meiotically heritable changes in gene function that cannot be explained by changes in the DNA sequence.[36] Epigenetic inheritance involves the transmission of epigenetic marks not encoded in the DNA sequence, from parent cell to daughter cells and from generation to generation. Epigenetic regulation of genome expression is mediated by three main mechanisms: (1) **DNA methylation**, (2) **histone modification and chromatin conformation change**, and (3) **regulation of gene expression by ncRNAs**. DNA methylation involves the covalent addition of a methyl group to the carbon-5 position of cytosine to form 5-methylcytosine (5-mC) in CpG dinucleotides. Methylation is catalyzed by three major DNA methyltransferases (DNMTs), and the methyl group donor is S-adenosylmethionine (SAM). The de novo methylation establishes the parent-specific methylation pattern, and maintenance methylation replicates the methylation pattern of the parent strand to the daughter strand during DNA replication. This is accomplished by first recognizing the hemi-methylated CpG sites at the replication foci, followed by the addition of methyl groups to cytosines on the nascent DNA strand to re-establish the parent-specific methylation pattern. The de novo methyltransferases are DNMT3A and DNMT3B, whereas the maintenance methyltransferase is DNMT1.

Methylation of the C of CpG is associated with transcriptional silencing, and the absence of methylation is associated with active transcription. Thus, unmethylated CpG islands are associated with the promoters of transcriptionally active genes, such as housekeeping genes and genes showing tissue-specific expression. Transcriptional silencing by DNA methylation is mediated by a condensed state of chromatin. Conversely, transcriptionally active genes maintain an open state of chromatin.

Covalent histone modification—such as acetylation, methylation, phosphorylation, ubiquitination, or sumoylation of specific amino acid residues, such as lys (K), arg (R), ser (S) and others, but mainly lys residues of different histone subunits—can either upregulate or downregulate gene expression. All known histone acetylation and phosphorylation modifications are transcription-activating, whereas all known sumoylations are transcription-silencing. Histone methylation and ubiquitination can be transcription-activating or silencing, depending on the specific residue modified. Table 1.2 shows some transcriptional-activating and repressing histone modifications. Epigenetic orchestration of genome expression is a tightly regulated process and it involves the cross-talk between DNA methylation and histone modifications.[37]

Regulation by small ncRNAs (e.g. miRNAs, siRNAs) is another means of epigenetic regulation of gene and genome expression. Small ncRNA-mediated silencing of gene expression, known as **RNA interference** (**RNAi**), is achieved either by translational repression (by miRNA) or by mRNA degradation (by siRNA).[22]

Some of the relatively well studied examples of epigenetic phenomena regulating gene and genome expression are **transvection** (observed in dipteran insects), **genomic imprinting**, **X-chromosome inactivation**, **paramutation**, and **heterochromatin spread and position effect variegation**.[38] Although epigenetic mechanisms can edit the language of DNA written in its

[g]Sometimes, indiscriminate propagation of heterochromatin into adjacent euchromatin results in silencing of genes located in close proximity to the propagating heterochromatin. The silencing is often not complete; the genes are silenced in some cells, but in other cells they are expressed, resulting in a so-called variegated (patchy) expression pattern. Because this expression pattern is brought about by the proximity of the genes to the heterochromatin, the phenomenon is called **position-effect variegation (PEV)**.

TABLE 1.2 Some Transcription-Activating and Repressing Histone Modifications

Some Transcription-Activating Modifications
Acetylation
Histone H2A: K5, K9, K13; **Histone H2B:** K5, K12, K15, K20; **Histone H3:** K9, K14, K18, K23, K56; **Histone H4:** K5, K8, K13, K16
Phosphorylation
Histone H3: T3, S10, S28, Y41; **Histone H2AX:** S139 (for DNA repair)
Methylation (me1/me2/me3)
Histone H3: K4, K9 (me1), K36, K79, R17, R23; **Histone H4:** R3
Ubiquitination
Histone H2B: K120, K123 (yeast)
Some Transcription-Silencing Modifications
Methylation (me1/me2/me3)
Histone H3: K9 (me2, me3), K27; **Histone H4:** K20
Ubiquitination
Histone H2A: K119
Sumoylation
Histone H2A: K126 (yeast); **Histone H2B:** K6, K7 (yeast); **Histone H4:** K5, K8, K12, K16, K20

base sequence, thereby altering genome expression, epigenetic modulation of gene and genome expression needs further characterization. For example, much needs to be understood in terms of the correlative versus causal effects between exposure to various environmental factors and epigenetic changes. Additionally, we are not yet able to distinguish between adaptive and adverse epigenetic changes. Normal epigenetic changes associated with age and different life stages need to be thoroughly characterized as well. Some preliminary data are available but more work is underway.

1.10.3.1 Histone Code

Strahl and Allis[39] coined the term **histone code** to describe the concept that specific histone modifications could act sequentially or in combination to form a recognizable "code" that could regulate transcription as well as the state of chromatin condensation. Turner[40] used the term **epigenetic code**, which was conceptually same as the histone code. For example, phosphorylation of histone H3 serine 10 (H3S10) stimulates acetylation of histone H3 lysine 14 (H3K14), which is a transcription-activating modification; monoubiquitination of histone H2B lysine 120 (H2BK120) stimulates methylation of histone H3 lysine 4 (H3K4), which is also a transcription-activating modification.[41] See Box 1.12 regarding symmetrical and asymmetrical histone code.

BOX 1.12

ASYMMETRICAL MODIFICATION OF HISTONE AND ASYMMETRICAL HISTONE CODE

The traditional view assumes that histone code is symmetrical; that is, both molecules of the same histone in a nucleosome are modified in the same way. However, recent experimental evidence challenges this long-held view.[42] Using preparations of chromosomal mononucleosomes from embryonic stem cells, mouse embryonic fibroblasts, and cultured HeLa cells, the authors showed the existence of di- and trimethylation of lysine 27 of histone H3 (H3K27me2/3) both symmetrically and asymmetrically in native chromatin in approximately equal proportions. When the H3K27me2/3 mark occurred asymmetrically there was a different methylation mark on the sister histone, either H3K4me3 or H3K36me2/3. In other words, in a nucleosome, one of the two H3 molecules contains one mark, while the other H3 contains a different mark. Whereas H3K4me3 or H3K36me2/3 are transcription-activating modifications, H3K27me2/3 is transcription-repressing modification. The coexistence of such antagonizing histone modification marks might facilitate rapid and efficient regulation of transcription because the removal of one of these marks may be sufficient to rapidly induce transcriptional activation or repression. The existence of asymmetric histone modifications also shows that histone code could be symmetric or asymmetric. The possibility of existence of asymmetric histone modification marks throughout the genome significantly expands the scope of epigenetic regulation, particularly when the combinatorial aspect of such modifications and their effect on transcription are taken into account.

1.10.3.2 The Dynamics of Epigenetic Changes

Epigenetic modifications, particularly DNA methylation, have been traditionally regarded as static modifications. Progress in epigenetics during the past few years has demonstrated that epigenetic modifications of the genome are lot more dynamic than initially thought. A recent study in mice[43] suggests that epigenetic modifications can even control circadian rhythms of gene expression, thereby regulating circadian-rhythm-driven physiological processes. The authors observed circadian oscillations of several antisense RNA, long noncoding RNA, and microRNA transcripts coupled with rhythmic histone modifications in promoters, gene bodies, or enhancers in adult mouse livers. Promoter DNA methylation levels were relatively stable. The authors identified a set of 1262 (9% of expressed) oscillating transcripts, of which 1160 were protein-coding, including genes implicated in metabolic regulation, such as *Arntl*, *Cry1*, *Per1*, *Per2*, *Per3*, *Rorc*, *Foxo3*, and many others. The five investigated histone modifications—H3K4me1, H3K4me3, H3K9ac, H3K27ac, and H3K36me3—were enriched in actively transcribed genes and correlated with transcript levels. The oscillating expression of an antisense transcript (*asPer2*) to the gene encoding the circadian oscillator component *Per2* was also identified. Robust transcript oscillations often accompanied rhythms in multiple histone modifications and recruitment of multiple chromatin-associated clock components. The findings of this study, as well as some other studies before it, demonstrate that epigenetic modifications could be very dynamic and may even control rapid and short-term regulation of gene expression.

1.10.4 Lessons Learned from the Second Phase of the ENCODE Project about the DNA Elements in the Human Genome and its Epigenetic Modifications

The Encyclopedia of DNA Elements (ENCODE) project has been a logical continuation of the big science that was launched with the human genome sequencing project. ENCODE aims to delineate all functional elements encoded in the human genome. *A functional element is defined as a discrete genome segment that either encodes a product (e.g. protein or noncoding RNA) or displays a reproducible biochemical signature (e.g. protein binding, or a specific chromatin structure).* Following the initial success of the first phase of ENCODE, initiated in 2003 to characterize 1% of the human genome, the scope of ENCODE has been broadened since 2007 to study DNA elements in the whole human genome. The work in the second phase involved integration of results from experiments involving 147 different cell types, and all ENCODE data, with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions.[44,45]

Based on the analysis, about 80% of the genome was assigned some kind of genetic function, either RNA-associated or chromatin-associated. About 95% of the genome was found to lie within 8 kb of a DNA−protein interaction, and 99% within 1.7 kb of at least one of the biochemical events measured by ENCODE. The analysis annotated 8801 small RNA and 9640 long noncoding RNA-coding loci. Greater than 62% of the genomic bases were found to be represented in > 200-nt-long RNA molecules. Most transcribed bases were found to be within annotated genes or in overlapping annotated gene boundaries; that is, in noncoding DNA. Also, 11,224 pseudogenes were annotated, of which 863 are transcribed and associated with active chromatin.

An initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features were annotated. A total of 62,403 transcription start sites were identified, of which 27,362 (44%) are within 100 bp of the 5′-end of an annotated or known transcript. The remaining regions predominantly lie across exons and 3′-UTRs, some exhibiting cell-type-restricted expression, representing possible start sites of novel cell-type-specific transcripts. The binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types were mapped using chromatin immunoprecipitation followed by deep sequencing (ChIP-seq); 87 (73%) were sequence-specific transcription factors. Overall, 636,336 binding regions covering 231 megabases (8.1%) of the genome were found to be enriched for regions bound by DNA-binding proteins across all cell types.

Statistical models to analyze genome-wide transcription-factor-binding data identified six different types of genomic region, based on the binding data of transcription-related factors (TRFs). These six different types of genomic region form three pairs: (1) binding-active regions (BARs) and binding-inactive regions (BIRs), (2) promoter-proximal regulatory modules (PRMs) and gene-distal regulatory modules (DRMs), and (3) high-occupancy of TRF (HOT) regions and low-occupancy of TRF (LOT) regions. Region types from different pairs may overlap. For example, DRMs are subsets of BARs, and some HOT regions overlap with PRMs and DRMs. Each of these regions, however, exhibits some unique properties. The six types of region were found to occupy from about 15.5 Mbp (equivalent to 0.50% of the human genome) to 1.39 Gbp (equivalent to 45% of the human genome) in the different cell lines. Expectedly, the distribution of BARs correlates with gene density. Also, about 70 to 80% of the HOT regions were mapped within 10 kb of annotated coding and noncoding genes.

Assay for histone modifications and variants in 46 cell types showed a great deal of variability across cell types, in accordance with changes in transcriptional activity. For example, monomethylation of lysine 4 of histone H3 (H3K4me1) was found as a mark of regulatory elements associated with enhancers and other distal elements, H3K4me2 was found as a mark of regulatory elements associated with promoters and enhancers, whereas H3K4me3 was found as a mark of regulatory elements primarily associated with promoters/transcription starts. In contrast, H3K9me3 is the repressive mark found associated with constitutive heterochromatin and repetitive elements.

In conclusion, the map created by ENCODE reveals that cell type is important. In other words, cell-type-specific regulation of genome expression in multicellular organisms might hold the key to explaining not only differential regulation of gene expression, but also the development of disease.

References

1. Graci JD, Cameron CE. *Antiviral Chem Chemotherap* 2004;**15**:1–13.
2. Segal DJ, et al. *Proc Natl Acad Sci USA* 1999;**96**:2758–63.
3. Le Doan T, et al. *Nucl Acids Res* 1987;**15**:7749–60.
4. Beal PA, Dervan PB. *Science* 1991;**251**:1360–3.
5. Pilch DS, et al. *Biochemistry* 1991;**30**:6081–8.
6. Grigoriev M, et al. *Proc Natl Acad Sci USA* 1993;**90**:3501–5.
7. Chin JY, et al. *Front Biosci* 2007;**12**:4288–97.
8. Kumar A. *Eukaryot Cell* 2009;**8**:1321–9.
9. Kay RA, et al. *Cancer Res* 2005;**65**:10742–9.
10. Hawkins JD. *Nucl Acids Res* 1988;**16**:9893–908.
11. Sterner DA, et al. *Proc Natl Acad Sci USA* 1996;**93**:15081–5.
12. Choudhuri S, et al. *Biochem Biophys Res Commun* 2000;**274**:79–86.
13. Belshaw R, Bensasson D. *Heredity* 2006;**96**:208–13.
14. Lambowitz AM, Zimmerly S. *Annu Rev Genet* 2004;**38**:1–35.
15. Chorev M, Carmel L. *Front Genet* 2012;**3**:Article 55.
16. Sauna ZE, Kimchi-Sarfaty C. *Nat Rev Genet* 2011;**12**:683–91.
17. Edwalds-Gilbert G, et al. *Nucl Acids Res* 1997;**25**:2547–61.
18. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 3–48.
19. Leontis NB, Westhof E. *RNA* 2001;**7**:499–512.
20. Abu Almakarem AS. *Nucl Acids Res* 2012;**40**:1407–23.
21. Choudhuri S. *Biochem Biophys Res Commun* 2009;**388**:177–80.
22. Choudhuri S. *J Biochem Mol Toxicol* 2010;**24**:195–216.
23. Mercer TR, Mattick JS. *Nat Struct Mol Biol* 2013;**20**:300–7.
24. Amaral PP, et al. *Nucl Acids Res* 2011;**39**:D146–51 (Database Issue)
25. Tay Y, et al. *Cell* 2011;**147**:344–57.
26. Salzman J, et al. *PLoS ONE* 2012;**7**(2):e30733.
27. Memczak S, et al. *Nature* 2013;**495**:333–8.
28. Hansen TB, et al. *Nature* 2013;**495**:384–8.
29. Isom DG, et al. *Proc Natl Acad Sci USA* 2011;**108**:5260–5.
30. Tompa P. *Trends Biochem Sci* 2012;**37**:509–16.
31. Fussner, et al. *EMBO Rep* 2012;**13**:992–6.
32. Phesant M, Mattick JS. *Genome Res* 2007;**17**:1245–53.
33. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 49–99.
34. Lander ES. *Nature* 2011;**470**:187–97.
35. Valenzuela L, Kamakaka RT. Chromatin insulators. *Annu Rev Genet* 2006;**40**:107–38.
36. Riggs AD, et al. Introduction. In: Russo VEA, et al., editors. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor, NY: CSHL Press; 1996. p. 1–4.
37. Choudhuri S, et al. *Toxicol Appl Pharmacol* 2010;**245**:378–93.
38. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 101–28.
39. Strahl BD, Allis CD. *Nature* 2000;**403**:41–5.
40. Turner B. *Bioessays* 2000;**22**:836–45.
41. Choudhuri S. *Toxicol Mech Methods* 2011;**21**:252–74.
42. Voigt P, et al. *Cell* 2012;**151**:181–93.
43. Vollmers C, et al. *Cell Metab* 2012;**16**:833–45.
44. The ENCODE Project Consortium. *Nature* 2012;**489**:57–74.
45. Yip KY, et al. *Genome Biol* 2012;**13**:R48.

CHAPTER

2

Fundamentals of Molecular Evolution*

OUTLINE

2.1 BIOINFORMATICS, MOLECULAR EVOLUTION, AND PHYLOGENETICS

Probably, the shortest classical definition of evolution is *descent with modification* from the ancestor. Evolutionary changes lead to changes in the inherited characters in a population[a]. The ultimate outcome of evolution is the formation of new species (**speciation**), but evolution can generate diversity at all possible levels of biological organization including at the level of macromolecules, such as DNA and proteins.

Molecular evolution is a relatively recent discipline that has developed since DNA and protein sequence information became available. Simply stated, molecular

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

[a]A population is composed of members of a species occupying a geographic area. A community is composed of members of different populations occupying the same geographic area.

evolution is evolution at the level of nucleic acids and proteins. At the molecular level, the primary cause of evolution is the accumulation of changes in genomic sequence (hence proteins as well[b]). Therefore, evolution results in alteration of the genetic composition (**gene pool**) of a population over time. Changes in gene pool are associated with changes in gene frequency in a population[c].

The work of Emile Zuckerkandl and Linus Pauling between 1960 and 1965, particularly their seminal publication in 1965,[1] is credited with ushering in a change in evolutionary thinking from the level of species to the level of macromolecular sequence. Such a paradigm shift in evolutionary thinking from population to macromolecular sequence essentially paved the way for the birth of a new field, molecular evolution. The classical definition of evolution as *descent with modification* refers to the event of speciation—that is, the formation of new species from an ancestral species. The same definition and concepts also apply to molecular evolution except for the fact that the targets of molecular evolution are nucleic acid and protein sequences. The causes of molecular evolution, such as mutation, recombination, gene conversion, duplication and divergence of genes, de novo origin of new genes, and structural and functional evolution of genomes, as well as changes in gene frequency in a population, are also at the heart of evolution at the level of species and beyond.

The availability of the complete genome sequence of many species provides a wealth of data and information for molecular evolutionary studies and comparative genomics. *Evolutionary biology provides the scientific context and bioinformatic analysis utilizes the analytical tools for comparative genomics*. In the context of evolutionary biology, the goal of various applications of bioinformatics, such as sequence alignment, sequence identity/similarity search, motif analysis, sequence homology analysis, chromosomal synteny analysis, and making phylogenetic trees, is to trace the signature and determine the rate of molecular evolution, as well as study the relatedness of taxa. Following the spirit of the now-famous statement by Dobzhansky that "nothing in biology makes sense except in the light of evolution, "Higgs and Attwood (2005) have stated, "nothing in bioinformatics makes sense except in the light of evolution".[2] This is a very astute way of summarizing the relationship between bioinformatics and molecular evolution.

It has become a standard practice in studies involving DNA or protein sequence to obtain a phylogenetic tree and assess sequence divergence. Freely available software on the web has made it almost effortless to input the data and quickly get an output. Because of such widespread use of DNA and protein sequence analysis and phylogenetic inference, it is important to understand the principles of molecular evolution. The following narrative summarizes some fundamental concepts of molecular evolution that help in understanding the evolutionary foundations of bioinformatics.

2.2 BIOLOGICAL EVOLUTION AND BASIC PREMISES OF DARWINISM

Biological evolution is most simply defined as *descent with modification*; the modification may be small scale (e.g. changes in gene/protein sequence) or large scale (e.g. speciation). After life had originated on Earth about 3.6 billion (3600 million) years ago, it evolved from simple to progressively complex forms, all from one primordial ancestral form, called the **last universal common ancestor** (**LUCA**). The evolutionary history of the descendants of LUCA constitutes the **tree of life**.

Evolution of life is a continuous process involving splitting of lineages, divergence of the descendants, and adaptive radiation into different environments (ecological niches) creating phenotypic diversity, and ultimately leading to reproductive isolation and the formation of new species (**speciation**). It is important to note in this context that even though "species" is an accepted taxonomic category, the concept of species and speciation is a hotly debated issue even 150 years after the publication of Darwin's *On the Origin of Species*. We will follow the most widely used definition of species, provided by the **biological species concept**.

Two pioneering architects of the biological species concept were Theodosius Dobzhansky and Ernst Mayr. According to Mayr's classical definition of species, "species are groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups"[d].[3] In other

[b]Changes in genomic sequence include changes in the sequence of protein-coding genes, non protein-coding genes, and regulatory sequences, as well as intergenic regions. Such changes may result in altered gene expression and trigger genome evolution.

[c]A small-scale change within a population below the species level, such as a change in allele frequencies, is called microevolution. Microevolution can be observed over a short period of time, such as across a few generations (e.g. development of resistance). In contrast, large-scale changes and evolution at or above the species level and over a long period of time are called macroevolution.

[d]This definition of species was originally proposed in Mayr's now-classic book *Systematics and the Origin of Species* (1942, Columbia University Press, New York). However, Mayr's definition of species owed its origin to the concept of species proposed by Dobzhansky in his famous book *Genetics and the Origin of Species* (1937, Columbia University Press, New York). Dobzhansky conceptualized species as "that stage in the evolutionary process at which the once actually or potentially interbreeding array of forms becomes segregated in two or more separate arrays which are physiologically incapable of interbreeding."

words, a species is a reproductive community that represents a unique gene pool. Genetic exchange between members of two different gene pools is usually not successful in producing fertile offspring that could perpetuate the existence of the species. When populations within a species become isolated by geography, mate selection, or other means that interfere with mating, they may start to diverge and over time may evolve into new species.

Darwin's theory of evolution by natural selection states that (1) variations exist among the organisms of a population, (2) the resources (food and space) are limited, (3) the scarcity of resources would lead to competition among individuals, and (4) individuals with favorable variations are more likely to survive in the competition whereas those that do not have the favorable variations simply die out. Those that survive will reproduce, increase in number, and occupy a specific environment. This process, which removes some organisms from the population but favors (selects) others, is called **natural selection** and it is a **passive process** acting like a sieve. Natural selection could be **purifying** (**negative**) **selection** that removes deleterious variations, and **positive** (**Darwinian**) **selection** that fixes the beneficial variations in the population and promotes the emergence of new phenotypes. When the organisms with favorable variations reproduce, the variations spread in the population and help the population to better adapt to the environment. Over many generations, the population adapted to a specific environment evolves into a new species that becomes reproductively isolated from other such groups. The coupling of Darwinism with modern genetics transformed classical Darwinism into **neo-Darwinism** (also known as **modern synthesis** or the **synthetic theory of evolution**).

The Darwinian evolutionary process predicts that the pace of evolution is gradual because an evolving population accumulates small variations over a long period of time. Hence, the divergence of lineages is slow, steady, and stepwise. For example, for a species A to evolve into species B, it should go through many stages, such as A_1, A_2, A_3 ... A_n until it evolves into B. This gradual pace of evolution through incremental changes is known as **phyletic gradualism**. However, the fossil records for most species are incomplete and they do not show the existence of small incremental changes on the way to the new species[e]. To account for the lack of fossil records showing phyletic gradualism, paleontologists Stephen J. Gould and Niles Elredge[4] put forth a competing hypothesis, which claims that species are generally stable, changing little over long periods of time. This condition of little or no change is called **stasis**. The stasis is punctuated by rapid bursts of evolutionary changes that result in the formation of new species. As a result, this process leaves few fossils behind, which can explain the absence of many intermediate forms in the fossil record. Gould and Elredge termed this phenomenon **punctuated equilibrium**. In reality, both phyletic gradualism and punctuated equilibrium could have played a role in evolution.

A basic assumption of the Darwinian theory is that new mutations, both advantageous and deleterious, constantly arise in the population independent of need, and *evolution is caused by natural selection acting through beneficial mutations by fixing them in the population*. Darwinian evolution does not consider neutral mutations that do not confer any selective advantage or disadvantage to be of any importance in the evolutionary process. This long-held view of Darwinian evolution was challenged by the neutral theory of molecular evolution. The neutral theory is discussed later in this chapter.

2.2.1 First Experimental Demonstration of Evolutionary Principles in the Test Tube

Sol Spiegelman and colleagues[5] first demonstrated that **Darwinian evolutionary principles**—that is, **variation**, **selection**, and **amplification**—could lead to the evolution of biological macromolecules in the test tube in an extracellular environment. Spiegelman and coworkers explored the evolutionary consequences for a self-duplicating nucleic acid molecule put under selection pressure for faster growth. Bacteriophage Qβ is an RNA phage with an RNA genome (~3500 nucletotides (nt)) that codes for four proteins: viral coat protein, attachment protein, maturation protein, and β1 replicase, also called Qβ-replicase, which is an RNA-dependent RNA polymerase. When Qβ-replicase is incubated with Qβ-RNA template in the presence of ribonucleotides, it synthesizes new Qβ-RNA molecules.

The goal of the experiment was to determine how molecules evolve if the selection pressure is allowed to only select for molecules that can multiply increasingly faster. The experimental procedure involved serial transfer of the reaction mix in which the incubation time was progressively reduced over time. The first

[e]Among living species, the fossil record of the modern-day horse from *Hyracotherium* (previously known as *Eohippus*) to *Equus*, spanning a period of about 55 million years, is one of the better-preserved fossil records that show macroevolutionary changes. Most fossil records are not as well preserved.

reaction was allowed to proceed for 20 minutes, after which an aliquot was used to start the second reaction, and so on for the first 13 reactions. After the first 13 reactions, the incubation periods were reduced to 15 min (transfers 14–29), 10 min (transfers 30–38), 7 min (transfers 39–52), and 5 min (transfers 53–74). The progressive reduction in the incubation intervals between transfers maintained the selection pressure for the evolution of the most rapidly multiplying RNA template molecules. As the experiment progressed, the rate of RNA synthesis increased and the product became smaller. By the 74th transfer, the size of the replicating molecule had become ~17% of its original size by deleting most of the original genome, and replicated 15 times faster than the complete viral RNA. This short RNA template variant was found to have experienced a significant change in base composition as well. The fact that this RNA template variant replicated 15 times faster than the complete viral RNA suggested that in addition to becoming smaller, the variant increased the efficiency with which it interacted with the replicase. Therefore, the RNA molecules adapted to the new conditions by throwing away anything not needed for fast replication[f].

It should be emphasized in this context that Spiegelman's experiment was a demonstration of **directed evolution** because selection pressure was applied to achieve a predetermined evolutionary outcome. The goal of Spiegelman's experiment as stated by Mills et al. was, "What will happen to the RNA molecules if the only demand made on them is the Biblical injunction, multiply, with the biological proviso that they do so as rapidly as possible?" In contrast, natural evolutionary processes are not directed. Genetic variations are random and spontaneous; hence they arise in the population independent of need. The advantages or disadvantages of such variations become apparent only when selection pressure arises. Thus, the natural evolutionary process works as a **blind watchmaker**, as Richard Dawkins calls it to underscore the lack of purpose and direction in the process. However, in recent years, the concept of directed (adaptive) mutation and directed evolution in bacteria, originally proposed in 1988 by John Cairns and coworkers,[6] has garnered some support. This idea is still not mainstream in evolutionary biology and is beyond the scope of this book.

Since the experiment of Spiegelman, many more extracellular Darwinian experiments have been conducted to direct the evolution of desired traits in biological macromolecules, and many laboratories have reported some remarkable findings.

2.3 MOLECULAR BASIS OF HERITABLE GENETIC VARIATIONS—THE RAW MATERIALS FOR EVOLUTION

Genetic variations in a population evolve irrespective of need. Most genetic variations are deleterious or at best neutral, but some may be beneficial in a specific environment. It is the selection pressure that reveals the utility of a beneficial genetic variation. Four important sources of molecular genetic variations are mutation, recombination, gene flow, and creation of new genes.

2.3.1 Molecular Basis of Mutation

Mutation is the change of genomic sequence. Mutation can be a **point mutation** (alteration of just one nucleotide), a **frameshift mutation** (alteration of the open reading frame (ORF) of the gene), or a **chromosomal mutation**—that is, large-scale alterations of the chromosomal DNA (**insertion**, **deletion**, **inversion**, **duplication**, **translocation**) (Figure 2.1A). Chromosomal mutations can result in gene duplication and divergence, exon shuffling, retrotransposition, gene fission/fusion, and gene deletion; each of these events creates genetic diversity.

Based on the effect on the polypeptide product, a point mutation can be missense, nonsense, or silent. A **missense point mutation** changes an amino acid in the polypeptide; a **nonsense point mutation** creates a stop codon, thereby prematurely truncating the ORF and ending translation of the polypeptide; a **silent point mutation** does not change the amino acid sequence of the polypeptide (Figure 2.1B). Splice donor or acceptor site mutations as well as splicing signal site mutations can result in the exonization of a previous intron sequence or intronization of a previous exon sequence; these types of mutations frequently have pathological consequences. There are a number of reports in the literature describing such mutations.

Based on the type of base altered, a point mutation can be classified as a **transition** or a **transversion** mutation. A pyrimidine replaced by another pyrimidine (C→T or T→C) or a purine replaced by another purine (A→G or G→A) is a transition mutation. A common mechanism of transition mutations is the formation of tautomeric forms (amino→imino tautomer as occurs in A and C; and keto→enol tautomer as occurs in G and T), and mispairing of bases (Figure 2.1C). If the mispairing survives the DNA repair machinery (e.g. if the mispairing occurs during replication), then by the following replication cycle the

[f]The small, rapidly duplicating RNA template variant was later termed the **Spiegelman monster**.

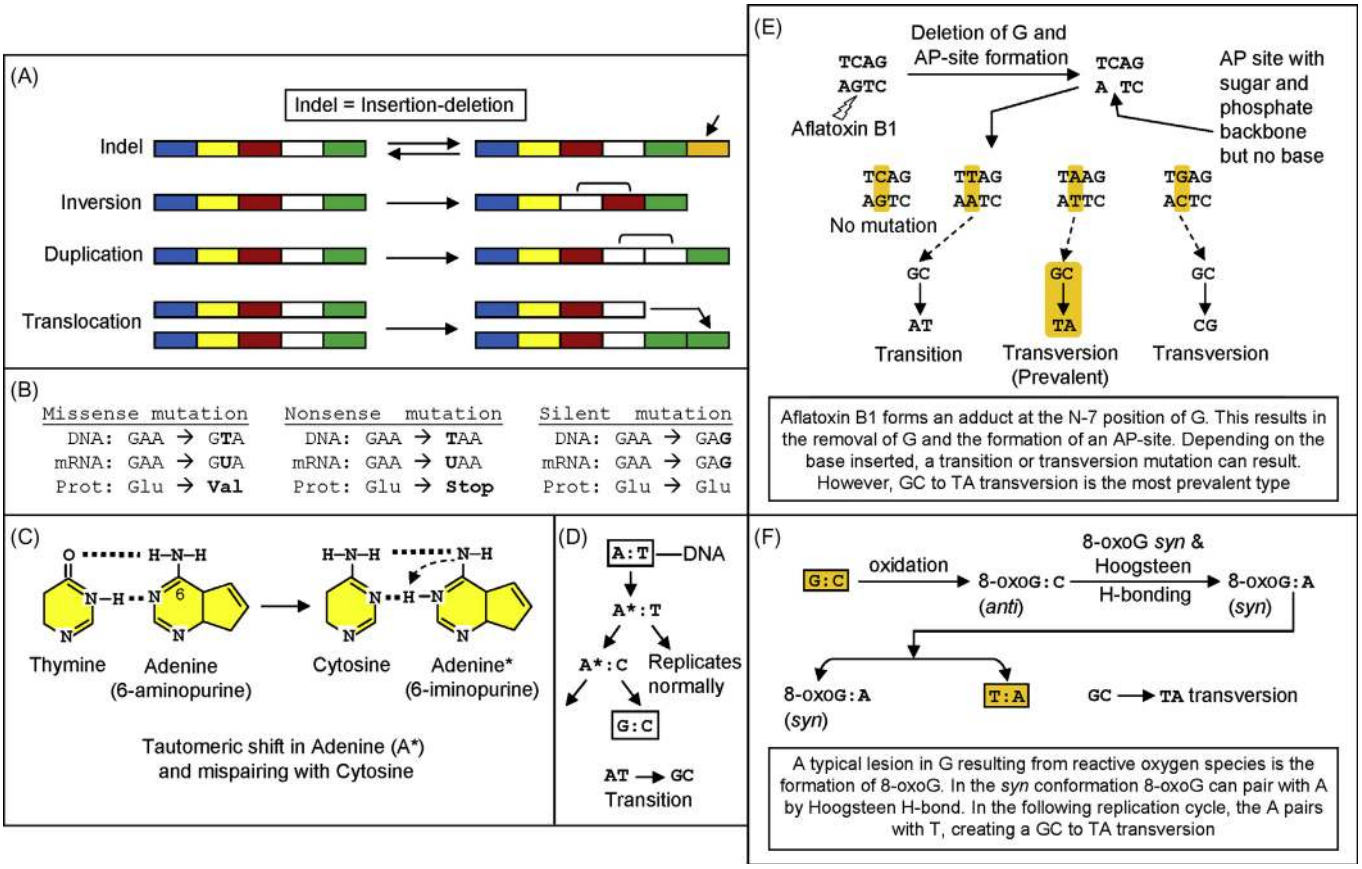


FIGURE 2.1 **Molecular basis of mutation.** (A) Various types of mutations affecting long DNA fragments, i.e. a chromosome. (B) Various effects of a one-base-pair mutation in DNA (only sense strand is shown). A missense mutation alters the amino acid sequence of a protein; a nonsense mutation disrupts the ORF and prematurely stops translation, whereas a silent mutation does not change the amino acid sequence of the protein. (C) Mechanism of transition mutation due to tautomeric shift in adenine resulting in 6-iminopurine from 6-aminopurine. (D) Wrong base pairing by imino tautomer of adenine results in AT-to-GC transition mutation in two replication cycles. (E) The mechanism of aflatoxin-B1-mediated transversion mutation (see text for details). (F) The mechanism of 8-oxoG-mediated transversion mutation (see text for details).

affected position of DNA has the base pair replaced by transition mutation (Figure 2.1D). Another mechanism of transition mutation in genomes is the spontaneous oxidative deamination of methylated C to form T, resulting in CG→TA transition over time. In contrast to transition mutation, a purine replaced by a pyrimidine or a pyrimidine replaced by a purine is a transversion mutation. Chemicals such as aflatoxin B1 can cause transversion mutation through adduct formation. Aflatoxin B1 forms an adduct at the N-7 position of guanine. This ultimately results in the removal of G and the formation of an AP-site (apurinic site). Depending on the base inserted for repair, a transition or transversion mutation can result. However, GC→TA transversion is the most prevalent type (Figure 2.1E).[7] Oxidation of guanine can also lead to transversion. A typical lesion in guanine resulting from oxidative stress is the formation of 8-oxoG. The 8-oxoG lesion in DNA is normally repaired by the dedicated enzyme 8-oxoG DNA glycosylase, which removes the oxoG with the concomitant cleavage of the DNA backbone. If the removal fails to take place, 8-oxoG tends to form the *syn* conformer, which then pairs with A by Hoogsteen H-bond during replication. In the following replication cycle, the A pairs with T, creating a GC→TA transversion (Figure 2.1F).[8] As mentioned above, *transition mutations are far more prevalent than transversion mutations*. In earlier literature, a point mutation was called a **single nucleotide polymorphism** (**SNP**) if it occurred in at least 1% of the population, but currently, any point mutation is regarded as an SNP. In the human genome, >65% of all SNPs are C→T transition mutations. SNPs and **copy number variations** (**CNVs**, also called **copy number polymorphisms or CNPs**) together constitute a significant source of inter-individual variation in a population.

In addition to the classical mutations described above, expansion or contraction of repeat sequences

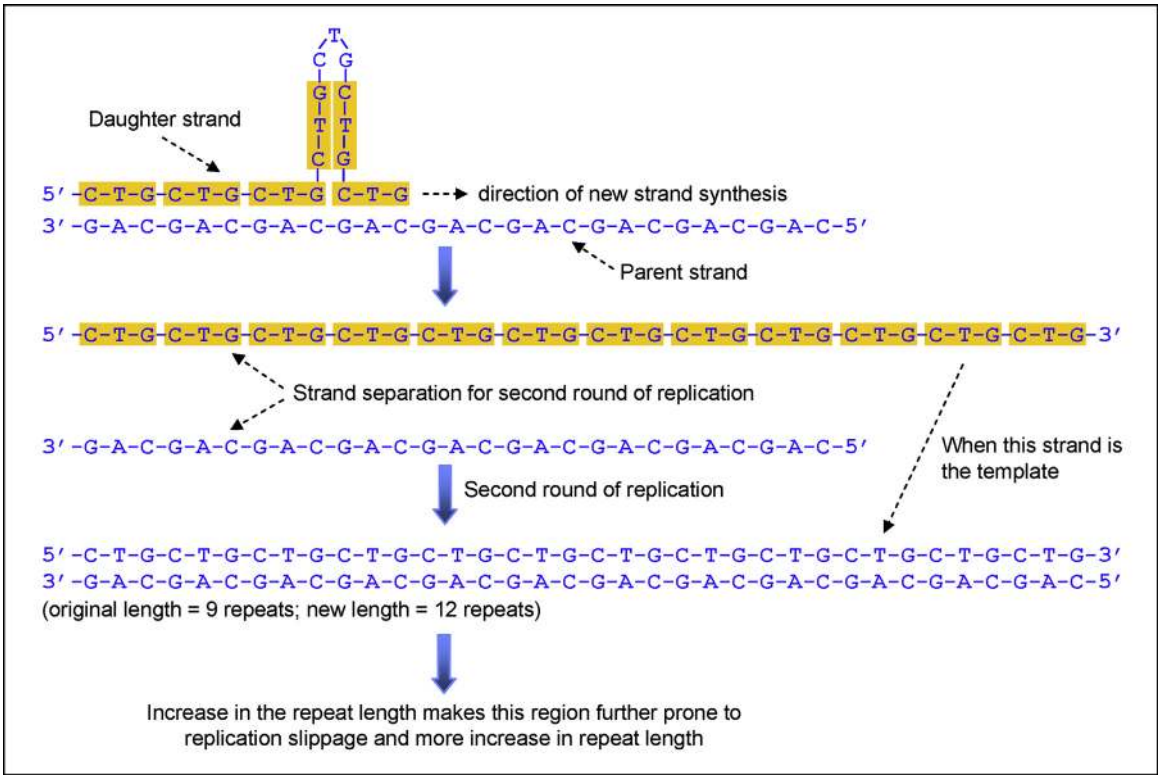


FIGURE 2.2 **Mechanism of expansion of triplet repeats through replication slippage.** The −C−T−G− triplet repeats in the gene are highlighted except the one forming loop. The increase in the number of repeats through replication slippage is a random process; it may be as few as one triplet or it may be multiple triplets. The figure shows an increase of three −C−T−G− triplet repeats in the gene in two rounds of replication. The strand of DNA containing the −C−T−G− triplets (highlighted) is the sense strand; therefore, the mRNA will have the same repeats as −C−U−G−.

constitutes another class of mutations. Repeat sequences in DNA can be expanded during replication. Two mechanisms can result in the expansion of repeat sequences: **replication slippage** (also called **slipped strand mispairing**) and **unequal crossing over**. In replication slippage, a long stretch of repeat sequences in the DNA folds back and pairs on itself, forming an internal hairpin or stem−loop structure, during replication. As a result, there is a net increase in the repeat sequences following replication in the daughter strand while the repeat length in the parent strand remains the same. The increased length of one strand propagates through subsequent rounds of replication (Figure 2.2). Misalignment of DNA involving blocks of the same repeat sequences may also occur during crossing over (unequal crossing over). As a result, in one chromosome the repeat length increases (insertion) while in the other chromosome it decreases (deletion), as shown in Figure 2.3.

The presence of uninterrupted trinucleotide repeats (triplet repeats) makes the sequence unstable and prone to further expansion through replication slippage. Increased numbers of triplet repeats are associated with a number of heritable genetic disorders in humans, such as Huntington's disease (CAG repeats), myotonic dystrophy (CTG repeats), fragile-X syndrome (CGG repeats). A higher number of uninterrupted triplet repeats is usually correlated with an earlier onset and a greater severity of the disease. *In contrast, interruption of the triplet repeats may reduce the predisposition of the carrier to the disease*. For example, fragile-X syndrome in humans is associated with the expansion of the CGG triplet repeats in the *FMR1* (fragile-X mental retardation 1) gene. However, if these CGG repeats are interspersed with AGG triplet repeats, the predisposition towards developing the disease is significantly reduced.[9] Populations that have a disproportionately large number of uninterrupted CGG-repeat-containing alleles, such as the Tunisian Jews, have a much higher incidence of fragile-X syndrome.[10]

Most mammals possess a small number of the CGG repeats in the *FMR1* gene (mean = 8 ± 0.8), but primates have a greater number of repeats (mean = 20 ± 2.3). Interestingly, nonhuman primates do not have fragile sites in the *FMR1* gene because they have many more interruptions in the CGG sequences.[11]

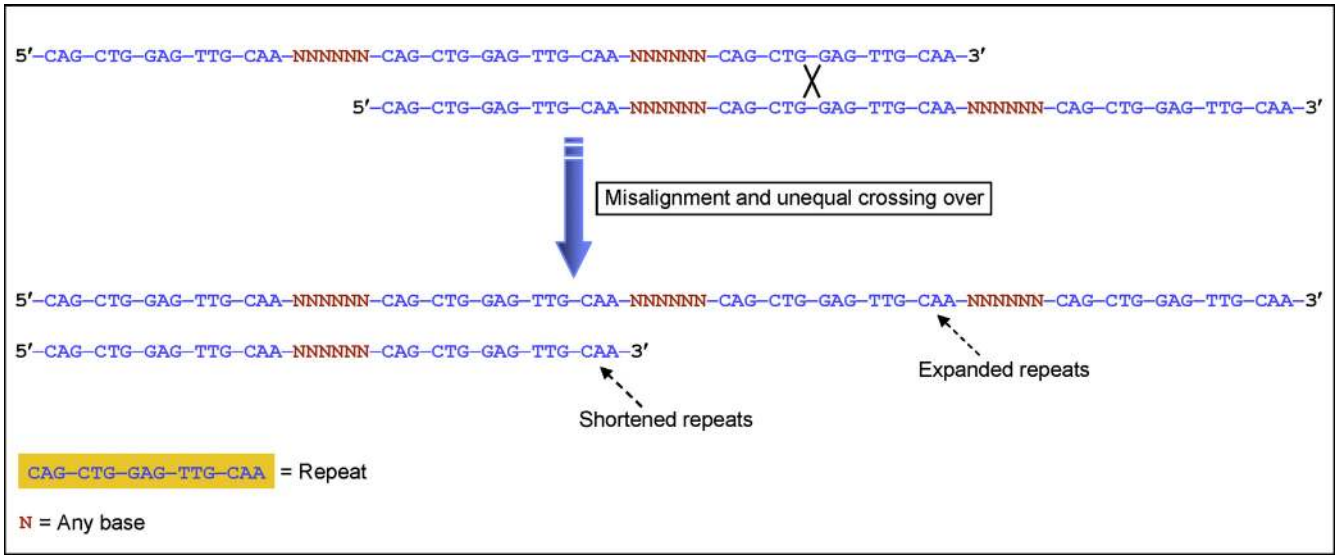


FIGURE 2.3 **Unequal crossing over altering the repeat length.** The block of repeat sequence used here as an example is –CAG–CTG–GAG–TTG–CAA–. The presence of blocks of the same repeat sequence makes the chromosomal misalignment and unequal crossing over possible.

2.3.2 Recombination and Generation of Genetic Diversity

In sexually reproducing organisms, meiotic recombination during gamete formation provides a means of creating genetic variation. In genetic recombination, a DNA segment moves from one DNA molecule to another DNA molecule. Recombination can take place between two homologous sequences or two nonhomologous sequences. Recombination between two homologous sequences is called **homologous recombination** and it occurs during meiosis between two homologous DNA molecules (homologous chromosomes) by crossing over. *The frequency of homologous recombination is low*. Recombination between two nonhomologous sequences can be mediated by **site-specific recombination**. Site-specific recombination occurs when two nonhomologous DNA molecules have only a small region of sequence identity; recombination occurs using this small region. *Recombination apparently depends on short stretches (could be as short as ~30 bp) of complete identity rather than long stretches of general similarity*.[12] Site-specific recombination helps in the integration of phage DNA into a bacterial chromosome; it can also help integrate transposable elements into the host DNA. Therefore, site-specific recombination provides a mechanism for introducing genetic diversity in the recipient genome.

Recombination between homologous chromosomes begins with double-strand breaks (DSBs). Because the non-sister chromatids of homologous chromosomes may not be identical in terms of their DNA sequence, mismatch repair synthesis during recombination may result in **gene conversion**. The mismatch repair enzyme corrects the sequence mismatch by partial resection of the broken DNA molecule followed by resynthesis of one of the strands using the corresponding DNA strand of the non-sister chromatid as the template. This results in a unidirectional transfer of the donor sequence to the acceptor sequence. It is easy to contemplate that if an allele is removed during resection, that allele is created during resynthesis based on the sequence of the allele of the donor strand. This phenomenon leads to gene conversion. Therefore, gene conversion involves nonreciprocal exchange of genetic material in which one sequence remains unchanged and the other sequence is altered.

Homologous recombination can also take place between two stretches of DNA that are not allelic. This is called **non-allelic homologous recombination** (**NAHR**). NAHR is driven by sequence identity, and it results in deletion in one chromosome and duplication in the other chromosome. Duplicated segments are predisposed to further NAHR. NAHR may lead to loss or increased copy number of specific genes, resulting in copy number variations (CNVs) of specific genes within the deleted or duplicated region. Such CNVs have major implications in health and disease as well as genome evolution. *In general, repeats provide hotspots of major structural alterations in the genome, ranging from microduplication and microdeletion to major segmental duplication and deletion, as well as repeat expansion and contraction*.

2.3.3 Gene Flow and Introduction of Genetic Diversity

Gene flow is also called **gene migration**. Gene flow is the transfer of genetic material from one population to another. Gene flow can take place between two populations of the same species through migration, and is mediated by reproduction and **vertical gene transfer** from parent to offspring. Alternatively, gene flow can take place between two different species through **horizontal gene transfer** (HGT, also known as **lateral gene transfer**), such as gene transfer from bacteria or viruses to a higher organism, or gene transfer from an endosymbiont to the host. HGT is discussed in detail later in this chapter. Gene flow within a population can increase the genetic variation of the population, whereas gene flow between genetically distant populations can reduce the genetic difference between the populations. Because gene flow can be facilitated by physical proximity of the populations, gene flow can be restricted by physical barriers separating the populations. Incompatible reproductive behaviors between the individuals of the populations also prevent gene flow.

2.3.4 Origin of New Genes, Creation of Genetic Diversity and Genome Evolution

Generation of new genes is an important mechanism for creating genetic novelties; hence, it is an important driving force of evolution in all organisms. New genes can be created by two major processes, (1) processes that use coding sequences (pre-existing genes) as the raw materials, and (2) processes that use noncoding sequences as the raw material.

2.3.4.1 Origin of New Genes from Coding Sequences (Pre-existing Genes)

These processes are better understood and include gene duplication, exon shuffling, gene fusion and fission, and lateral gene transfer.

2.3.4.1.A GENE DUPLICATION AND THE 2R HYPOTHESIS

Gene duplication creates paralogs. Susumu Ohno's seminal book *Evolution by Gene Duplication* (1970)[13] popularized the concept that gene duplication plays an important role in evolution. By comparing the genome size of different groups of non-vertebrate chordates and vertebrates, Ohno argued that the complexity of vertebrate genomes during evolution was achieved by whole-genome duplications in the lineage leading to vertebrates. Analysis of **orthologous genes** (**orthologs**[g]) showed that compared to urochordates (e.g. sea squirts), the genomes of jawless vertebrates, such as lamprey and hagfish, contain at least two orthologs and the genomes of mammals contain three or more orthologs. Ohno proposed that the ancestors of reptiles, birds, and mammals had experienced at least one tetraploid evolution either at the stage of fish or at the stage of amphibians. Since the turn of the millennium, the modern version of Ohno's hypothesis, known as the **two rounds (2R) hypothesis**, has resurfaced and gained popularity. There are disagreements regarding the stages of evolution when genome duplications took place. The most popular version of the 2R hypothesis proposes that one round of genome duplication took place at the root of the vertebrate lineage—that is, after the emergence of urochordates—followed by another around the time Agnatha (jawless vertebrates, e.g. lamprey and hagfish) and Gnathostomata (jawed vertebrates) split—that is, before the radiation of jawed vertebrates.[14–16] There are, however, debates about the 2R hypothesis, but that is beyond the scope of this section.

Ohno considered whole-genome duplication to be more important as an evolutionary mechanism than individual gene duplication, but gene duplication is now known to be a major mechanism for the creation of novel genetic material and an important driver of genome evolution. Genome sequencing shows that gene duplication is prevalent in all three domains of life (Bacteria, Archaea, Eukarya). In multicellular eukaryotes, including humans, ~40–60% genes have been produced through duplication, depending on the species. Several publications have reported on the rate of gene duplication in various eukaryotic species, but the results vary significantly. For example, based on observations from the genomic databases for several eukaryotic species, Lynch and Conery estimated that in eukaryotes the average rate of gene duplication is approximately 0.01 per gene per million years (i.e. the probability of duplication of a eukaryotic gene is at least 1% per million years[h,i]).[17,18] However, Cotton and Page estimated a gene duplication rate that is one order of magnitude lower than the estimate of Lynch and Conery.[19] Many duplicated genes are inactivated

[g]Orthologous genes or orthologs are homologs in different species—that is, they evolved from a common ancestral gene through speciation. Orthologs often retain the same or similar function(s).

[h]The duplication event per gene per million years was estimated to be 0.0023 for *Drosophila melanogaster*, 0.0083 for *Saccharomyces cerevisiae*, and 0.0208 for *Caenorhabditis elegans*, the average being ~0.01. So, it was the highest for *C. elegans*.

[i]The duplication event per gene per million years was estimated to be 0.009 for humans. In this publication, the rates calculated were slightly lower for *Drosophila*, yeast, and *C. elegans*, but the average was still ~0.01.

by accumulating degenerative mutations and become pseudogenes. Gene duplication can result from unequal crossing over, retrotransposon insertion, segmental duplication, and chromosomal (whole-genome) duplication.

If the rate of gene duplication is assumed to be somewhere in between the two estimates cited above, then it becomes close to the rate of fixed nucleotide substitutions, particularly in protein-coding genes. Using data from human and rodents, and assuming 80 million years as the time of divergence between the two lineages, the average fixed nucleotide substitution rate in protein-coding genes was calculated to be 0.74 per nonsynonymous site and 3.51 per synonymous site per billion (10^9) years.[20] However, such average estimates could still vary significantly in different species.

Unequal crossing over usually generates tandem duplication, which could involve the entire gene or part of a gene. Figure 2.3 shows duplication of a section of the gene through unequal crossing over. Duplication of the entire gene involves duplication of the introns as well as the regulatory sequences. The insertion of **processed** (**retrotransposed**) **pseudogenes** can also introduce genetic variability to the genome, particularly if the retrotransposed pseudogenes recruit new promoters and become functional. Some expressed pseudogenes regulate the mRNA expression of the normal gene. For example, *Makorin1-p1* in mice is a transcribed pseudogene, which regulates the expression of the normal gene *Makorin1*.[21] Pseudogenes are of two main types: (I) **duplicated** (**nonprocessed**) and (II) **retrotransposed** (**processed**). Duplicated pseudogenes arise from genomic DNA duplication or unequal crossing over. They retain the original exon−intron organization of the functional gene (hence nonprocessed), but their protein-coding potential is lost because of the loss of transcription regulatory elements, such as promoters or enhancers, or mutations disrupting the ORF, such as frameshifts or premature stop codons. In contrast, processed pseudogenes result from retrotransposition—that is, they arise from reverse transcription of mRNA into complementary DNA (cDNA) followed by the integration of the cDNA into the genome. As a result, processed pseudogenes lack introns and promoter, and they typically contain the poly(A) tail. Because they are retrotransposed, they are flanked by direct repeats. Processed pseudogenes are usually nonfunctional unless they are integrated under the influence of an active promoter, or recruit new promoters over time to become functional. Another type of pseudogene is known as the **unitary pseudogene**. A unitary pseudogene is a regular gene that has lost the protein-coding potential because of spontaneous mutation in the coding region; so it is neither duplicated nor retrotransposed. Because most pseudogenes are nonfunctional, they are not under selection pressure and are free to accumulate further mutations and increasingly diverge from the parent sequence from which they were derived. Pseudogenes have been identified in all known genomes, but their numbers greatly vary. For example, the estimated number of pseudogenes is 10,000−20,000 in humans, but only 110 in *Drosophila*.[22]

Human genome sequencing has revealed the widespread occurrence of **segmental duplications**, which often involve blocks of 1−200-kb (or longer) sequences that have been copied from one region of the genome and integrated into another region. Hence, segmental duplications create paralogous loci. The duplicated regions represent low-copy repeats and have >90% identity. Such strong sequence identity suggests that they are relatively recent in origin. The finished sequence of the human genome reported about 5.3% of the genome as segmental duplications.

Chromosomal (**whole-genome**) **duplication** is thought to arise by the breakdown of the normal mitotic or meiotic process. If chromosomes duplicate but do not separate (chromosomal non-disjunction) and are maintained in the same cell, a diploid gamete is produced. Fertilization of a diploid gamete by a normal haploid gamete would produce a triploid organism. The same mechanism can produce tetraploidy and even higher ploidy. In addition to the above mechanism of polyploidy, termed **autopolyploidy**, genome duplication and polyploidy can also be produced by hybridization of two related species that produce viable offspring. Such polyploidy is called **allopolyploidy**, and allopolyploids produce a diverse set of gametes. During evolution, whole-genome duplication resulting in polyploidy occurred frequently in plants but infrequently in animals.

The **evolutionary fate of duplicated genes** involves either acquiring new function or becoming nonfunctional. In most cases, the duplicated genes are free to acquire degenerative mutations and become pseudogenes (**pseudogenization**) because there are no functional constraints and the genes are not under selection pressure. Thus, pseudogenization is a neutral process. In order for the gene to escape pseudogenization and functional death, selection pressure must force the duplicated gene to drift towards fixation through **neofunctionalization**. Gene duplication followed by neofunctionalization of the duplicated gene provides an important mechanism for the genome to diverge both structurally and functionally. Neofunctionalization involves acquiring new function by the duplicated gene at the expense of the ancestral function—that is, the duplicated gene acquires a function that was not present in the ancestral gene. For example, the type III antifreeze protein (*AFPIII*) gene in the Antarctic zoarcid fish evolved from a sialic acid synthase (*SAS*) gene after duplication,

divergence, and neofunctionalization. The SAS is an old cytoplasmic enzyme present in microbes through vertebrates, whereas AFPIIIs are secreted plasma proteins that bind to invading ice crystals and arrest ice growth to prevent fish from freezing. The *SAS* gene possesses both sialic acid synthase and rudimentary ice-binding activities. Following duplication, the N-terminal SAS domain was deleted and replaced by a nascent signal peptide needed for the extracellular export of the mature protein. Further optimization of the C-terminal domain's ice-binding ability through amino acid changes led to the evolution of AFPIII as a neofunctionalized secreted protein capable of non-colligative freezing-point depression.[23] Another example is the retinoic acid receptor (*RAR*) gene. Mammals have three *RAR* paralogs—*RAR*α, β, and γ—created by genome duplications at the time of origin of vertebrates. Using pharmacological ligands selective for specific paralogs, it was demonstrated that RARβ kept the ancestral RAR role, whereas RARα and RARγ diverged both in ligand-binding capacity and in expression patterns. Therefore, neofunctionalization occurred at both the expression and the functional levels to shape RAR roles during development in vertebrates.[24] Many other examples of neofunctionalization have been reported in the literature.

Neofunctionalization does not always have to arise following gene duplication. A beneficial mutation of the wild-type gene may create a mutant allele with new function. If the beneficial mutant allele is maintained by balancing selection, the carrier (heterozygote) will have increased fitness. If the beneficial mutant allele becomes the source of the duplicated gene, then the duplicated gene will be quickly fixed in the population by positive selection.[25]

Another functional outcome of gene duplication and divergence is **subfunctionalization**. Like pseudogenization, subfunctionalization is also a neutral process. Subfunctionalization occurs when the duplicated copies (paralogs) partition the attributes of the ancestral gene, such as function and/or expression. Following a duplication event, both paralogs experience a period of relaxed selection and accelerated evolution. This is because natural selection does not distinguish which paralog should be under selection and which paralog should be free from selective constraint. Thus, both genes might accumulate mutations that impair ancestral gene function. Under this condition, each paralog may retain one part of the function (subfunction) of the ancestral gene. Alternatively, each individual paralog may lose its ability to substitute for the ancestral gene function, but together the two paralogs may still be able to complement each other in producing ancestral gene function. Subfunctionalization has been proposed as an alternative mechanism driving duplicate gene retention in organisms with small effective population sizes.[26] A model to explain the high retention of duplicated genes through subfunctionalization was provided early on by the **duplication−degeneration−complementation (DDC) model**.[27] According to the DDC model, originally proposed in the context of *cis*-regulatory elements, subfunctionalization is driven entirely by degenerative mutations. Degenerative changes occur in regulatory sequences of both duplicated copies such that the expression pattern of the original gene can only be achieved when the two duplicated genes can complement each other. Therefore, degenerative mutations in the regulatory elements may increase the chance of duplicate gene retention. An implication of the DDC model is that the paralogs can not accumulate same inactivating mutations that would interfere with their ability of complementation. A number of examples of subfunctionalization have been reported in the literature. A common example is the normal human hemoglobin, which is composed of two α-chains and two β-chains (α2β2) encoded by α-globin and β-globin genes, respectively. The α- and β-globin genes are products of gene duplication and subsequent subfunctionalization because they complement each other in producing normal functional hemoglobin.[28] An example of subfunctionalization in terms of differential expression of paralogs is that of the *pax6a* and *pax6b* genes in zebrafish; these paralogs arose following a whole-genome duplication event about 350 million years ago. The expression patterns of *pax6a* and *pax6b* have diverged from each other since the duplication event. Whereas *pax6a* is widely expressed in the brain compared to *pax6b*, only *pax6b* is expressed in the developing pancreas. Such differential expression of *pax6b* in brain and pancreas is due to the loss of a brain-specific downstream regulatory element but gain of an upstream pancreas enhancer element.[29] An example of subfunctionalization has also been reported in Archaea. When Tocchini-Valentini and coworkers searched the genome of *Sulfolobus solfataricus* (Archaea; Crenarchaeota) for **homologs**[j] of *Methanocaldococcus jannaschii* (Archaea; Euryarchaeota) tRNA endonuclease, they found two paralogs of the tRNA endonuclease gene of *M. jannaschii* in the genome of the *S. solfataricus*. Characterization of these two paralogous gene products revealed that both are required for tRNA endonuclease activity, each complementing the other for complete

[j]Homologous genes, or homologs, are related to each other by descent from a common ancestral gene. Homologs may or may not have the same or similar function. Therefore, the orthologs and paralogs described above are two different types of homologous genes.

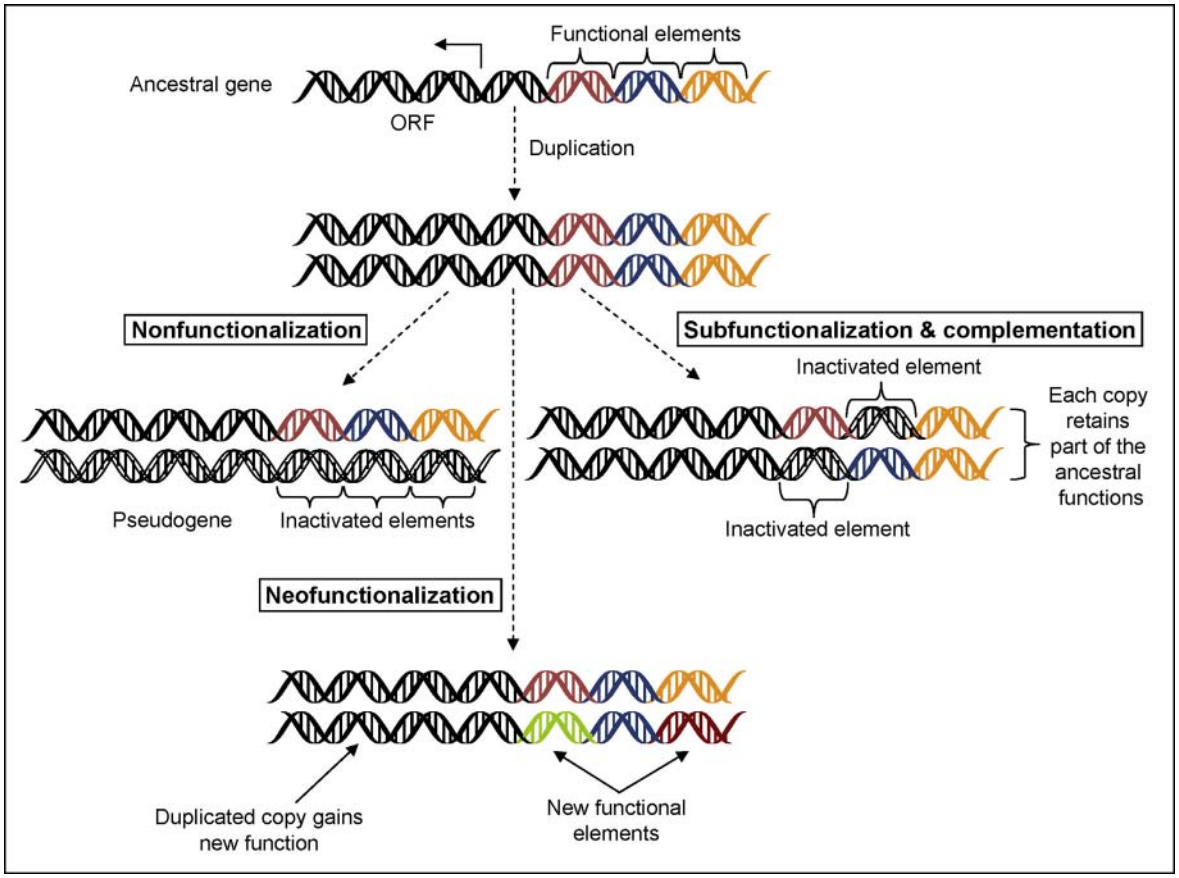


FIGURE 2.4 **Three possible fates of duplicated genes: pseudogenization (nonfunctionalization), neofunctionalization, and subfunctionalization using *cis*-regulatory modules as targets of divergence.** Duplicated genes are not under selection pressure; hence, there are no functional constraints and a duplicated gene is free to acquire degenerative mutations and become a pseudogene. Sometimes, the acquisition of new function by the duplicated gene (neofunctionalization) provides an important mechanism for the genome to diverge both structurally and functionally. The newly acquired function is not present in the ancestral gene. Subfunctionalization occurs when the duplicated copies (paralogs) partition the attributes of the ancestral gene, such as function and/or expression. The figure shows that degenerative changes occurred in regulatory sequences of both paralogs such that the expression pattern of the original gene can only be achieved when the two duplicated genes complement each other (see text for examples).

activity. Detailed analysis of the amino acid sequences of the two proteins demonstrated that these two sequences had evolved by duplication of the ancestral sequence followed by divergence and subfunctionalization of the sequences.[30] Figure 2.4 shows the three fates of duplicated genes discussed here (pseudogenization, neofunctionalization, subfunctionalization) using *cis*-regulatory modules as targets of divergence.

2.3.4.1.B EXON SHUFFLING

The natural process of creating new combinations of exons by intronic recombination is called exon shuffling.[31] Following the discovery of introns, Walter Gilbert suggested that the presence of introns allowed exon shuffling, which resulted in genomes being more complex and diversified. Exon shuffling is largely responsible for protein-domain shuffling.[32] The diversity of protein-domain combinations increased with the evolution of organismal complexity. However, most protein domains are ancestral; only few new domains have been invented in the vertebrate lineage. For example, about 7% of the protein families in human genome seem to be specific to vertebrates. The majority of the proteins necessary for the maintenance of basic cellular functions evolved early. Hence, the evolution of proteome complexity was driven by the reshuffling of pre-existing components into a richer collection of domain architectures.[33] Therefore, protein-domain shuffling, which refers to the duplication of a domain or the insertion of a domain from one gene into another, has been a major factor in the evolution of human phenotypic complexity. Kaessmann et al.[34] systematically analyzed intron phase distributions in the coding sequence of human protein domains to identify signatures of exon shuffling resulting in domain shuffling. Introns of symmetrical phase combinations

(i.e. 0−0, 1−1, and 2−2[k]) were found to be predominant at the boundaries of domains, whereas non-boundary introns showed no excess symmetry, suggesting that exon shuffling primarily involved rearrangement of structural and functional domains. Domains flanked by phase 1 introns (i.e. 1−1 symmetrical domains) were found to have dramatically expanded in the human genome due to domain shuffling. The observation of predominance and extracellular location of 1−1 symmetrical domains among metazoan protein-specific domains suggested an association with the evolution of multicellularity. In contrast, 0−0 symmetrical domains were found mostly overrepresented among ancient protein domains that are shared between the eukaryotic and prokaryotic kingdoms. Franca et al.[35] investigated the intron phase distribution in 10 genomes to generate a catalog of putative exon shuffling events in several eukaryotic species, including non-metazoans (choanoflagellate *Monosiga brevicollis*), early branching metazoans (the sea anemone *Nematostella vectensis*), the smallest chordate (urochordate *Ciona intestinalis*), and representative species from all vertebrate lineages except reptiles (zebrafish, *Xenopus*, chicken, mouse, and human). They confirmed previous observations that exon shuffling mediated by phase 1 introns (1−1 exon shuffling) is the predominant kind in multicellular animals, whereas exon shuffling mediated by phase 0 introns (0−0 exon shuffling) is the predominant type in non-metazoan species. They also concluded that such a pattern was achieved since the early steps of animal evolution.

Intronic recombination generating exon shuffling was most likely facilitated by two important events at a later stage during the evolution of eukaryotes: the emergence of spliceosomal introns, and the insertion of repetitive sequences within spliceosomal introns.[36] Although the presence of repetitive sequences in introns could facilitate intron recombination, insertion of repetitive sequences in self-splicing introns would not have been tolerated because self-splicing introns encode an essential function. In contrast, insertion of repetitive sequences would have been tolerated in spliceosomal introns because of the lack of such functional constraints. Hence, recombination involving self-splicing introns early in life's evolution could not have played an important role in exon shuffling, and consequently in the evolution of ancient proteins. Exon shuffling most likely increased in parallel with the evolution and expansion of spliceosomal introns and the concomitant appearance of less compact genomes.

Patthy analyzed the evolutionary distribution of some proteins that could be identified as modular proteins (containing specific functional modules) and seemingly evolved by intronic recombination. His analysis revealed that modular multidomain proteins produced by exon shuffling are restricted in their evolutionary distribution[l]. The majority of these proteins are functionally linked to the evolution of multicellularity of animals, such as constituents of the extracellular matrix, proteases involved in tissue remodeling, various proteins of body fluids, and proteins associated with cell−cell and cell−matrix interactions. Some examples include selectins, interleukin-2 receptor, cartilage link protein, follistatin, C-type lectin, and tolloid. The results suggest that exon shuffling acquired major significance at the time of metazoan radiation.

2.3.4.1.C GENE FUSION AND FISSION

During evolution, many complex proteins were apparently produced by gene fusion and less complex proteins by gene fission. Gene fusion results in the creation of a composite protein. In contrast, gene fission results in the creation of two or more smaller, split proteins. For example, the basic biochemistry of fatty acid synthesis is very similar from *E. coli* to mammals. However, the six enzymes and the acyl carrier protein involved in fatty acid synthesis exist as independent polypeptides in *E. coli*, whereas in mammals these exist as one composite polypeptide containing all the activities because of the fusion of genes encoding them.

Snel and coworkers[37] analyzed all ORFs of 17 completely sequenced bacterial genomes using the Smith−Waterman sequence comparison algorithm; the analysis showed evidence for numerous cases of gene fusion and fission. In general, they observed that

[k]As mentioned in Chapter 1, introns can be divided into three types based on phases: phase 0, phase 1, and phase 2. A phase 0 intron does not disrupt a codon, a phase 1 intron disrupts a codon between the first and the second bases, and a phase 2 intron disrupts a codon between the second and third bases. An exon flanked by two introns of the same phase (e.g. 0−0, 1−1, 2−2) is called a symmetrical exon, whereas an exon flanked by two introns of different phases (e.g. 0−1, 1−2, 2−0, etc.) is called an asymmetrical exon. Legitimate alternative splicing involves the removal of a symmetrical exon. In contrast, alternative splicing involving an asymmetrical exon results in a change of the ORF downstream of the 3′-splice site (Figure 1.5), but this is very rare.

[l]In the analysis, protein modules were considered to be generated through exon shuffling if: (1) the modules were homologous (i.e. modules derived from a common ancestor) but present in otherwise nonhomologous proteins, and (2) the transposition of the module was mediated by exon shuffling through intronic recombination. Evidence of exon shuffling through intronic recombination was considered if the module was flanked by introns of same phase. Thus, the introns of these modular proteins were shown to have a marked intron-phase bias.

fusion occurred more often than fission. Using the same approach (sequence-based comparison) Enright and Ouzounis[38] identified 7224 components and 2365 composite unique proteins across the 24 species considered in the study. These 24 genomes included those of bacteria and eukaryotes, including *Drosophila melanogaster* and *Caenorhabditis elegans*. They found a number of functional associations. For example, MXR1 (peptide methionine sulfoxide reductase, involved in antioxidative processes) and YCL033C (function unknown) were predicted to be functionally associated by virtue of gene fusion in three species—*Helicobacter pylori*, *Haemophilus influenzae*, and *Treponema pallidum*—and this observation was supported by experimental results. Likewise, Yanai et al.[39] identified groups of closely related proteins that have undergone fusion or fission. For example, the genes for glycolytic enzymes triosephosphate isomerase (TPIA), phosphoglycerate kinase (PGK), and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in the parasitic bacterium *Mycoplasma genitalium*, are linked by fusion events in other species, such as TPIA + PGK in *Thermotoga maritima* and TPIA + GAPDH in *Phytophthora infestans*.

Using domain architecture comparison, Kummerfeld et al.[40] performed a comprehensive analysis of divergent sequences in distantly related organisms to identify evidence of gene fusion and fission during evolution. The authors considered proteins at the level of domain architecture because structural domains reveal more about distant evolutionary relationships than simple sequence alignment. The domain information was collected from the Structural Classification of Proteins (SCOP) database, which provides an evolutionary definition of domains based on three-dimensional structure. The authors studied proteins across 131 genomes (17 Archaea, 98 Bacteria, and 16 Eukarya), and investigated 7116 domain architectures to identify protein domains that evolved by fusion or fission. In order to do that, the authors looked for domain architectures that were present as a single protein (i.e. the composite form) in at least one genome, and as a set of shorter proteins (i.e. the split forms) in other genomes, which would suggest that the composite protein was split by fission or the split proteins were fused at some stage during evolution. The authors identified 2869 groups of multidomain proteins as a single protein in certain organisms and as two or more smaller proteins with equivalent domain architectures in other organisms. They also found that fusion events were approximately four times more common than fission events, which is consistent with the observation by Snel et al. The authors discussed the possible contribution of horizontal gene transfer in the evolution of composite proteins, which is more prevalent in Bacteria and Archaea.

2.3.4.1.D HORIZONTAL GENE TRANSFER

Horizontal gene transfer, also known as **lateral gene transfer**, refers to nonsexual transmission of genetic material between unrelated genomes; hence, horizontal gene transfer involves gene transfer across species boundaries. The phenomenon of horizontal gene transfer throws a wrench in the concepts of last common ancestor, syntenic relationship between genomes, phylogeny and the evolution of discrete species units, taxonomic nomenclature, etc.[m] The majority of examples of horizontal gene transfer are known in prokaryotes. In bacteria, three principal mechanisms can mediate horizontal gene transfer: **transformation** (uptake of free DNA), **conjugation** (plasmid-mediated transfer), and **transduction** (phage-mediated transfer). In plants, **introgression** can mediate horizontal gene transfer; this means gene flow from one gene pool to another gene pool—that is, from one species to another species by repeated backcrossing between an interspecific hybrid and one of its parent species. Therefore, introgression depends on the extent of reproductive isolation between the two species. Introgression has also been reported between duck species, between butterfly species involved in mimicry, and between human and Neanderthal.[41]

Horizontal gene transfer in animals is not common, but there are some reports. For example, Acuña et al.[42] identified the gene *HhMAN1* from the coffee berry borer beetle, *Hypothenemus hampei*, which shows clear evidence of horizontal gene transfer from bacteria. *HhMAN1* encodes the enzyme mannanase, which hydrolyzes galactomannan. Phylogenetic analyses of the mannanase from both prokaryotes and eukaryotes revealed that mannanases from plants, fungi, and animals formed a distinct eukaryotic clade, but *HhMAN1* was most closely related to prokaryotic mannanases, grouping with the *Bacillus* clade. *HhMAN1* was not detected in the closely related species *H. obscurus*, which does not colonize coffee beans. The authors hypothesized that the acquisition of the *HhMAN1* gene from bacteria was likely an adaptation in response to need in a specific ecological niche.

[m]During evolution, different lineages split from a common ancestor (the last common ancestor of those lineages) and evolve to ultimately form reproductively isolated groups (species). However, lineages descending from a common ancestor still maintain many ancestral genes in groups and in the same order but scattered in different chromosomes (syntenic relationship between genomes). This scenario of evolution does not consider the possibility of exchange of genetic material between groups belonging to different lineages. The phenomenon of horizontal gene transfer is an exception to this paradigm.

There are also some examples of horizontal gene transfer from fungi to arthropods, such as aphids (insects) and mites (arachnids). Phylogenetic analysis revealed the evidence of horizontal transfer of genes encoding carotenoid desaturase and carotenoid cyclase–carotenoid synthase from fungi to pea aphid,[43] and to spider mite.[44] Notably, the fused carotenoid cyclase–carotenoid synthase gene is characteristic of fungi but not of plants or bacteria. The authors discussed the possible mechanism of such gene transfer. Gene transfer into a single arthropod ancestor of both spider mites and aphids is not likely because it would require subsequent loss of these genes in most other living arthropod taxa. The most likely scenario is the transfer of these genes through symbiosis, which probably occurred independently in both aphids and spider mites. It has been suggested that the frequent association of mites with viruses makes them ideal horizontal gene transfer vectors, including incorporation of mobile genes into their own genomes.

2.3.4.2 *Origin (de Novo) of New Genes from Noncoding Sequences*

The processes of how a new gene is created de novo from noncoding sequence are not well understood. For a noncoding DNA to give birth to a protein-coding gene, two features are needed: the DNA must be transcription-competent, and the DNA must acquire an open reading frame. It is being increasingly appreciated that a rare but consistent feature of eukaryotic genomes is the evolution of new genes de novo. Every genome contains genes that lack homologs in other taxonomic lineages. These new genes are called **orphan genes**. Orphan genes may arise by duplication and rearrangement followed by rapid divergence, but their de novo origin from noncoding DNA appears to be a very important mechanism.[45] If orphan genes are born through a duplication–divergence mechanism, they have to diverge beyond recognition as paralogs. In contrast, the de novo origin of orphan genes from noncoding DNA requires the emergence of sequence features forming functional signals, such as transcription initiation signal, polyadenylation signal, splice signal, etc., and finally the sequence would have to come under regulatory control in order for the gene to be expressed. Further accumulation of additional regulatory elements can expand the tissue expression pattern of a newly evolved orphan gene. *One characteristic of genes originated de novo is that these genes are usually simple (mostly single exon) so that their evolution de novo would be possible.*

In recent years, following the sequencing of many genomes, there have been multiple reports of identification of genes born de novo from noncoding DNA. Begun and coworkers,[46,47] reported de novo origin of orphan genes from noncoding DNA in *Drosophila*. By comparing the genome sequences of various species of *Drosophila*, Levine et al. described five novel genes in *D. melanogaster* that were derived from noncoding DNA. These genes have no homologs in any other species. Begun et al. subsequently used testis-derived expressed sequence tags (ESTs) from *D. yakuba* to identify genes that have likely arisen either in *D. yakuba* or in the *D. yakuba*/*D. erecta* ancestor. They identified eleven such genes. The genes described in these two publications are mostly X-linked, expressed in the testis, and have male germ-line functions. Zhou et al.[48] identified nine genes that originated de novo, and estimated that about 12% of the new genes that originated in the *Drosophila* lineage had arisen de novo. In recent years, efforts have turned to the human genome in order to find genes that most likely originated de novo. By building blocks of conserved synteny between human and chimpanzee genome and using 1:1 orthologs identified as BLASTP hits (hits in the protein database using Basic Local Alignment Search Tool (BLAST)) with no other similarly strong hits, Knowles and McLysagh reported three human protein-coding genes—*CLLU1*, *C22orf45*, and *DNAH10OS*—that seemingly had de novo origin in the human genome. Each of these three genes is a single-exon gene; however, they do contain introns in the untranslated regions. In order to minimize the chance that the genes could be annotation artifact, the authors only considered human genes that are classified as "known" by Ensembl and that have expressed sequence tag (EST) support for transcription.[49] Another de novo protein-coding gene, *C20orf203*, which is associated with brain function in humans, was reported in 2010.[50]

More recently, the identification of the most extensive set of human genes born de novo from noncoding DNA was reported by Wu et al.[51] Using a similar approach as that of Knowles and McLysaght, they reported 60 new protein-coding genes that apparently originated de novo in the human lineage since its divergence from the chimpanzee. Their data are supported by both transcriptional and proteomic evidence. Using RNA sequencing, the highest expressions of these genes were found to be in the cerebral cortex and testes, suggesting that these genes may contribute to phenotypic traits that are unique to humans, including the development of cognitive ability. Interestingly, the earlier finding of Knowles and McLysagh on the three human genes identified as having a de novo origin (*CLLU1*, *C22orf45*, and *DNAH10OS*) was not supported by the findings of Wu et al. The discrepancy was due to changes in gene annotation in the different versions of the databases used by these two groups (version 46 used by Knowles and McLysaght versus version 56 used by Wu et al.). This discrepancy also underscores the fundamental challenge of identifying

genes of de novo origin accurately based on annotated genome. A major challenge remains to demonstrate the functionality of these genes.

Exonization of previous intron sequences through mutation and abolition of splice sites is another mechanism of increasing the proportion of coding sequences derived from noncoding sequences in the genome. Examples include exonization of intronic *Alu* sequences,[52,53] and of intronic sequences in the collagen IV gene.[54] However, exonization of introns may also be associated with pathological outcomes.[55,56]

2.4 FACTORS THAT AFFECT GENE FREQUENCY IN A POPULATION

The mechanism of molecular evolution also involves the accumulation of genetic diversity, which leads to changes in gene frequency and genetic structure of the population. Changes in allele frequency initially result in microevolution, which introduces genetic variations in a population through processes such as mutation, migration, selection, genetic drift, population bottlenecks, and even relaxation of purifying selection.

A simple model for calculating gene frequency in a diploid population is provided by the **Hardy–Weinberg equilibrium** principle (see Box 2.1). It states that *the gene frequency in a diploid population remains constant through generations provided five conditions are met: no mutation, no migration, no selection, no genetic drift, and panmixis (random mating)*. For example, two alleles A_1 and A_2 can produce three possible genotypes: A_1A_1, A_1A_2, and A_2A_2. According to the Hardy–Weinberg principle, if the frequency of A_1 is p, and the frequency of A_2 is q ($q = 1 - p$, because $p + q = 1$, i.e. 100%), then the frequencies of A_1A_1, A_1A_2, and A_2A_2 are p^2, $2pq$, and q^2, respectively, and $p^2 + 2pq + q^2$ will also be 1 (i.e. 100%). A population in which the genotypic ratios are maintained is said to be in Hardy–Weinberg equilibrium.

BOX 2.1

Hardy–Weinberg Equilibrium at a Single Locus with Two Alleles

		Sperm	
		A_1 (p)	A_2 (q)
Egg	A_1 (p)	A_1A_1 (p^2)	A_1A_2 (pq)
	A_2 (q)	A_1A_2 (pq)	A_2A_2 (q^2)

Hence, the frequencies are: $A_1A_1 = p^2$, $A_1A_2 = 2pq$, $A_2A_2 = q^2$.

The sum of the frequencies of alleles as well as the genotypes is always 1.

Hence, for the alleles, $p + q = 1$ (=100%), and for the genotype, $(p + q)^2 = 1$, or $p^2 + 2pq + q^2 = 1$ (=100%).

Example: If the frequency of $A_1 = 0.7$ and the frequency of $A_2 = 0.3$ (=1 − 0.7), then the frequencies of the genotypes in the population are as follows:

$A_1A_1 = (0.7)^2 = 0.49 = 49\%$;

$A_1A_2 = 2(0.7)(0.3) = 0.42 = 42\%$;

$A_2A_2 = (0.3)^2 = 0.09 = 9\%$.

Hardy–Weinberg Equilibrium at a Single Locus with Three or More Alleles (Multiple Alleles)

If the locus under study has three or more alleles (multiple alleles), the derivation of frequencies is similar to that used for two alleles. If the alleles are A_1, A_2, and A_3, and the frequencies are, p, q, and r respectively, then:

The gene frequency p (A_1) + q (A_2) + r (A_3) = 1.

The genotype frequency $(p + q + r)^2 = 1$, or

p^2 (A_1A_1) + q^2 (A_2A_2) + r^2 (A_3A_3) + $2pq$ (A_1A_2) + $2pr$ (A_1A_3) + $2qr$ (A_2A_3) = 1.

Hardy–Weinberg Equilibrium at Two or More Loci

Let's assume, at one locus, the alleles are A_1 and A_2 and their frequencies are p and q, respectively.

At a separate, independently assorting locus, the alleles are B_1 and B_2, and their frequencies are r and s, respectively. Hence, $p + q = 1$, and $r + s = 1$.

The four types of allelic combinations in the gametes are: A_1B_1, A_1B_2, A_2B_1, and A_2B_2; their frequencies will be pr, ps, qr, and qs, respectively, and $pr + ps + qr + qs = 1$.

If all the alleles are at equilibrium, then the genotype frequencies will be $(pr + ps + qr + qs)^2$. The genotype frequencies of offspring can also be easily calculated using the Punnett square; for example, a cross $A_1A_2B_1B_2 \times A_1A_2B_1B_2$ will yield p^2r^2 $A_1A_1B_1B_1$; $2p^2rs$ $A_1A_1B_1B_2$; $2pqr^2$ $A_1A_2B_1B_1$; ... q^2s^2 $A_2A_2B_2B_2$.

The Hardy–Weinberg equilibrium principle is a very simplistic representation of the maintenance of gene frequencies in a population, and it does not take into account most of the complexities associated with actual populations. The conditions that need to be met for a population to remain in Hardy–Weinberg equilibrium also underscore the conditions that can introduce genetic variations in a population and cause microevolution, as discussed below.

2.4.1 Mutation

Genetic variation in a population is derived from a wide assortment of different alleles. Mutation or change in the genetic material is one of the primary sources of generation of genetic diversity in the population. As discussed above, a mutation can be a point mutation, a change in the open reading frame of a gene, or a chromosomal mutation. Chromosomal mutations are large-scale changes in chromosomal structure and organization, exemplified by insertion–deletion (indel), inversion, duplication, and translocation (Figure 2.1A).

The spontaneous point mutation rate (see Box 2.2) varies depending on the gene and the species. The mutation rate can be expressed differently. Studies utilizing breeding of control mice and monitoring mutations in five coat-color loci demonstrated an average mutation rate of $\sim 12 \times 10^{-6}$ per locus per gamete for forward mutations from the wild type, and $\sim 2 \times 10^{-6}$ per locus per gamete for reverse mutations from recessive alleles.[57,58] Mouse mutation data summarized from different radiation experiments showed a forward mutation rate of 6.6×10^{-6} per locus per generation.[59] The average forward mutation rate of the hypoxanthine phosphoribosyltransferase (*HPRT*) gene of the human promyelocytic leukemia cell line HL-60 was reported to be $\sim 2\text{–}6 \times 10^{-7}$/cell/generation.[60] When the mutation rate is calculated based on the evolution of pseudogenes, it turns out to be one or two orders of magnitude higher. This is expected because pseudogenes are mostly free from selective constraints. For example, the mutation rate based on the evolution of pseudogenes in humans was estimated to be $\sim 2 \times 10^{-8}$ per base per generation.[61] However, a different estimate, based on determining the substitution rate in pseudogenes, calculated the average mutation rate in mammalian nuclear DNA to be $3\text{–}5 \times 10^{-9}$ nucleotide substitutions per nucleotide site per year.[62]

Therefore, changes in allele frequency due to mutations alone are very small. Nevertheless, for a large population, the cumulative effect of mutation over many generations can be significant. Recently, it was demonstrated that natural genetic variations in the human genome are caused by small insertions and deletions.[63] The authors reported almost 2 million small insertions and deletions (indels) ranging from 1 to 10,000 bp in length in the genomes of 79 diverse humans. These variants include 819,363 small indels that map to human genes. Small indels were frequently found in the coding exons of these genes, and several lines of evidence indicate that such variations are a major determinant of human biological diversity.

BOX 2.2

ESTIMATION OF MUTATION RATE

The mutation rate in haploid organisms can be directly measured because the mutation will be expressed and the mutant phenotype can be observed.

Determination of the mutation rate in diploid organisms is more challenging because a recessive mutation can be masked by the dominant allele. Hence, the expression of the mutant phenotype and the actual occurrence of the mutation can be separated by many generations. Some major contributions on the estimation of mutation rate in mammals were made by a number of different groups from the 1950s to the 1970s. The contributions of Gunther Schlager and Margaret Dickie (cited above) of the Jackson Laboratory, Bar Harbor, Maine, are worth mentioning simply because of the volume of the work they did. They analyzed in excess of 7 million mice over many years for five coat-color loci (*nonagouti*, *brown*, *albino*, *dilute*, *leaden*) for estimating the average mutation rate.

For direct estimation, as done by Schlager and Dickie, the mutation rate in a single generation is used. In this scenario, the parental genotypes are known. If the offspring shows a mutant phenotype, it is backcrossed with the parents, and also crossed with a mouse homozygous for that mutation, and with a mouse that does not carry the mutation, in order to confirm the mutation. The mutation rate is calculated as follows:

$$\mu = x/2N,$$

where μ = mutation rate, x = number of mutant offspring, and N = total number of offspring examined. The factor 2 is used because each offspring develops from fertilization involving two haploid gametes. Each haploid gamete contains one allele that can potentially be the mutant allele. Therefore, the mutation rate calculated this way is expressed as "per locus per gamete." When using cell

BOX 2.2 (*cont'd*)

culture, the mutation rate can also be expressed "per cell division."

Example: If eight offspring are born with a mutant phenotype out of 1 million (10^6) progeny, and if three of those offspring had affected parents, then five offspring were born with the new mutation. Therefore, the mutation rate will be $5/(2 \times 10^6) = 2.5 \times 10^{-6}$ per locus per gamete.

Because an accurate estimation of mutation rate involves using animals with known genotype, many forward crosses and backcrosses with parents, and careful analysis of a large number of progeny, it may be difficult to determine the true mutation rate if parental genotype information is not available. In this situation, the **mutation frequency** (instead of mutation rate) can be calculated using the same formula. The mutation frequency does not tell when the mutation first appeared in the population; however, mutation frequency can provide an approximation of the true mutation rate.

2.4.2 Migration (Gene Flow)

Migration is the movement of organisms from one location to another. It involves movement from one subpopulation to another subpopulation, or dispersal of groups of individuals from one central population into different geographic locations. The various subpopulations of a species that has broad geographic distribution do not have the same genetic makeup; therefore, the relative frequency of various alleles may differ significantly. In such cases, migration of individuals from one subpopulation to another can add significant genetic variation to the receiving subpopulation. If the individuals from the two subpopulations then mate (panmixis), the relative frequencies of various alleles and genotypes eventually change and come to equilibrium again. In contrast, if groups of individuals move out of one central population into different geographic locations, then over time those subpopulations accumulate genetic variations independently and consequentially genetically diverge form one another.

The gene frequencies in the resulting population can be calculated by taking into account the fraction of the migrant subpopulation, the fraction of the native subpopulation, and the gene frequencies in those subpopulations, as exemplified in Box 2.3.

2.4.3 Natural Selection

Natural variations exist among the individuals in any population. Many of these differences do not affect survival or reproductive fitness (e.g. the eye color variations in humans), but some differences may improve the chances of survival of a particular group of individuals. Natural selection results in the fixation of these advantageous variations in the population, leading to greater adaptability to and reproductive success in the environment. Thus, natural selection drives the evolutionary engine.

Natural selection can be of two types, based on its effect on the fate of genetic variations: purifying (negative) selection and positive (Darwinian) selection. Purifying selection removes deleterious variations, whereas positive selection fixes beneficial variations in the population and promotes the emergence of new phenotypes. As a result, natural selection acts on populations to determine the allele frequency and distribution of quantitative traits[n] over generations. The principal types of selection determining the distribution of traits across a population are directional, stabilizing, disruptive, and balancing selection.

Directional selection favors the advantageous allele so that its proportion (and the associated phenotype) increases in the population. As a result, both the allele frequency and the phenotype are skewed in one direction and away from the average phenotype (Figure 2.5A). A popular example is the phenomenon of **industrial melanism** in the peppered moth (*Biston betularia*). This species has both light- and dark-colored phenotypes. Before the industrial revolution in England, the light-colored phenotype was predominant. During the industrial revolution, the trees on which the peppered moths

[n]A **quantitative trait** is a phenotype that is influenced by multiple genes as well as by the environment. Each gene involved in influencing a quantitative trait segregates according to Mendel's law. Because of polygenic influence, quantitative traits vary over a continuous range; hence, they are also known as **continuous traits**. As the name implies, quantitative traits can be measured. Some examples of quantitative trait phenotype in humans are skin color, height, blood pressure, and IQ. The (statistical) analysis that helps find the association between the phenotype and the molecular data in order to explain the genetic basis of complex traits is known as **quantitative trait locus** (**QTL**) analysis.

BOX 2.3

EFFECT OF MIGRATION ON GENE AND GENOTYPE FREQUENCIES

If a migrant subpopulation M migrates into a native subpopulation N, forming the resulting population R, the fraction of the migrant population in the resulting population is M/R, and that of the native population is N/R; hence, M/R + N/R = 1 (i.e. 100%).

If:

The frequency of $A_1 = p_M$ and that of $A_2 = q_M$ in subpopulation M
The frequency of $A_1 = p_N$ and that of $A_2 = q_N$ in subpopulation N
The frequency of $A_1 = p_R$ and that of $A_2 = q_R$ in the resulting population R

then:

$p_R = [(M/R \times p_M) + (N/R \times p_N)]$
$q_R = [(M/R \times q_M) + (N/R \times q_N)]$.

Example: If 300 individuals from a subpopulation (M) migrate into a native subpopulation (N) of 700 individuals, the resulting population (R) will contain 1000 individuals.

So, M/R = (300/1000) = 0.3 (i.e. 30% of the resulting population is migrant population); N/R = (700/1000) = 0.7 (i.e. 70% of the resulting population is native population).

Originally, if:

The frequency of A_1 in subpopulation M (p_M) = 0.45, and that of A_2 (q_M) = 0.55
The frequency of A_1 in subpopulation N (p_N) = 0.75, and that of A_2 (q_N) = 0.25

then:

The frequency of A_1 in the resulting population R (p_R) = $[(M/R \times p_M) + (N/R \times p_N)] = [(0.3 \times 0.45) + (0.7 \times 0.75)] = 0.66$
The frequency of A_2 in the resulting population R (q_R) = $[(M/R \times q_M) + (N/R \times q_N)] = [(0.3 \times 0.55) + (0.7 \times 0.25)] = 0.34$

Therefore, the frequencies of A_1 and A_2 in the resulting population are different from those of both the migrant and native populations.

With the change in gene frequencies, the genotype frequencies of A_1A_1, A_1A_2, and A_2A_2 in the resulting population R would change as well, and can be calculated following the Hardy–Weinberg equilibrium principle.

rested were blackened by soot. The darker background gave the dark-colored moths an advantage in hiding from predatory birds and at the same time made the light-colored moth more visible and prone to predation. As a result, over time the dark-colored moths proliferated and became the predominant phenotype while the light-colored moth population was significantly reduced. Through regulation and legislation, the environment started clearing up. As a result, the balance between light-colored and dark-colored varieties was reversed and the light-colored variety proliferated again.

Stabilizing selection is known to be the most prevalent type of natural selection; it favors the intermediate (average) phenotype of the trait, and in doing so it removes the extreme phenotypes of the trait from the population (Figure 2.5B). Thus, stabilizing selection reduces genetic variability in the population. *It is generally accepted that stabilizing selection maintains the DNA and protein sequences over evolutionary time*. However, Kimura[64] demonstrated that under stabilizing selection, extensive neutral evolution can occur through random genetic drift. In other words, many cryptic neutral genetic changes may occur in natural populations while maintaining the phenotype unchanged. A common example of stabilizing selection is the mortality and birth weight in human babies. It is well known that both very large and very small human babies suffer high mortality rates; hence, the intermediate weight is the most favored phenotype for survival.

Disruptive selection (**diversifying selection**) favors the two extreme phenotypes of the trait and minimizes the average phenotype. Thus, disruptive selection creates a bimodal distribution of a trait in the population; consequently, it is the opposite of stabilizing selection in the outcome (Figure 2.5C). Disruptive selection is an important driving force behind sympatric speciation[o]. An example of disruptive selection is provided by the mimicry and survival of the African butterfly *Pseudacraea eurytus*. In this species, the coloration

[o]**Sympatric speciation** is the process by which new species evolve from an ancestral species through the evolution of reproductive barriers while inhabiting the same geographic region. This is in contrast to **allopatric speciation**, in which geographical isolation separates two populations of a species resulting in reproductive isolation and speciation.

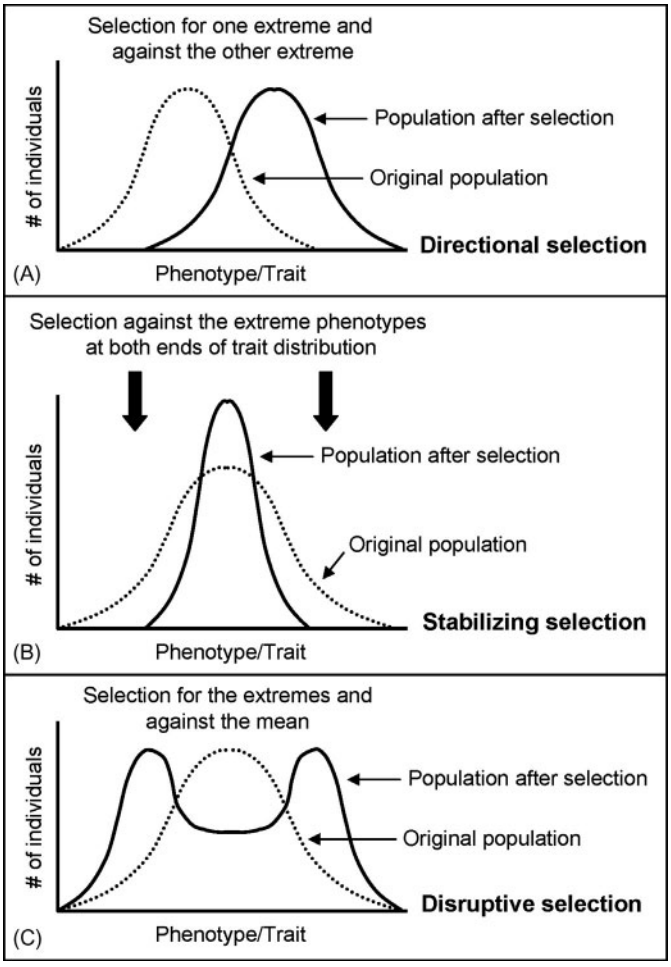


FIGURE 2.5 **Three types of natural selection.** (A) Directional selection; (B) stabilizing selection; (C) disruptive selection. See text for details.

ranges from reddish yellow to blue, with some intermediate colors. The extreme colors mimic other butterflies that are not normally preyed upon by the local predatory birds. In contrast, butterflies with intermediate coloration are devoured by the predators in greater numbers. Therefore, butterflies with extreme coloration survive in greater proportion compared to those with intermediate coloration. Another example of disruptive selection is the selection of the two extreme trophic phenotypes in the spadefoot toad (*Spea multiplicata*). Using a mark-recapture experiment in a natural pond, Martin and Pfennig[65] showed that the spadefoot toad can have different trophic phenotypes depending on the resource availability. However, disruptive selection favors the two extreme phenotypes, the small-headed "omnivore phenotype," which feeds mostly on detritus, and a large-headed "carnivorous" phenotype, which feeds on and whose phenotype is induced by the fairy shrimp. By foraging more effectively on the two alternative resource types, these extreme phenotypes avoid competition for food resources and are favored by disruptive selection, whereas the intermediate phenotypes are reduced in number.

Balancing selection (**balanced polymorphism**) maintains polymorphism in the population with respect to an allele of a trait. Therefore, balancing selection maintains genetic diversity in the population. A classic example of balancing selection is the **heterozygote advantage** in areas in Africa with high incidence of malaria. Sickle cell anemia reduces life expectancy and is caused if an individual is homozygous for a variant of hemoglobin (HbS/HbS). A red blood cell (RBC) containing HbS becomes sickle-shaped and is extremely sensitive to oxygen deprivation. However, the malarial parasite *Plasmodium* cannot survive in such sickle-shaped RBCs. Thus, heterozygous individuals, containing one normal copy and one variant copy of the hemoglobin gene (HbA/HbS), are at a survival advantage in areas with high incidence of malaria. In contrast, individuals homozygous for normal hemoglobin (HbA/HbA) are at an increased risk of death by malaria. Thus, selection maintains the apparently deleterious HbS allelic variant in the population, and balances between strong selection against both HbA/HbA and HbS/HbS genotypes by providing a selective advantage to the HbA/HbS genotype.

Based on the scale of changes, selection can lead to microevolution and macroevolution. **Microevolution** means small changes in the genome and is also associated with changes in gene frequency in a population. Over time, the accumulated small changes collectively can be significant enough to create certain new traits so that the group possessing those traits could be assigned an infra-species category, such as a **subspecies** or **variety** under the original species. In contrast, **macroevolution** means evolutionary changes leading up to the formation of species or higher taxa. The mechanisms for both micro- and macroevolutionary processes are generally the same.

2.4.4 Genetic Drift

Genetic drift (**also called random genetic drift**) means a change in the gene pool strictly by chance fixation of alleles. The effects of genetic drift can be acute in small populations and for infrequently occurring alleles, which can suddenly increase in frequency in the population or be totally wiped out. The alleles thus fixed by chance (genetic sampling error) may be neutral—that is, they may not confer any survival or reproductive advantage. Therefore, for small populations, genetic drift can result in a significant change in gene frequency in a short period of time.

Genetic drift can be caused by a number of chance phenomena, such as differential number of offspring left by different members of a population so that certain genes increase or decrease in number over generations independent of selection, sudden immigration or emigration of individuals in a population changing gene frequency in the resulting population, or population bottleneck. Of these, population bottleneck can cause a radical change in allele frequencies in a very short time. A population bottleneck occurs when a population suddenly shrinks in size owing to random events, such as sudden death of individuals due to environmental catastrophe, habitat destruction, predation, or hunting. When the small number of surviving individuals gives rise to a new population, there is a radical change in the gene frequency in the resulting population, in which certain genes (including rare alleles) of the original population may radically increase in proportion while others may radically decrease or be wiped out completely, independently of selection. Additionally, the resulting population contains a small fraction of the genetic diversity of the original population. The **founder effect** is a severe case of population bottleneck and happens when a few individuals migrate out of a population to establish a new subpopulation. Random genetic drift accompanies such founder effect, to severely reduce the genetic variation that exists in the original population. In the new population, the founder effect can rapidly increase the frequency of an allele whose frequency was very low in the original population. If the allele is a disease-related allele, the founder effect can lead to the prevalence of the disease in the new population. An increase in a specific disease in a human population due to the founder effect is seen in the Old Order Amish of eastern Pennsylvania,[66] and in the Afrikaner population of South Africa.[67]

The current Amish population has descended from a small number of German immigrants who settled in the United States during the eighteenth century. The incidence of Ellis–van Creveld syndrome (a form of dwarfism with polydactyly, abnormalities of the nails and teeth, and heart problems) is many times more prevalent in this Amish population than in the American population in general. The origin of this disease can be traced back to one couple, Samuel King and his wife, who came to the area in 1744. The mutated gene that causes the syndrome was passed along from the Kings and their offspring. The Amish population practices endogamy (individuals tend to mate within their own subgroup). Additionally, in this community the **gene flow is centrifugal**—that is, members may leave the community but outsiders do not join the community—therefore, there has been no introduction of exogenous genes into the Amish gene pool. As a result, the frequency of the disease gene has rapidly increased over generations.

Another example of founder effect comes from the Afrikaner population of South Africa, which is mainly descended from one group of European (mainly Dutch, but also German and French) immigrants that landed there in 1652. The present-day Afrikaner population has a very high prevalence of Huntington's disease; over 200 affected individuals in more than 50 supposedly unrelated families have been found to be ancestrally related through a common progenitor in the seventeenth century. Thus, the root of the disease can be traced back over 14 generations to a common progenitor who supposedly carried the gene for Huntington's disease. Huntington's disease is an autosomal dominant disease caused by triplet (CAG) repeat expansion in the gene (and the mRNA), containing 40 to >100 CAG triplets. The onset and severity of the disease is directly correlated with the number of repeats.

2.4.5 Nonrandom Mating

Changes in gene frequency by genetic drift are influenced in a large part by the breeding structure of the population—that is, whether the population practices random mating or nonrandom mating. **Inbreeding** is the most common form of nonrandom mating. Inbreeding occurs when genetically related individuals preferentially mate with each other (e.g. mating between relatives). The most extreme form of inbreeding is **self-fertilization**. Inbreeding produces a larger excess of homozygotes in the population than would be expected from random mating. Consequently, inbreeding also increases the frequency of homozygotes of rare alleles, including rare recessives, which will be subject to selection. If a rare allele is deleterious, its frequency can rise through homozygosity because of significant inbreeding in a normally outbreeding population. This phenomenon is called **inbreeding depression**.

Inbreeding is measured by the **inbreeding coefficient** (F), which is a measure of the probability that two alleles are identical by descent. This means the degree to which two alleles are more likely to be homozygous than heterozygous simply because the parents are genetically related. The value of F can theoretically range from 0 (0%; hence no inbreeding, completely random mating) to 1 (100%; hence complete inbreeding, all alleles are identical by descent).

If the frequency of allele A is p and the frequency of allele a is q, and the value of F is known, then the frequencies of genotypes AA, Aa and aa are determined as follows:

$$AA = p^2 + Fpq; \quad Aa = 2pq - 2Fpq; \quad Aa = q^2 + Fpq. \quad (2.1)$$

2.5 THE NEUTRAL THEORY OF EVOLUTION

The Darwinian theory of evolution by natural selection is based on the assumption that new mutations that constantly arise in the population are mostly adverse but some are beneficial. Natural selection filters out the adverse mutations, while fixing beneficial mutations in the population. In other words, evolution is caused by natural selection acting through beneficial mutations fixed in the population. Thus, it is an underlying assumption by Darwinian evolutionists that neutral mutations that do not confer any selective advantage or disadvantage are very rare, if they exist at all. A corollary to this assumption is that genetic drift, which causes chance fixation of neutral alleles, could not have played any role in evolution.

This long-held view of molecular evolution was challenged by the neutral theory of molecular evolution, proposed by Kimura.[68] In brief, the neutral theory postulates that evolutionary changes at the molecular level are not caused by natural selection alone acting only on advantageous mutations, but are mostly caused by random chance fixation of selectively neutral or near-neutral alleles (genetic drift). Therefore, genetic drift plays an important role in molecular evolution. To expand the concept, according to neutral theory, the majority of new mutations are either deleterious or neutral. Deleterious mutations adversely affect the fitness of the carrier whereas neutral mutations do not affect the fitness of the carrier (hence, selectively neutral). *Fitness in the context of evolution means the ability to reproduce, and contribute to the gene pool of the next generation*. Deleterious mutations that adversely affect fitness are removed from the population by purifying selection. In contrast, neutral mutations are subject to chance sampling and random fixation in every generation. In this process, some neutral mutations are fixed randomly by sheer chance while others are removed from the population. Once a neutral mutation is fixed by chance, its frequency increases by genetic drift, which leads to genetic polymorphism in the population. These genetic variations in the population provide the raw materials for molecular evolution. The allele carrying the new fixed mutation is called a **derived allele**, as opposed to the **ancestral allele** from which it is derived. As mentioned above, extensive neutral evolution can occur through random genetic drift while the phenotype is still maintained unchanged under stabilizing selection.[64]

It should be remembered that neutral theory does not deny the role of natural selection in evolution—that is, it does not deny the importance of positive selection in the origin of adaptations—it simply complements the Darwinian view by emphasizing the role of neutral mutations as additional raw materials for evolution and genetic drift as an additional mechanism of evolution. The neutral theory also predicts that purifying selection is ubiquitous, but positive selection is rare.[69]

2.5.1 Synonymous and Nonsynonymous Substitutions, Constraints on Changes in Gene and Protein Sequence, and Evolution

A nucleotide substitution that changes the corresponding amino acid in the protein is called a **nonsynonymous substitution** (denoted as K_A), whereas a nucleotide substitution that does not change the amino acid in the protein is called a **synonymous substitution** (denoted as K_S).

The neutral theory predicts that synonymous substitutions will be tolerated, but nonsynonymous substitutions will be removed by purifying selection. Consequently, nonsynonymous substitutions will be fewer than synonymous substitutions. Consistent with this prediction, it is known that synonymous substitutions typically exceed nonsynonymous substitutions in protein-coding genes, and functionally constrained regions of genes evolve at a slower rate than regions that are not functionally constrained. However, if a nonsynonymous substitution confers some selective advantage, then it will be rapidly fixed in the population by positive selection. The average rates of synonymous and nonsynonymous substitutions previously calculated were 4.7 substitutions/synonymous site versus 0.88 substitutions/nonsynonymous site per 10^9 (billion) years, respectively.[70] This estimate was subsequently revised to 3.51 substitutions/synonymous site versus 0.74 substitutions/nonsynonymous site per 10^9 (billion) years in rodents and humans, as stated earlier in this chapter.

2.5.2 Signatures of Positive Selection

A prediction of the neutral theory is that if the substitutions are all neutral, then for a given protein-coding gene the K_A/K_S ratio between two species should be very similar to the same ratio within species (null hypothesis), and it is the deviation from this prediction that provides support for positive selection (with some exceptions, such as relaxation of purifying selection and population bottleneck). McDonald and Kreitman[71] proposed a simple method to determine signatures of positive selection in protein sequence (see Box 2.4). The test relies on determining statistically significant deviation from the prediction of the neutral theory (the null hypothesis) that if the

BOX 2.4

THE MCDONALD–KREITMAN TEST

The McDonald–Kreitman method tests the neutral theory as the null hypothesis (H_0) against the (positive) selection hypothesis as the alternative hypothesis (H_1). In this test, two DNA sequences are aligned. Nucleotide substitutions in the coding region are classified in two ways: (1) **synonymous** versus **replacement**, and (2) **fixed difference** versus **polymorphic**.

1. **Synonymous** versus **replacement** substitutions: Synonymous substitutions result in a synonymous codon and no amino acid change in the protein, whereas replacement (or **nonsynonymous**) substitutions result in a nonsynonymous codon and amino acid change.
2. **Fixed difference** versus **polymorphic** substitutions: Polymorphic substitutions show variations within species, whereas fixed difference (also called **fixed divergence**) substitutions differ between species but not within species. Such dual classification allows the use of a 2×2 table. McDonald and Kreitman studied the sequence evolution of the *Adh* gene in *Drosophila melanogaster*, *Drosophila simulans*, and *Drosophila yakuba*. Tabulating the alignment data provided the following table:

	Fixed Difference (between species)	Polymorphism (within species)
Synonymous (K_S) (no amino acid change)	17	42
Replacement (K_A) (amino acid change)	7	2

$G = 7.43$; $P = 0.0006$.

McDonald and Kreitman used the *G*-test for statistical independence to determine if the cells in the 2×2 table were independent. In other words, whether the proportion of replacement versus synonymous changes was independent of whether the changes were fixed or polymorphic; similarly whether the proportion of fixed difference versus polymorphism was independent of whether the changes were synonymous or replacement.

The replacement/synonymous substitution ratio (K_A/K_S) of the fixed differences between species is 7/17 (= 0.41), whereas the same ratio of the polymorphic sites within species is 2/42 (= 0.048). *Thus, there is a more than eight-fold excess of replacement mutations between species compared to polymorphic mutations within species*. Similarly, the fixed difference/polymorphic substitution ratio among synonymous sites is 17/42 (= 0.40), whereas the same ratio among replacement sites is 7/2 (= 3.5). *Thus, there is a more than eight-fold excess of replacement substitutions compared to synonymous substitutions between species*. If all these substitutions were neutral, no such statistically significant differences would be expected. Therefore, the result of the *G*-test of independence indicates deviation from the assumptions of neutral evolution, thereby signifying a strong signature of positive selection.

substitutions are all neutral, then for a given protein-coding gene, the K_A/K_S ratio at divergent sites **between** species should be very similar to the same ratio at polymorphic sites **within** species. Deviation from the null hypothesis will constitute evidence of positive selection.

Signatures of positive selection, however, are not very widespread, except in some select groups of genes, such as genes important in host–pathogen interactions, as well as in sex-related genes. For example, strong signatures of positive selection, with K_A/K_S ratios ranging from 1.36 to 5.15, were observed when two proteins (16 and 18 kDa) in the acrosomal vesicle of abalone spermatozoa were compared. These values were among the highest for full-length sequences analyzed so far.[72]

2.5.3 Selective Sweep and the Hitchhiking Effect

If a new mutation offers increased fitness to the carrier, it is fixed in the population through positive selection, and its frequency rapidly increases. Such rapid fixation of an advantageous mutation is called **selective sweep**. As the frequency of the new mutation

BOX 2.5

NEUTRAL EVOLUTION–MUTATION RELATIONSHIP

1. The probability of fixation of a mutation (p) in a diploid population of size N is $1/2N$ (i.e. $p = 1/2N$).
2. The rate of substitution per unit time (k) in a diploid population of size N = the number of mutations fixed per unit time in a diploid population of size $N \times$ the probability of fixation of a mutation (p).
3. Because the number of mutations fixed per unit time is the mutations rate μ, and the number of any gene in a diploid population of size N is $2N$, the number of mutations fixed per unit time in a diploid population of size $N = 2N \times \mu$.
4. Hence, point (2) stated above can be expressed as $k = 2N \times \mu \times p$.
5. Because $p = 1/2N$, p can be substituted for $1/2N$ and point (2) can be rewritten as $k = 2N \times \mu \times 1/2N$; or $k = \mu$.
6. In other words, the rate of substitution per unit time—i.e. the rate of neutral evolution (k)—is equal to the mutation rate (μ) of neutral alleles, and is independent of the population size.

increases, the frequency of the genes/sequences around it that are very closely linked and not easily separated by recombination also increases. The net result is a loss of sequence variability around the newly fixed mutation in the population. The increase in frequency of the neighboring genes/sequences, simply because of their close proximity to the newly fixed mutation, is called the **hitchhiking effect**, or **genetic hitchhiking**. Selective sweep and the hitchhiking effect are the results of strong positive selection. The hitchhiking effect may also lead to an increase in the proportion of somewhat disadvantageous or deleterious mutations in the population.[73]

2.6 MOLECULAR CLOCK HYPOTHESIS IN MOLECULAR EVOLUTION

Kimura's neutral theory derived support from the **molecular clock hypothesis**. The molecular clock hypothesis states that the rate of molecular evolution of a gene (the rate of nucleotide substitution) or a protein (the rate of amino acid substitution) is approximately constant over evolutionary time. In other words, the number of replacements in the gene or protein is proportional to the time since their origin—that is, the number of replacements per unit time is similar. The hypothesis was based on the initial observation of amino acid substitutions in human and horse hemoglobin by Zuckerkandl and Pauling in 1962. This was followed by similar observations on cytochrome c from seven different eukaryotic species: horse, human, pig, rabbit, chicken, tuna, and baker's yeast.[74] The term "molecular clock hypothesis" was coined by Zuckerkandl and Pauling in 1965. The concept of the molecular clock fits well with Kimura's neutral theory because the rate of neutral evolution is equal to the mutation rate of neutral alleles, as shown in Box 2.5.

However, after more protein sequences were studied in the 1970s, it was realized that the rate of substitution could differ significantly in different proteins and different organisms. Nonetheless, the molecular clock represents a valuable tool in studies of evolution and molecular systematics, and it has been widely used in estimation of divergence times and reconstruction of phylogenetic trees.

2.7 MOLECULAR PHYLOGENETICS

Phylogeny refers to the evolutionary history of organisms or populations. **Phylogenetics** is the study of phylogenies—that is, the study of the evolutionary relationships among various organisms and populations. According to evolutionary theory, the similarity among organisms and groups of organisms is attributable to their descent from a common ancestor. This similarity extends even to the structure and function of molecules, such as DNA and proteins. Traditional phylogenetics considered morphological features. Modern phylogenetics uses information from DNA and protein sequences. The use of DNA and protein sequence information and their change over evolutionary time in order to infer the evolutionary relationship among a set of homologous genes or proteins is referred to as **molecular phylogenetics**. The goal of molecular phylogenetics is to estimate the evolutionary divergence of the DNA and protein sequences from a common ancestral sequence, and thus reconstruct the correct evolutionary relationships among these sequences in the form of a phylogenetic

tree. With the advent of molecular biology techniques, particularly DNA sequencing, molecular phylogenetic studies have become very common. Sometimes molecular phylogenetics is used to infer the evolutionary relationships among organisms. *In general, inference on evolutionary relationships based on protein sequences is preferred to that based on nucleic acid sequences.*

2.7.1 From Systematics and Biological Classification to Molecular Phylogenetics

> Systematics is the scientific study of the kinds and diversity of organisms and of any and all relationships among them ... Classification of organisms is an activity that belongs exclusively to systematics. ***G. G. Simpson***[75]

Biological classification is concerned with ordering (arranging) organisms or groups of organisms, both **living** (**extant**) and **fossil** (**extinct**), into hierarchical and multilevel categories based on their evolutionary relationships. Therefore, the conceptual foundation of the science of systematics and the activity of biological classification is the evolutionary (phylogenetic) relationship among taxa. The expression **phylogenetic systematics** (also known as **cladistics**, discussed in Section 2.7.2.2) underscores the link between systematics and phylogeny. Because classification of organisms takes into consideration their evolutionary relationships, the revision of older classification schemes with modern data, particularly ancestral and derived characters and homology (discussed later under cladistics), has affected only minor details.[76] With the availability of the vast amount of molecular data and analytical tools, molecular phylogenetics has become the norm for studying the evolutionary relationships. Nevertheless, for historical reasons it is appropriate to consider molecular phylogenetics against the backdrop of systematics and biological classification.

The first systematic way of classifying organisms was introduced by the Swedish botanist Carl Linnaeus. Linnaeus's classification scheme involved categorizing organisms based solely on morphological characters without any evolutionary context. He published his work as a book called ***Systema Naturae***. The 10th edition of *Systema Naturae*, published in 1758, is considered to be the beginning of biological classification and the **binomial nomenclature** system in biology. In binomial nomenclature, an organisms is given a name composed of two parts, usually using latinized expression; the first part identifies the genus to which the species belongs and the second part identifies the species within the genus. The original Linnaean classification scheme is called **Linnaean hierarchy**, and it had seven categories: kingdom, phylum, class, order, family, genus, and species. These categories are called **taxonomic categories**. Organisms that are the subjects of classification are called **taxa** (singular: **taxon**). Modern biological classification systems have many more taxonomic categories compared to the seven originally proposed by Linnaeus.

Linnaeus introduced his system of classification 100 years before the theory of evolution was proposed by Darwin; hence, it had no evolutionary context. Linnaeus's classification scheme was based on choosing "similar" characters, and such choice was more or less arbitrary. With a greater understanding of genetics—including population genetics, mechanism of evolution, and relationships among the living and extinct organisms at the biochemical and molecular levels—it became apparent that biological classification should reflect the relationships among organisms or groups of organisms by their descent from a common ancestor during evolution. *The meaning of "similarity" in modern biological classification is ancestral similarity (homology).*

2.7.2 Systems of Biological Classification

The three main systems of modern biological classification are **phenetics**, **cladistics**, and **evolutionary classification**. For all practical purposes, phenetics is no longer used as a phylogenetic method, whereas cladistics has become the most widely used method for molecular phylogenetic analysis.

2.7.2.1 Phenetics and Phenograms

Phenetics, also known as **numerical taxonomy**, was introduced in the 1950s.[77] Phenetics attempts to group species into higher taxa based on overall similarity, usually in morphology or other observable traits, and regardless of their phylogeny or evolutionary relationships. Many different characteristics are used to calculate a similarity coefficient, varying between 0 (no similarity) to 1 (highest similarity), between all pairs of organisms that are subjects of phenetic classification. Similarity coefficients are used to create a similarity matrix and develop a **phenogram**, which is a tree-like network expressing phenetic relationships. According to the proponents of phenetics, similarity is expected among the descendants of a common ancestor; therefore, grouping together the most similar taxa automatically produces phylogenetic classification. Although phenetics is not used anymore, its historical importance lies in introducing computer-based numerical algorithms, which are now essential in all modern phylogenetic analyses.

2.7.2.2 Cladistics, Clades, and Cladograms

The main proponent of cladistics was the German entomologist Willi Hennig in the mid-twentieth century.

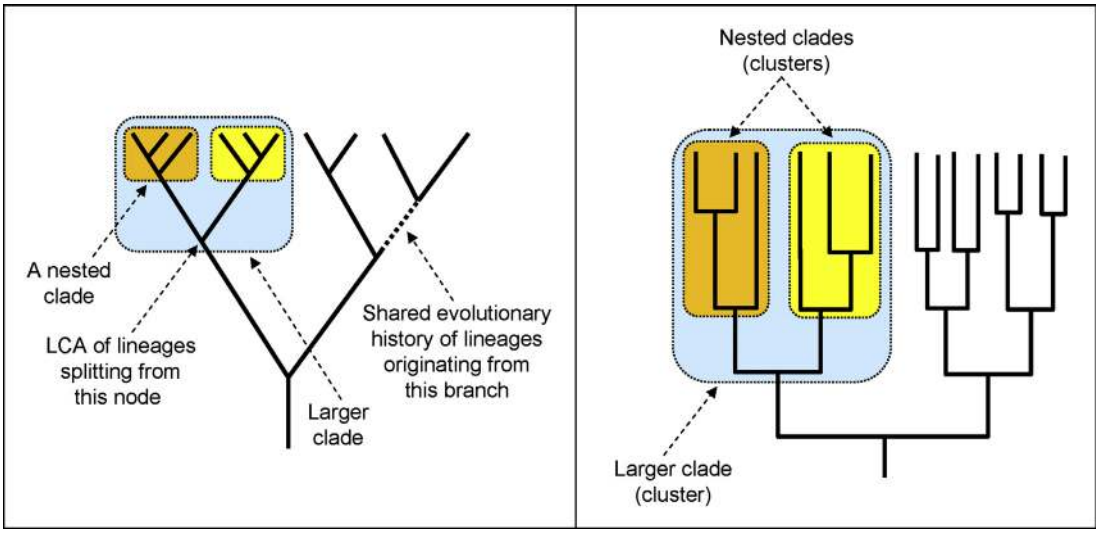


FIGURE 2.6 **Nested clades within a larger clade in a phylogenetic tree.** A typical cladogram on the left and a typical dendrogram on the right. In a phylogenetic tree, each branching point (node) represents the LCA of the lineages (including nodes) arising from this point. A branch preceding a node represents the shared evolutionary history of lineages that split from the node.

Cladistics is also known as **phylogenetic systematics** or **phylogenetic classification**. Cladistics classifies organisms based on shared derived characters. Therefore, taxa that share specific derived characters are grouped more closely together than those who do not. The groups are called **clades**; each clade consists of an ancestor and all of its descendants. The relationships between clades are shown in a branching hierarchical tree called a **cladogram**. Depending on the branching of the cladogram, it is possible to identify smaller clades within a larger clade; the smaller clades are called **nested clades.** Figure 2.6 shows nested clades within a larger clade in a phylogenetic tree. The phylogenetic tree has been represented as a typical cladogram on the left and as a typical **dendrogram** on the right. The dendrogram is sometimes loosely called a cladogram. In a phylogenetic tree (cladogram), each branching point (node) represents the **last common ancestor** (**LCA**) of the lineages (including nodes) arising from this point. The separation of taxa along the cladogram is driven by evolutionary innovation of new characters (evolutionary novelties or apomorphies, discussed below).

2.7.2.2.A SOME IMPORTANT TERMINOLOGY OF CLADISTICS

Terms used to describe various character states that are relevant in the discussion of cladistics include **apomorphy**, **synapomorphy**, **plesiomorphy**, **symplesiomorphy**, **autapomorphy**, and **homoplasy**. The terms are described below with examples.

A **primitive** or **ancestral character state** is called **plesiomorphy** (**plesiomorphic character**), and a shared plesiomorphy is called a **symplesiomorphy**. For example, hair is a unique mammalian character that evolved with the evolution of mammals. Mammalian evolution was followed by further evolution of various mammalian groups and subgroups based on evolutionary novelties. For example, primates form a more recently evolved mammalian group. Therefore, hair is a plesiomorphy (ancestral character) for primates. Because hair, as an ancestral mammalian character, is shared by all primates, it is also a symplesiomorphy (shared plesiomorphy) for primates in general.

In contrast to an ancestral character state, a **derived character state** (**evolutionary novelty**) is called **apomorphy** (**apomorphic character**), and a shared apomorphy is a **synapomorphy**. For example, hair is an apomorphy for mammals as a group because it distinguishes mammals from other vertebrate clades, such as reptiles. Because hair is shared by all mammals, it is also the synapomorphy (shared apomorphy) for mammals in general. Among mammals, different groups have their own apomorphies. For example, an opposable thumb is an apomorphy for primates because it is an evolutionary novelty for primates and is not found in non-primate mammals. Similarly, the feather is an apomorphy for birds. Therefore, an apomorphy for a larger clade can be a plesiomorphy for a smaller nested clade within that larger clade.

An apomorphy that is unique to a taxon is called **autapomorphy**. An example of a non-anatomical autapomorphy in modern humans is speech, which is unique to humans.

A character state that evolved because of **convergent evolution** but was not acquired through common evolutionary lineage is called **homoplasy**, and the

character is called a **homoplastic character**. Homoplastic characters evolve independently in multiple taxa in different evolutionary lineages in response to adaptation; these characters are not present in their common ancestor. For example, fins evolved independently in sharks (cartilaginous fish) and dolphins (mammals) to perform the same function, but they are structurally different and were not derived from their common ancestor. Hence, the fin is a homoplastic character for sharks and dolphins. In contrast to homoplasy, **homology** is a character state shared by a set of species and is present in their common ancestor. The term homology is pervasive in the evolutionary literature, including molecular evolution.

2.7.2.3 *Evolutionary Classification*

The third system of modern biological classification is referred to as **evolutionary classification**, also known as **Darwinian classification**, **evolutionary taxonomy**, and **evolutionary systematics**. It is actually the oldest of the three approaches and its strongest proponents include renowned evolutionary biologists such as Ernst Mayr, George Gaylord Simpson, and Julian Huxley. Mayr and Bock[78] emphasized that, contrary to the general belief, not all biological classifications are evolutionary classifications. They opined that evolutionary classification is more inclusive than ordering systems (e.g. phenetics and cladistics), which are based on just the pattern of branching points. Nevertheless, ordering systems producing dendrograms and cladograms are still useful phylogenetic classification schemes. Proponents of evolutionary classification maintain that classifications should reflect the two aspects of evolutionary change: (1) the splitting of the phyletic lineages—that is, the branching in the phylogenetic tree—and (2) the invasion of new environmental niches—that is, adaptation and evolutionary divergence. Therefore, the amount of evolutionary change after the branching points is an important consideration in evolutionary classification. In order to take account of this, evolutionary classification weighs the evolutionary innovations (apomorphic characters) that determine the branching point in the tree. Major evolutionary innovations that help a new phyletic lineage adapt to a new environment and drive adaptive evolution are given greater weight. Therefore, evolutionary classification tries to tell the evolutionary history of the taxonomic group.

Each of the three methods discussed above has its own strengths and shortcomings, and the proponents of each method claim that their method is the best. However, cladistics has become the method of choice for molecular phylogenetic analysis because of the molecular (sequence) data used to measure divergence from an ancestral taxon. This is probably why the use of cladistics has progressively increased with the increase in the number of entries in DNA and protein sequence databases, and has now become commonplace in molecular phylogenetic analysis.

2.7.3 Phylogenetic Tree

A **phylogenetic tree** or **evolutionary tree** is a diagrammatic representation of the evolutionary relationship among various taxa. The phylogenetic tree, including its reconstruction and reliability assessment, is discussed in more detail in Chapter 9. The terms **evolutionary tree**, **phylogenetic tree**, and **cladogram** are often used interchangeably to mean the same thing—that is, the evolutionary relationships among taxa. The term dendrogram is also used interchangeably with cladogram, although there are subtle differences, discussed in Chapter 9. Thus, it is important to be aware that usage of the vocabulary is not always consistent in the literature, although the context is the same, that is, representation of the evolutionary relationships of taxa.

References

1. Zuckerkandl E, Pauling L. *J Theor Biol* 1965;**8**:357–66.
2. Higgs PG, Attwood TK. *Bioinformatics and molecular evolution*. MA: Blackwell; 2005. pp. 1–11
3. Mayr E. *Populations, species, and evolution*. New York: Belknap Harvard; 1970. pp. 10–20
4. Gould SJ, Eldredge N. *Paleobiology* 1977;**3**:115–51.
5. Mills DR, et al. *Proc Natl Acad Sci USA* 1967;**58**:217–24.
6. Cairns J, et al. *Nature* 1988;**335**:142–5.
7. Trottier Y, et al. *Mol Carcinogen* 1992;**6**:140–7.
8. Bruner SD. *Nature* 2000;**403**:859–66.
9. Eichler EE, et al. *Hum Mol Genet* 1995;**4**:2199–208.
10. Falik-Zaccai TC, et al. *Am J Human Genet* 1997;**60**:103–12.
11. Eichler EE, et al. *Nat Genet* 1995;**11**:301–8.
12. Cabot EL, et al. *Genetics* 1993;**135**:477–87.
13. Ohno S. *Evolution by gene duplication*. New York: Springer-Verlag; 1970.
14. Hokamp K, et al. *J Struct Funct Genom* 2003;**3**:95–110.
15. Kasahara M. *Curr Opin Immunol* 2007;**19**:547–52.
16. Makałowski W. *Genome Res* 2001;**11**:667–70.
17. Lynch M, Conery JS. *Science* 2000;**290**:1151–5.
18. Lynch M, Conery JS. *J Struct Func Genom* 2003;**3**:35–44.
19. Cotton JA, Page RDM. *Proc R Soc Lond B* 2005;**272**:277–83.
20. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 99–164
21. Hirotsune S, et al. *Nature* 2003;**423**:91–6.
22. Thibaud-Nissen F, et al. *BMC Genomics* 2009;**10**:317.
23. Deng C, et al. *Proc Natl Acad Sci USA* 2010;**107**:21593–8.
24. Escriva H, et al. *PLoS Genet* 2006;**2**:e102.
25. Lynch M. *The origins of genome architecture*. Sunderland, MA: Sinauer; 2007. pp. 193–235
26. Rastogi S, Liberles DA. *BMC Evol Biol* 2005;**5**:28.
27. Force A, et al. *Genetics* 1999;**151**:1531–45.
28. Hahn MW. *J Heredity* 2009;**100**:605–17.
29. Kleinjan DA, et al. *PLoS Genet* 2008;**4**:e29.

30. Tocchini-Valentini GD, et al. *Proc Natl Acad Sci USA* 2005;**102**:8933−8.
31. Kolkman JA, Stemmer WP. *Nat Biotechnol* 2001;**19**:423−8.
32. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 249−322
33. International Human Genome Sequencing Consortium. *Nature* 2001;**409**:860−921.
34. Kaessmann H, et al. *Genome Res* 2002;**12**:1642−50.
35. Franca GS, et al. *Genetica* 2012;**140**:249−57.
36. Patthy L. *Gene* 1999;**238**:103−14.
37. Snel B, et al. *Trends Genet* 2000;**16**:9−11.
38. Enright A, Ouzounis C. *Genome Biol* 2001;**2**:34.1−7.
39. Yanai I, et al. *Proc Natl Acad Sci USA* 2001;**98**:7940−5.
40. Kummerfeld SK, Sarah A. *Trends Genet* 2005;**21**:25−30.
41. Syvanen M. *Annu Rev Genet* 2012;**46**:341−58.
42. Acuña R, et al. *Proc Natl Acad Sci USA* 2012;**109**:4197−202.
43. Moran NA, Jarvik T. *Science* 2010;**328**:624−7.
44. Altincicek B, et al. *Biol Lett* 2012;**8**:253−7.
45. Tautz D, Domazet-Lošo T. *Nat Rev Gen* 2011;**12**:692−702.
46. Levine MT, et al. *Proc Natl Acad Sci USA* 2006;**103**:9935−9.
47. Begun DJ, et al. *Genetics* 2007;**176**:1131−7.
48. Zhou Q, et al. *Genome Res* 2008;**18**:1446−55.
49. Knowles DG, McLysaght A. *Genome Res* 2009;**19**:1752−9.
50. Li C-Y, et al. *PLoS Comput Biol* 2010;**6**:e1000734.
51. Wu D-D, et al. *PLoS Genet* 2011;**7**:e1002379.
52. Sorek R, et al. *Mol Cell* 2004;**14**:221−31.
53. Schmitz J, Brosius J. *J Biochim* 2011;**93**:1928−34.
54. Butticè G, et al. *J Mol Evol* 1990;**30**:479−88.
55. Purevsuren J, et al. *Mol Genet Metab* 2008;**95**:46−51.
56. Raponi M, et al. *Hum Mut* 2006;**27**:294−5.
57. Schlager G, Dickie MM. *Genetics* 1967;**57**:319−30.
58. Schlager G, Dickie MM. *Mutat Res* 1971;**11**:89−96.
59. Russell LB, Russell WL. *Proc Natl Acad Sci USA* 1996;**93**:13072−7.
60. Monant Jr. RJ. *Cancer Res* 1989;**49**:81−7.
61. Drake JW, et al. *Genetics* 1998;**148**:1667−86.
62. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 5−38
63. Mills RE, et al. *Genome Res* 2011;**21**:830−9.
64. Kimura M. *Proc Natl Acad Sci USA* 1981;**78**:5773−7.
65. Martin RA, Pfennig DW. *Am Nat* 2009;**174**:268−81.
66. McKusick VA. *Nat Genet* 2000;**24**:203−4.
67. Hayden MR, et al. *S Afr Med J* 1980;**58**:197−200.
68. Kimura M. *Nature* 1968;**217**:624−6.
69. Hughes AL. *Heredity* 2007;**99**:364−73.
70. Li WS, et al. *Mol Biol Evol* 1985;**2**:150−74.
71. McDonald JH, Kreitman M. *Nature* 1991;**351**:652−4.
72. Vacquier VD, et al. *J Mol Evol* 1997;**44**(Suppl. 1):S15−22.
73. Chun S, Fay JC. *PLoS Genet* 2011;**7**:e1002240.
74. Margoliash E. *Proc Natl Acad Sci USA* 1963;**50**:672−9.
75. Simpson GG. *Principles of animal taxonomy*. New York: Columbia University Press; 1961. pp. 1−33
76. Mayr E, Ashlock PD. *Principles of systematic zoology*. 2nd ed. New York: McGraw-Hill. 1991, pp. 113−58.
77. Mayr E, Ashlock PD. *Principles of systematic zoology*. 2nd ed. New York: McGraw-Hill. 1991, pp. 195−205.
78. Mayr E, Bock WJ. *J Zool Syst Evol Res* 2002;**40**:169−94.

CHAPTER

3

Genomic Technologies*

OUTLINE

3.1 ADVANCES IN GENOMICS

Advances in genomics have broadened the scope of many already existing techniques from the gene scale to the genome scale with a concomitant drop in cost; DNA-sequencing and gene-expression-measurement technologies being the greatest beneficiaries. Genomics has two broad aspects: structural and functional. Structural genomics attempts to study the three-dimensional (3D) structure of proteins encoded by a genome. Therefore, the structural genomics approach requires the knowledge of the genome sequence, which is integrated with experimental and modeling data to predict the 3D structure of proteins. As the name implies, functional genomics aims to study gene (and protein) functions and interactions. Thus, functional genomics focuses on processes, such as transcription, translation, and protein−protein interaction. In reality, structural and functional aspects of genomics have overlaps simply because they both require knowledge of the genome sequence.

With the advancement of genomics, traditional molecular biology techniques—such as cloning, nucleic acid amplification, sequencing, mutagenesis, mutation detection, gene and protein interaction and expression studies—have been significantly improved in terms of their efficiency, cost, and high-throughput nature. Of these techniques, DNA-sequencing and gene-expression technologies have been revolutionized the most, and the scope of these techniques has been improved from the gene scale to the genome scale.

3.2 FROM SANGER SEQUENCING TO PYROSEQUENCING

Genome sequencing is the most direct method of detecting mutations, such as single nucleotide

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

polymorphisms (SNPs) and copy number variations (CNVs). The development of the dideoxy method of DNA sequencing was a major step forward for the science of molecular biology. The dideoxy method of DNA sequencing was published by Sanger and colleagues in 1977.[1] The technique is based on the **chain-termination** principle—that is, when DNA polymerase elongates the DNA chain, the incorporation of a dideoxynucleotide causes the termination of further chain elongation. This technique is not discussed any further because it is now the subject of textbooks. About 20 years after the development of Sanger's dideoxy sequencing, Pal Nyren introduced the **pyrosequencing** technique.[2] The pyrosequencing technique paved the way for the development and commercialization of large-scale, high-throughput, massively parallel sequencing technology, popularly referred to as **next-generation sequencing** or **next-gen sequencing** (**NGS**) **technology**.

3.3 PYROSEQUENCING, MUTATION DETECTION, AND SNP GENOTYPING

Pyrosequencing is based on the **sequencing by synthesis** principle. When DNA polymerase elongates the DNA chain, pyrophosphates are released. Each released pyrophosphate triggers a series of reactions that generates a detectable quantum of light. Therefore, pyrosequencing enables real-time detection of the sequence of a gene. Consequently, this technique is useful in the rapid detection of point mutations in the sequence and in SNP genotyping, including genotyping of microbes.

The DNA template that needs to be sequenced is first amplified by polymerase chain reaction (PCR). The amplicon (double-stranded amplified fragment) length is usually less than 200 bp for efficient pyrosequencing, but could be longer. While the number of cycles in regular PCR is around 30, the number of cycles in PCR for pyrosequencing is around 50. This is to ensure that the primers and the free nucleotides are utilized as much as possible. One of the two PCR primers is biotinylated at the 5′-end. The PCR amplicon containing a biotinylated end is captured on streptavidin-coated sepharose beads, denatured by alkali, and purified prior to pyrosequencing. The biotinylated strand is used as the template for pyrosequencing. A pyrosequencing primer (the third primer) is added to the purified biotinylated PCR strand and pyrosequencing is carried out.

Pyrosequencing is conducted in 96-well plates. During this process, the sequencing primer is first allowed to anneal with the DNA template in the presence of four enzymes—DNA polymerase, ATP sulfurylase, luciferase, and apyrase—and two substrates—adenosine 5′-phosphosulfate (APS) and luciferin—but without the deoxynucleotide triphosphates (dNTPs). Then, individual dNTPs are added to the reaction sequentially in a fixed order, which is programmed before the run. Out of the four dNTPs, only dATP is replaced by deoxyadenosine alpha-thio triphosphate (dATPαS). If the added dNTP is complementary to the base in the template strand, it is incorporated by the DNA polymerase and a pyrophosphate (PP_i) is released. ATP sulfurylase uses this PP_i and APS to generate ATP. The ATP is utilized by luciferase to oxidize luciferin into oxyluciferin with the concomitant emission of light, which is recorded by a charge-coupled device (CCD) camera in the form of a peak. Because of the stoichiometry of the reaction, the peak height is directly proportional to the number of nucleotides incorporated in tandem. Thus, if two of the same bases are incorporated back to back, the peak height becomes double, and so on. If the injected dNTP is not complementary to the template base, no signal is produced. Unutilized dNTPs are degraded by apyrase. The apyrase reaction is very important to keep the background noise level low. The readout of the pyrosequencing is called a **pyrogram** (Figure 3.1).

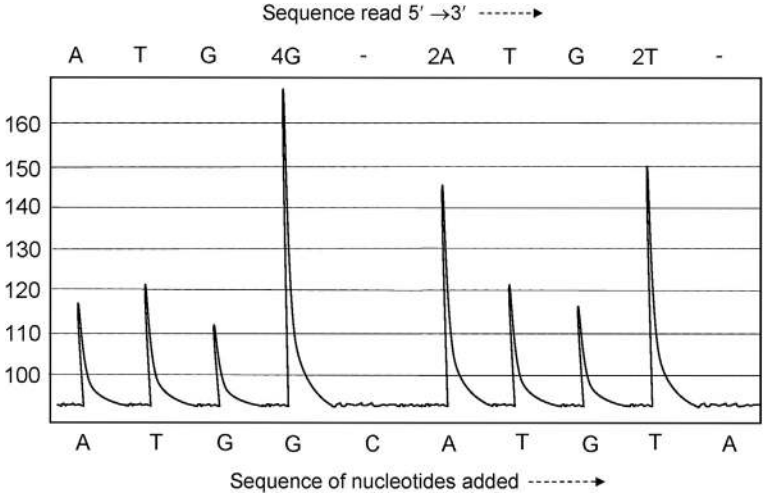


FIGURE 3.1 **A hypothetical pyrogram showing the sequence determination.** The peak height is proportional to the number of contiguous bases. There are four "G"s, two "A"s and two "T"s in this sequence. No peak was found at C in the middle and at A at the far right. The sequence for this window is ATGGGGGAATGTT.

By comparing the pyrogram of the query DNA (sample) with that of the wild-type DNA (reference), SNPs can be detected. The algorithm involves statistical analysis for significance. The enzymatic reactions of pyrosequencing are:

1. $DNA_n + dNTP \rightarrow DNA_{n+1} + PP_i$ (catalyzed by DNA polymerase)
2. $PP_i + APS \rightarrow ATP$ (catalyzed by ATP sulfurylase)
3. $ATP + luciferin + O_2 \rightarrow oxyluciferin + light\ quanta$ (catalyzed by luciferase)
4. Unincorporated $dNTP \rightarrow dNMP + 2\ P_i$ (catalyzed by apyrase)

3.4 NEXT-GENERATION SEQUENCING PLATFORMS

Next-generation sequencing (NGS) is high-throughput, massively parallel sequencing. NGS is also referred to as **second-generation** sequencing technology (the first generation being the original sequencing techniques of Sanger, and Maxam and Gilbert). The proposed cost of the first human genome sequencing was $3 billion ($3000 million). The sequencing of the genome of Dr J. Craig Venter reportedly cost $100 million, whereas the sequencing of the genome of Dr James Watson cost less than $1 million.[3] It is obvious that since the turn of the millennium, there has been a tremendous improvement in sequencing technology in terms of automation, high-throughput nature, and lowering the cost. The ultimate dream is to bring the sequencing cost down to $1000 per genome so that the genome of an individual can be sequenced for the purpose of personalized medicine and personalized nutrition.

Essentially, all NGS platforms discussed below utilize the following steps: DNA (sequencing) library preparation, immobilization of library fragments on a solid support, amplification of the fragments, massively parallel sequencing of the fragments, and computer-aided assembly of the sequence[a]. In this process, each nucleotide base incorporated is detected by a "wash-and-scan" method; millions of reactions are imaged per run to achieve the massively parallel sequencing; each read length is short. A DNA-sequencing library for use in NGS platforms is a collection of surface-anchored single-stranded fragments. The preparation of the sequencing library is a crucial step. *Therefore, the NGS technology does not need the DNA fragments to be cloned for sequencing*. Three popular NGS platforms discussed below are **Roche 454**, **Illumina Solexa**, and **ABI SOLiD**. All these technologies directly read the sequence of individual fragments without the need for cloning the fragments.

3.4.1 Roche 454

Roche 454 was the first NGS platform, introduced in the market in 2005. It is a high-throughput, large-scale, parallel pyrosequencing system. The 454 GS-FLX + system can sequence roughly 0.7 gigabases ($1\ Gb = 10^9$ bases) of DNA per run; the run time being 23 hours.[4] The coverage is $10\times$[b]. By 2013, the average read length was 700−800 bases. *These numbers are arbitrary because they keep improving with time*.

The 454 NGS platform represents a single-molecule improvement to standard pyrosequencing. In this technique, the sequencing library is amplified via **emulsion-PCR** (**em-PCR**), while **pyrosequencing chemistry** is used for sequencing the fragments. In em-PCR, a single DNA template molecule is clonally amplified in an oil/water emulsion (Figure 3.2). In brief, the technique comprises the following steps: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) one fragment−one bead complex formation, (3) fragment amplification by em-PCR, (4) purification, and (5) sequencing by synthesis.

The process begins with shattering of a large DNA molecule, such as genomic DNA, into approximately 800−1000-bp-long fragments. These double-stranded DNA (dsDNA) fragments are blunt ended (polished) and end ligated with universal adapters (A and B). These adapters provide priming sequences for both amplification and sequencing. The A/B-adapter-ligated dsDNA fragments are selected using streptavidin−biotin purification discussed before, denatured into single strands, and combined with an excess of micrometer-sized DNA capture beads or in a 1:1 DNA/bead ratio (but not an excess of DNA, in order to ensure generation of monoclonal beads). The surface of these beads carries oligonucleotides complementary to the adapter sequences on the fragment library. Next, the DNA

[a]If a genome is resequenced, the fragment assembly can be performed with the aid of the reference genome, called **reference assembly**. If a genome is sequenced for the first time, its assembly is called **de novo assembly**.

[b]Coverage denotes the number of times a genome (or a target sequence) has been sequenced. Thus, a $10\times$ coverage for a sequenced genome means that the entire genome has been sequenced 10 times over. So, the higher the coverage, the greater is the depth of sequencing (hence the term **deep sequencing**). A high coverage ensures that the base calling is accurate. Coverage (C) = [read length (L) × number of reads (N)]/G (haploid genome length). Thus, if a target sequence of 5000 bp is assembled from 100 reads with an average read length of 300 nucleotides, the coverage is $(300 \times 100)/5000 = 6\times$. Intuitively, a $6\times$ sequence coverage for the genome appears to mean that each base of the genome has been read 6 times over, but in reality that may not be the case because some parts of the genome of higher eukaryotes are not easily amenable to sequencing, such as intronic sequences and highly repeated sequences.

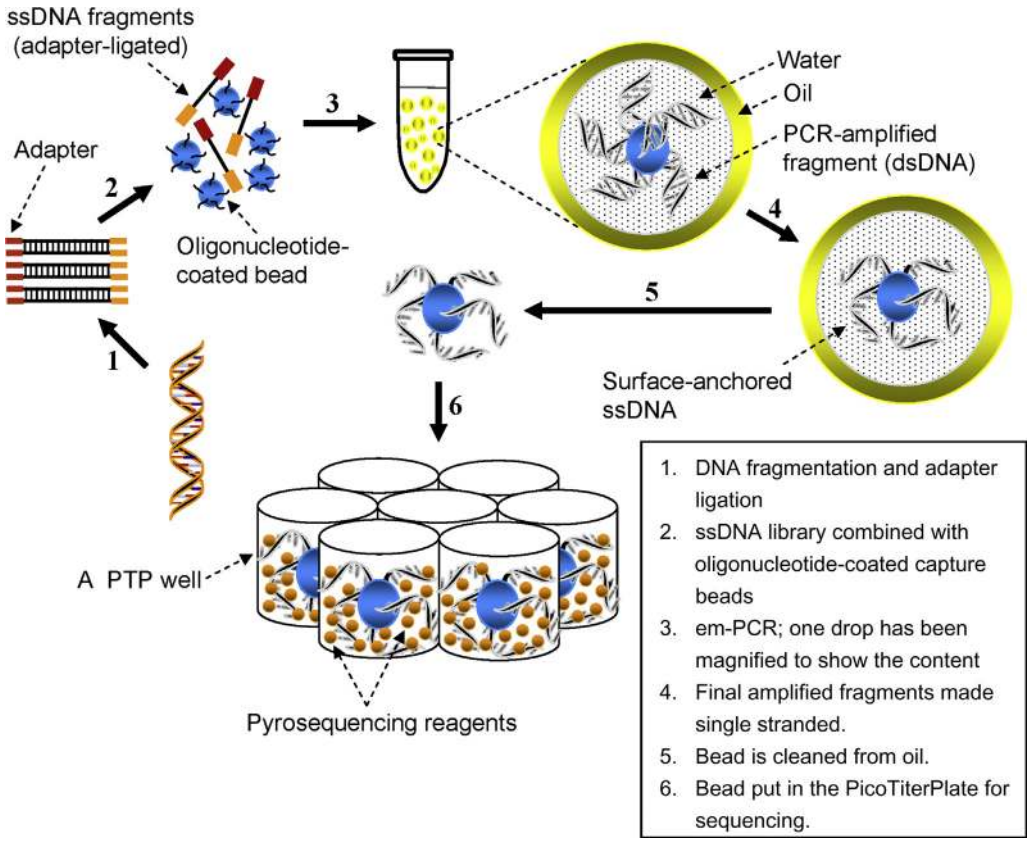


FIGURE 3.2 **Principles of 454 sequencing.** A DNA sequencing library is prepared by ligating adapters to end-polished DNA fragments. Single-stranded (ss) fragments are combined with DNA capture beads containing oligonucleotides complementary to the adapters. The DNA fragments, beads, and PCR reagents are combined within an aqueous mixture, mixed with synthetic oil, and vigorously shaken, which results in the formation of water-in-oil emulsion droplets. Typically, most droplets contain only one bead and one DNA fragment each. The DNA fragment is amplified in emulsion-PCR (em-PCR). The PCR products are purified, denatured, and sequenced in a picotiter plate (PTP) using pyrosequencing chemistry.

fragments, beads, and PCR reagents are combined within an aqueous mixture, which is then mixed with synthetic oil and vigorously shaken. The shaking results in the formation of water-in-oil emulsion droplets (micro-reactors). Typically, most droplets contain only one bead and one DNA fragment each, surrounded by the aqueous layer, which, in turn, is surrounded by the oil layer. The DNA fragment in each droplet is PCR amplified into clonally amplified copies. This PCR process is called **emulsion-PCR** (**em-PCR**). Thus, each bead will bear on its surface PCR products that have been amplified from a single molecule from the template library; these beads are therefore called monoclonal beads. In these bead-immobilized amplicons, the hybridized strand is washed away leaving the beads with surface-anchored single strands.

Next, the beads are screened from the oil and cleaned. The amplified DNA sequencing library, thus generated, is then loaded onto a picotiter plate (PTP) for pyrosequencing. The PTP contains 1.6 million wells; each well is approximately 44 μm in diameter and 75 picoliters in volume.[5,6] Each well can accommodate only a single capture bead. The pyrosequencing reaction mix is also packed into these wells. The PTP is loaded onto an automated pyrosequencing platform, such as the Roche 454 GS-FLX + system, and the DNA fragments are subjected to high-throughput parallel pyrosequencing. The beads that do not contain DNA are eliminated, and the beads that hold more than one type of DNA fragment (polyclonal beads) will be readily filtered out during sequencing signal processing.

3.4.2 Illumina Solexa

Solexa was founded in 1998 in the UK to develop high-throughput sequencing using fluorescently labeled nucleotides and a **sequencing-by-synthesis** approach,

like 454. However, while 454 employs pyrosequencing chemistry for sequencing, Solexa employs fluorescent **reversible terminator chemistry**[c]. The first Solexa sequencer (Genome Analyzer) was introduced in 2006, and could sequence 1 Gb in a single run. In 2007, Illumina acquired Solexa, and by 2011 this sequencing capability had increased to 600 Gb in a single run.[7] The coverage is 30×. By 2013, the run time in the HiSeq 2000/2500 platform was 11 days (regular mode) or 2 days (rapid run mode), and the average read length was ~100 bases.[4] *As indicated earlier, these numbers are arbitrary because they keep improving with time*. The main steps in the Solexa technology are the following: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) addition to flow-cell channels, (3) bridge amplification, (4) cluster generation, and (5) sequencing by synthesis.

For DNA-sequencing library preparation, long DNA is randomly fragmented by ultrasonication; fragments are blunt ended and adapter ligated at both ends. The adapter-ligated fragments are size selected for a length of 250−350 bp, and subjected to small-cycle (10−15 cycles) PCR to increase the yield, which is verified by gel analysis. The desired fragment size pool is isolated and used as the source of the DNA-sequencing library. The dsDNA fragments are denatured and added to the flow-cell channels. The flow-cell channels already contain surface-anchored oligonucleotide primers that immobilize these single-stranded fragments by hybridizing to the adapters. The next step is cluster generation. First, the immobilized fragments are subjected to standard PCR amplification so that many copies of the original fragment are produced and localized in a tight cluster. The double-stranded PCR products in the cluster are denatured and the original strands (hybridized to the surface-anchored primers providing the template for amplification) are washed away leaving the newly synthesized strands, which are now surface anchored. These surface-anchored single strands flip over to hybridize with their nearest surface-anchored primers, forming a bridge-like appearance. Polymerase in the PCR mix extends the hybridized primer, forming a double-stranded bridge. This process of PCR amplification is called **bridge amplification**. When the double-stranded bridge is denatured, two single-stranded molecules are obtained, each of which is now surface anchored. The bridge amplification PCR cycles are repeated to obtain dense clusters of amplified single-stranded products. In this way, several million dense clusters are generated in each channel of the flow cell. These initial clusters have both forward and reverse strand clusters. Next, the reverse strands are cleaved and washed away, leaving the forward strand clusters (Figure 3.3).

The strands are then sequenced using sequencing primers. The first sequencing cycle is initiated by adding all four fluorescently labeled reversible terminator bases (each base contains a different fluorophore), sequencing primers, and DNA polymerase to the flow cell. The polymerase can perform only single base extension; thus, only the base complementary to the template strand is incorporated and the extension stops because of the blocked 3′-end of the added base. Next, the unincorporated bases are removed and the added base is subjected to laser excitation. Following laser excitation, the emitted fluorescence is captured by a CCD camera. Thus, the first base is imaged. The first base of each fragment is similarly recorded and imaged. Then the fluorophore and the terminal 3′-OH end block of the first base are chemically removed, allowing the second cycle to take place. In a similar fashion, the second base added is imaged for all fragments. The cycle is repeated to determine the sequence of bases in each fragment, one base at time. The sequence is assembled by computer software using a reference genome (**reference assembly**). If there is no reference genome and the sequence is new, the sequence assembly is done by the **de novo assembly** method. To score SNPs, the sequence obtained is aligned and compared to a reference (e.g. reference genome) and sequence differences are identified.

3.4.3 ABI SOLiD

Applied Biosystems commercialized its SOLiD platform in 2008. The acronym SOLiD stands for **sequencing by oligonucleotide ligation and detection**. Unlike the 454 and Solexa platforms that use a sequencing-by-synthesis approach, the SOLiD platform uses a **sequencing-by-ligation** approach, and employs sequencing-by-ligation chemistry for sequencing.

Most recent SOLiD platforms, such as the SOLiD 4 system, produce 80−100 Gb of usable DNA data per

[c]In reversible terminator chemistry, each of the four types of dNTPs is labeled with a unique removable fluorophore at the base. Additionally, the 3′-OH end is chemically blocked, but the 5′-PO_4 end is free. After the fluorophore-conjugated dNTP is incorporated by DNA polymerase into the DNA chain, the fluorescence image of the fluorophore is captured using laser excitation. Next, the fluorophore and the 3′-OH block are chemically removed. The resulting 3′-OH end of the newly incorporated dNTP is ready to accept the next incoming nucleotide. This cycle is repeated.

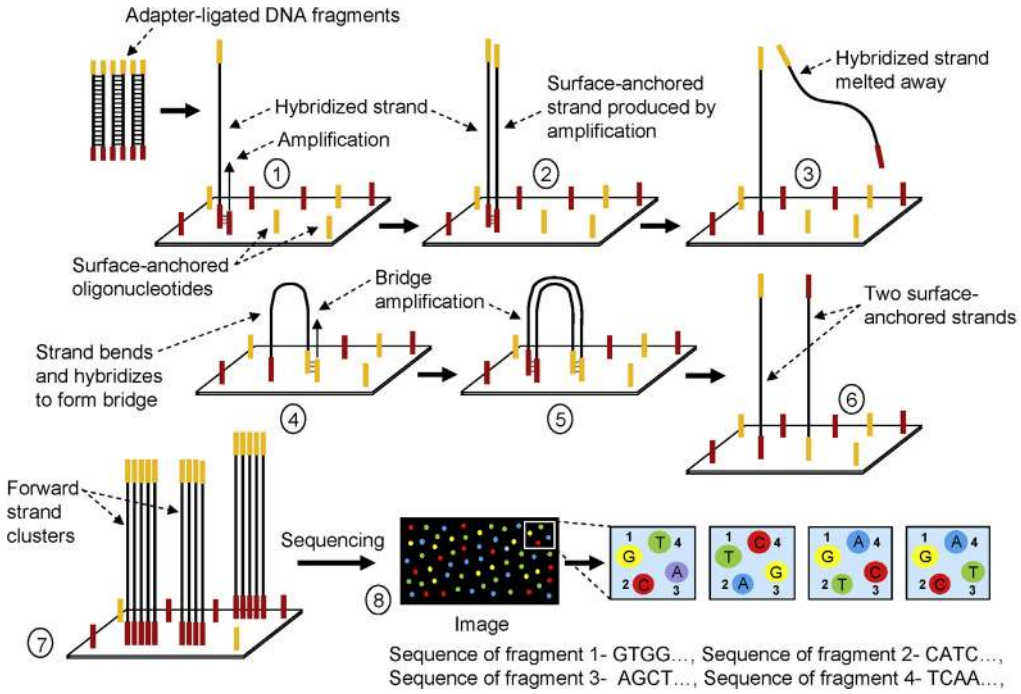


FIGURE 3.3 **Principles of Illumina Solexa sequencing.** The DNA-sequencing library is prepared by ligating adapters to the end-polished DNA fragments. The single-stranded fragments are allowed to hybridize with surface-anchored oligonucleotides that are complementary to the adapters. Initial PCR amplification of the strands followed by bridge (PCR) amplification results in the generation of single-stranded clusters. The strands are then sequenced using fluorescent reversible terminator chemistry (see text for details).

run.[8] The coverage is 30×. By 2013, the average read length of SOLiD sequencing was ~50 bases. *As indicated above, these numbers are arbitrary because they keep improving with time*. In brief, the technique comprises the following steps: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) one fragment−one bead complex formation, (3) fragment amplification by em-PCR, (4) purification, (5) bead immobilization on glass slide, and (6) sequencing by ligation.

The sequencing library preparation for SOLiD sequencing involves shearing of large DNA molecules into 400−600-bp fragments. The fragments are end repaired, adapter ligated, and immobilized on paramagnetic beads. The dilution and anchoring process ensures that only one template per location is tethered. The fragments on the beads are amplified by em-PCR, the beads with extended templates are separated out from undesired beads, the extended templates on the beads are 3′-end modified, and then the beads are immobilized on a glass slide.

The sequencing-by-ligation chemistry utilizes a di-base (two-base) query system for interrogating the sequence and a fluorescent dye for detection. This is also known as **two-base encoding**. The system uses four fluorescent dyes to interrogate all sixteen (4^2) possible two-base combinations. This system utilizes a number of probes; each probe is eight nucleotides (nt) long (8-mer), in which the first two bases at the 5′-end represent the unique two-base combination, and the fluorophore is at the 3′-end. The process begins when a sequencing primer is allowed to hybridize with the universal adapter. Next, a probe that contains the two-base combination complementary to the two bases immediately 3′ to the adapter hybridizes. The base pairing results in the ligation of the 8-mer to the sequencing primer, thereby extending the sequencing primer. The ligation step is followed by fluorescence detection and base calling. Next, a regeneration step removes three 3′ bases from the ligated 8-mer (including the fluorescent group). This prepares the extended primer for another round of ligation. This process is repeated until a specific read length is achieved. Then this extended hybridized sequence is melted away, and the process is repeated with new 8-mers (primer reset) (Figure 3.4).

There are even fully automated benchtop versions of these sequencing instruments available, such as the 454 GS Junior of Roche, MiSeq of Illumina, and Ion Personal Genome Machine and Ion Proton, both of Life Technologies (discussed below).

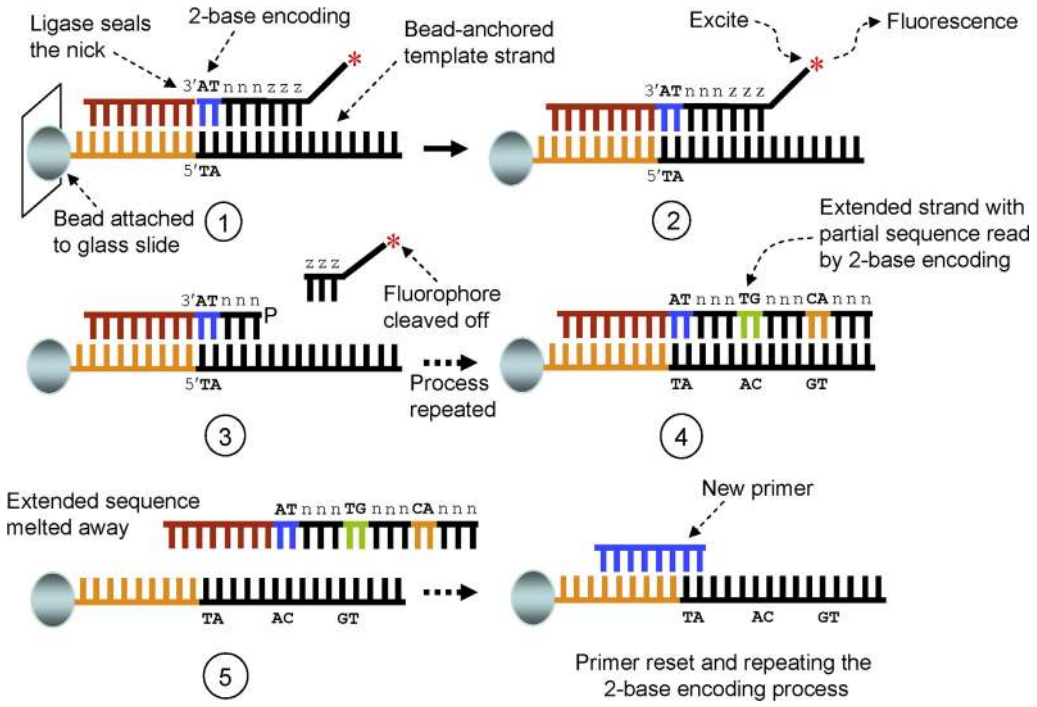


FIGURE 3.4 **Principles of SOLiD sequencing.** The DNA-sequencing library is prepared by ligating adapters to the end-polished DNA fragments, and immobilized on paramagnetic beads. The dilution and anchoring process ensures that only one template per location is tethered. The fragments on the beads are amplified by em-PCR, the extended templates on the beads are 3′-end modified, and the beads are immobilized on a glass slide. The sequencing-by-ligation chemistry utilizes a two-base encoding query system for interrogating the sequence and a fluorescent dye for detection (see text for details).

3.5 NEXT-NEXT-GENERATION SEQUENCING TECHNOLOGY

The invention of DNA sequencing technology was pioneered by Fred Sanger in the UK, and by Alan Maxam and Walter Gilbert in the USA. Sanger's dideoxy-chain-termination method ultimately became the sequencing method of choice because it was technically easier to perform and could be scaled up. These methods are popularly referred to as **first-generation sequencing technology**. The read lengths of these methods are typically 600−800 bp, but could be longer. **The original human genome sequencing project largely relied on the automated and scaled-up version of first-generation sequencing technology**. The main drawbacks of first-generation sequencing technology are the slow progress, because only a small amount of DNA could be sequenced per unit time (low throughput), and high cost (cost per base sequenced).

The introduction of **second-generation sequencing technology** (also known as **next-generation sequencing technology**), three popular platforms of which are discussed above, was an attempt to solve the two major problems of first-generation sequencing technology—that is, to introduce high-throughput sequencing technology for a lower cost of sequencing. However, the second-generation sequencing technology platforms have their own technical problems; for example, a PCR-generated DNA-sequencing library may have PCR-introduced bias and errors, fluorescent nucleotide labeling is not fully efficient, exonucleases are inefficient with labeled nucleotides, detection of single-molecule fluorescence has a high error rate because of the inherent noise in a fluorescence-driven base call, and the same strand can not be "re-read." The noise is due to the fact that the base addition is $< 100\%$ efficient; as a result, as the number of incorporation cycles increases, the population of molecules becomes asynchronous, which results in errors in sequencing read. Although the very high-throughput nature of these methods tends to alleviate some of these problems, the future goal is to develop next-next-generation sequencing technology that will be more efficient and free from the technical problems encountered in second-generation sequencing technology.

Next-next-generation sequencing technology[9] is **third-generation sequencing technology**, although the boundary between the second-generation and third-generation technologies may not be distinct. Ideal desired features of the true third-generation sequencing

technology will probably include the following: single-molecule sequencing technology, no PCR amplification, less complex sample preparation, no pausing of sequencing after each base incorporation (hence increase in sequencing rate), increased read length, and decreased cost. Some of the currently available sequencing technologies that are at the border between the current second-generation and the futuristic third-generation include Life Technologies' **Ion Torrent** semiconductor sequencer that employs a sequencing-by-synthesis approach and uses **pH change** (from the released hydrogen ion during the polymerization of nucleotides) to detect nucleotide incorporation, and **Helicose's** Genetic Analysis Platform that employs a sequencing-by-synthesis approach of a **single molecule** using a defined primer and works by imaging individual DNA molecules as they are extended. The Ion Torrent workflow involves generation of the sequencing library, amplification of the library fragments onto proprietary Ion Sphere particles by em-PCR, deposition of the Ion Sphere particles coated with template in the Ion chip, and sequencing. The average read length is up to 200 bases.

The only truly third-generation sequencing approach so far introduced seems to be the **single-molecule real-time** (**SMRT**) sequencing technology developed by Pacific Biosciences (PacBio). It employs a sequencing-by-synthesis approach and allows for direct observation of the synthesis of a single strand of DNA by DNA polymerase in real time. The SMRT technology of PacBio utilizes what is called a **zero-mode waveguide** (ZMW). A ZMW is a hole, tens of nanometers in diameter, fabricated in a 100-nm metal film deposited on a glass substrate. An active polymerase is immobilized at the bottom of each ZMW chamber. The ZMW, being so small, prevents visible laser light from passing entirely through it; the laser exponentially decays as it enters the ZMW. Because of this property, a laser passed through the glass into the ZMW only illuminates the bottom 30 nm of the ZMW chamber. Nucleotides are allowed to diffuse into the ZMW chamber; each base is labeled with a different fluorescent dye. The incorporated base can be recognized based on the fluorescence emission, which happens within the illuminated section of the nanochamber, and the synthesis of a single DNA molecule is directly recorded.[10] In this method, the same DNA molecule can be resequenced by creating a circular DNA template and separating the newly synthesized DNA strand from the template. In the PacBio *RS* platform, the average read length is about 3000 bases and the run time is very short, about 20 min[4] Various other approaches are being tested, such as **transmission electron microscopy** to directly image single DNA molecules, and a **nanopore-based** single-molecule sequencing approach. The sequencing community has been eagerly waiting to get their hands on third-generation sequencing technology.

3.6 HIGH-DENSITY OLIGONUCLEOTIDE-PROBE-BASED ARRAY TO INVESTIGATE GENOME EXPRESSION

Microarray and global gene-expression profiling is a crucial genomic technology. The term **microarray** is often used synonymously with DNA microarray and high-throughput gene-expression measurement. However, it can also be used in the context of expression profiling of proteins, carbohydrates, and tissues. The current discussion on microarray will focus on gene expression. Gene-expression microarray is a nucleic-acid-hybridization-based technique. Studies on nucleic-acid hybridization were pioneered independently by Paul Doty and Sol Spiegelman and their colleagues. The DNA−RNA hybridization principles were utilized to develop a number of widely used techniques to study gene expression, such as in situ hybridization, Northern blot, and solution hybridization.[11] These techniques mostly measure the expression of a single gene in multiple tissues and at multiple time points. Before the advent of genomics, a number of techniques were also developed to analyze differential gene-expression profiles, involving a large number of samples, multiple target sequences (a large number of transcripts), and many tissues at the same time; for example, ribonuclease (RNAse) protection assay (RPA), subtractive hybridization, differential display, serial analysis of gene expression (SAGE), and branched DNA (bDNA) signal amplification technique.[12]

However, global gene-expression profiling was revolutionized with the advent of the microarray. In 1996, Affymetrix commercialized its oligonucleotide-based DNA chip under the proprietary name GeneChip®. A microarray can be either a complementary DNA (cDNA) microarray or an oligonucleotide microarray. Currently, high-density oligonucleotide microarray is the method of choice. In an oligonucleotide microarray, an array of oligonucleotide probes (usually 20−80-mer) are synthesized either on-chip (on the platform) or by conventional synthesis followed by immobilization on the platform. An example of on-chip synthesis of oligonucleotides is the photolithographic technique, which is used by Affymetrix (Figure 3.5A). Another related technology uses an ink jet to spray oligonucleotide probes on the microarray. The fabrication of an oligonucleotide array is carried out by high-speed robotics. These robots rely on pins or needles to transfer the sample from a reservoir to the platform. The

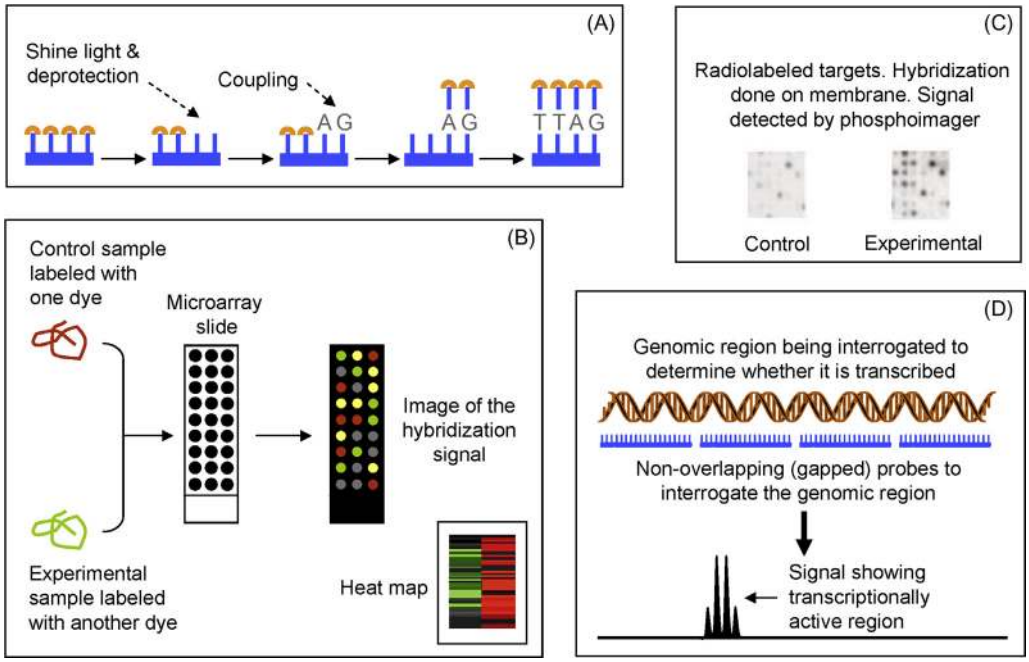


FIGURE 3.5 **High-density oligonucleotide-based array.** (A) Microarray fabrication by photolithographic synthesis, which involves repeated cycles of targeted deprotection, coupling, and protection of the coupled bases. (B) Microarray using fluorescent-dye-labeled targets and competitive hybridization of the two probes on the same array slide. The inset shows what a heat map could look like. (C) Microarray using radiolabeled targets. (D) Use of tiling array to identify a genomic region that was previously not known to be transcriptionally active.

pin diameter and shape, solution viscosity, and platform characteristics determine the volume transferred and how far the solution will spread. The number of spots on the microarray can vary between a few thousand to 30,000 on a 25×75-mm slide, each spot representing the product of a specific gene, and is generated by depositing between 1 and 10 nl ($1\ \text{nl} = 10^{-3}\ \mu\text{l}$) of PCR product representing that specific gene, usually at concentration of 100–500 μg/ml. The spot diameter can be between 75 and 200 μm, and the distance between spots is about 200 μm.[11] In a cDNA microarray format, customized cDNA probes are immobilized on a solid surface (glass or nylon membrane). The DNA fragments can be PCR amplified or be library clones. Thus, the array density is lower than in DNA chip, and the spotted cDNAs are longer than oligonucleotide probes.

To detect gene expression, the microarray is hybridized with the labeled **target**, which is the reverse-transcribed copy of the mRNA. The mRNA-derived cDNA is labeled, in most cases by fluorescent dyes, such as Cy3 and Cy5. Purified poly(A)$^+$ mRNA is usually recommended as the starting material for improving the signal/noise ratio—that is, for increased sensitivity and low background. Hybridization spots containing fluorescent dyes are detected by laser scanning of the microarray. The laser scanner is hooked to a confocal microscope and a CCD camera. The fluorescent tags are excited by the laser, while the microscope and the camera work together to create a digital image of the array. The results are then analyzed using special analysis software (Figure 3.5B).

For cDNA microarrays spotted on nylon membrane, the target cDNA population is radioactively labeled. Radiolabeled hybridization spots can be detected and analyzed by a phosphoimager (Figure 3.5C). Differences in the expression of specific sequences can be further validated using other conventional methods, such as Northern blot, reverse transcriptase-polymerase chain reaction (RT-PCR), RNAse protection assay, or bDNA assay.

Microarray data can be transformed into a colored graphical representation, the so-called **heat map** (Figure 3.5B inset). In the heat map, increased expression is displayed by the intensity of a certain color (such as red), whereas decreased expression is displayed with another color (such as green), and a third color (black, the absence of other colors) may represent no changes in expression pattern.

3.6.1 Tiling Array as a Versatile Tool to Interrogate the Whole Genome

A tiling array is an oligonucleotide-based whole-genome microarray, and has proved to be very useful for whole-genome functional analysis beyond simple

gene-expression profiling. Because the tiling array is a variation of the microarray, it is conducted in the same way as a regular expression microarray, the main difference being the probe design. Tiling arrays probe for known contiguous sequences, such as a genomic region whose expression is not known. The resolution power of tiling arrays depends on the probe design—that is, whether the probes are spaced apart (gapped) or overlapping.

Whole-genome tiling arrays can be used for the interrogation of genomic regions for transcription, antisense transcription, and alternative splicing; interrogation of transcription-factor-binding sites and genomic polymorphism, and mapping of genomic methylation sites; and comparative genomic hybridization (CGH).[13,14] Figure 3.5D shows just one application of the tiling array, how a tiling array can be used to detect a region of the genome that was not previously known to be transcriptionally active. Tiling arrays designed to detect SNPs utilize overlapping probes so that every base is interrogated for mutation. The number of oligoprobes used in a whole-genome tiling array can be many millions. For example, in order to comprehensively identify coding sequences in the human genome, Bertone et al.[15] used genome tiling arrays by designing about 52 million oligoprobes (36-nt long) positioned every 46 nt, on average. These probes cover 1.5 Gb of nonrepetitive genomic DNA, both sense and antisense strands.

Tiling array platforms are designed and fabricated in the same way as the regular expression microarray platforms described above.

3.7 GENOME-WIDE MUTAGENESIS, GENOME EDITING, AND INTERFERENCE OF GENOME EXPRESSION

The best way to study the function of a gene is to silence its expression and analyze the resulting phenotype. The principal method of silencing the expression of a gene is **gene targeting** (**gene knockout**) by homologous recombination in embryonic stem (ES) cells. Using homologous recombination, a specific genetic locus can be disrupted (knockout) or replaced with another functional open reading frame (ORF) (knock-in) in ES cells of mice. By replacing the endogenous mouse gene with a human ortholog, a humanized mouse model can also be produced. The targeting construct contains an expression cassette that is flanked by two long stretches of genomic DNA. These two stretches of genomic DNA, called **homology arms**, have the same sequence as that of the genomic DNA flanking the target locus. Thus, the homology arms facilitate recombination and integration of the construct into the locus, thereby disrupting the endogenous ORF (Figure 3.6). The gene-targeting technique is limited to the generation of mouse models because it requires knowledge of the ES cells in which the targeting is done to mutate the gene. Currently, the biology of mouse ES cells is well understood. As a result, gene knockout models are mouse models, and this technique cannot be routinely performed in other animal models.

The only organism where systematic targeting of a vast number (96%) of the annotated ORFs has been

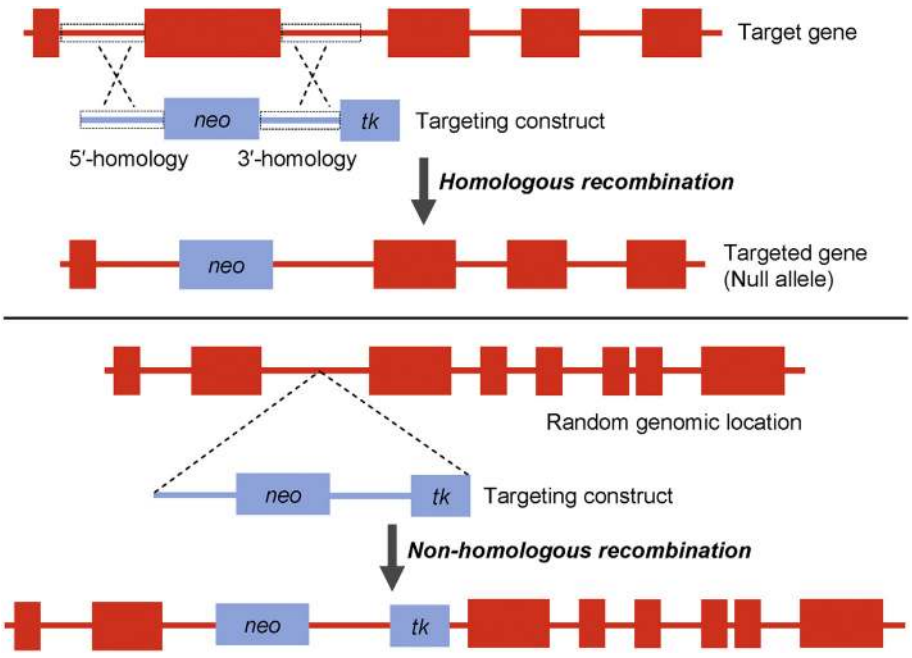


FIGURE 3.6 **Gene targeting.** The upper panel shows the generation of a null allele through gene targeting. The targeting construct is integrated through homologous recombination, which has a low frequency. In homologous recombination, the thymidine kinase (*tk*) gene, which is a negative selection marker, is not integrated. Only the *neo* gene, which is the positive selection marker, is integrated through legitimate recombination. The lower panel shows the random integration of the entire targeting construct by nonhomologous recombination, which has a higher frequency than homologous recombination.

achieved is yeast (*Saccharomyces cerevisiae*).[16] Each ORF was precisely targeted and replaced by mitotic recombination with the *KanMX* targeting cassette. The *KanMX* gene (which confers kanamycin resistance) in each cassette is flanked by yeast sequence that facilitates recombination and integration of the cassette in the yeast genome; in addition to the yeast sequence, the *KanMX* gene is also flanked by two distinct 20-nt sequences that serve as molecular barcodes to uniquely identify each deletion mutant.

Such an achievement could be a reality even for the mouse a few years from now. The International Knockout Mouse Consortium (IKMC) has been working to mutate all protein-coding genes in the mouse using a combination of gene trapping and gene targeting in C57BL/6 mouse ES cells.[17] **Gene trapping** is an insertional mutagenesis technique that randomly generates ES cells with well-characterized mutations. A gene-trap vector construct, called the **trap cassette**, contains a promoterless reporter cassette (such as *lacZ*). There is an upstream splice acceptor site and a downstream poly(A) sequence in the trap cassette. The splice acceptor sequence is not bypassed by the RNA-splicing machinery. The trap cassette reporter is used to identify the ES cells where the gene-trap construct is integrated. The gene-trap construct can be electroporated into the ES cells, or delivered using a retroviral vector. In some ES cells, the construct will be correctly integrated in an intron to produce incorrect splicing of the target gene, such that all exons downstream of the insertion site are not expressed. The endogenous functional promoter of the target gene will drive transcription producing fusion transcripts. The fusion protein translated from the fusion transcript provides a means of rapid identification of the disrupted gene. The targeted gene is identified by sequencing of the transcribed product. Figure 3.7 shows the gene-trap technique.

The limitations of classical gene targeting could soon be overcome by **zinc-finger nuclease** (**ZFN**) or **TAL effector nuclease** (**TALEN**) technology. A Zn finger is a small protein structural motif that has a Zn ion in a coordination complex with either four cysteines (Cys_4) or two cysteines and two histidines (Cys_2His_2) to stabilize the so-called finger-like fold (Figure 3.8 inset). A large class of transcription factors containing a Zn finger bind to the major groove of DNA through their Zn-finger DNA-binding domains; each domain actually recognizes a specific trinucleotide sequence in the DNA. A ZFN is an engineered synthetic protein that consists of an engineered Zn-finger DNA-binding domain fused to the cleavage domain of the *Fok*I restriction endonuclease. *Fok*I is a type IIS restriction endonuclease. Type IIS restriction endonucleases cleave the DNA outside of the recognition sequence, to one side. *Fok*I recognizes an asymmetric nucleotide sequence and cleaves one strand 9 nt downstream and the other strand 13 nt upstream of the recognition site, as follows: 5′-GGATG$(N)_9$▼-3′/ 3′-CCTAC$(N)_{13}$▲-5′. The *Fok*I cleavage domain induces double-strand breaks (DSBs) in specific DNA sequences, which triggers DNA repair. Eukaryotic cells repair

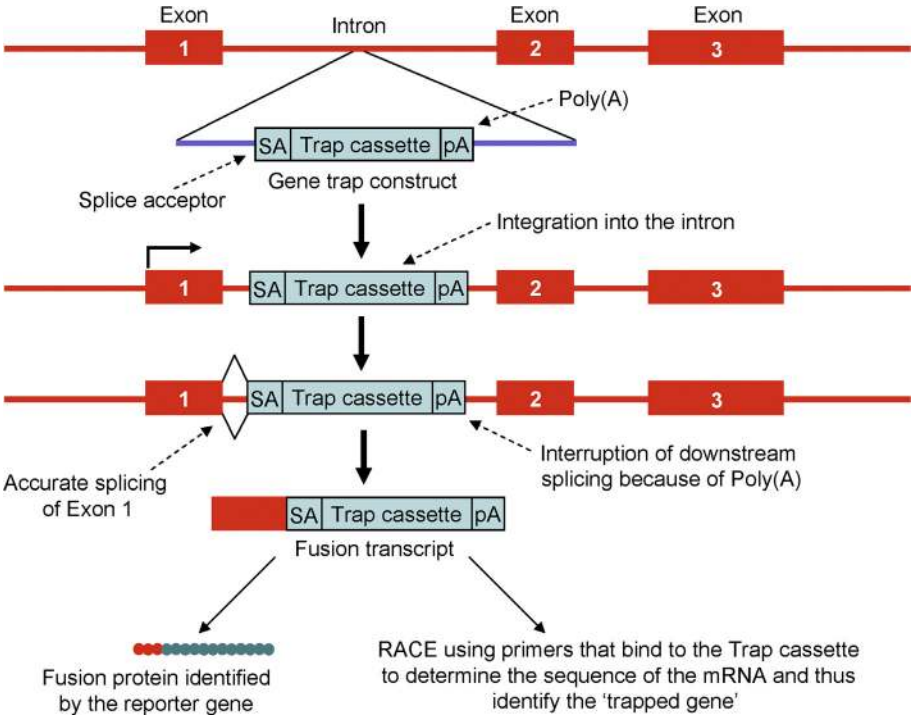


FIGURE 3.7 **Gene trapping is an insertional mutagenesis technique.** Random insertion of the trap cassette in the genome generates ES cells with well-characterized mutations. The trap cassette reporter is used to identify the ES cells where the gene-trap construct is integrated. Rapid amplification of cDNA ends (RACE) using trap-cassette-specific primers is employed to identify the trapped genes in the ES cells. Where the construct is correctly integrated into an intron, this produces incorrect splicing of the target gene, such that all exons downstream of the insertion site are not expressed.

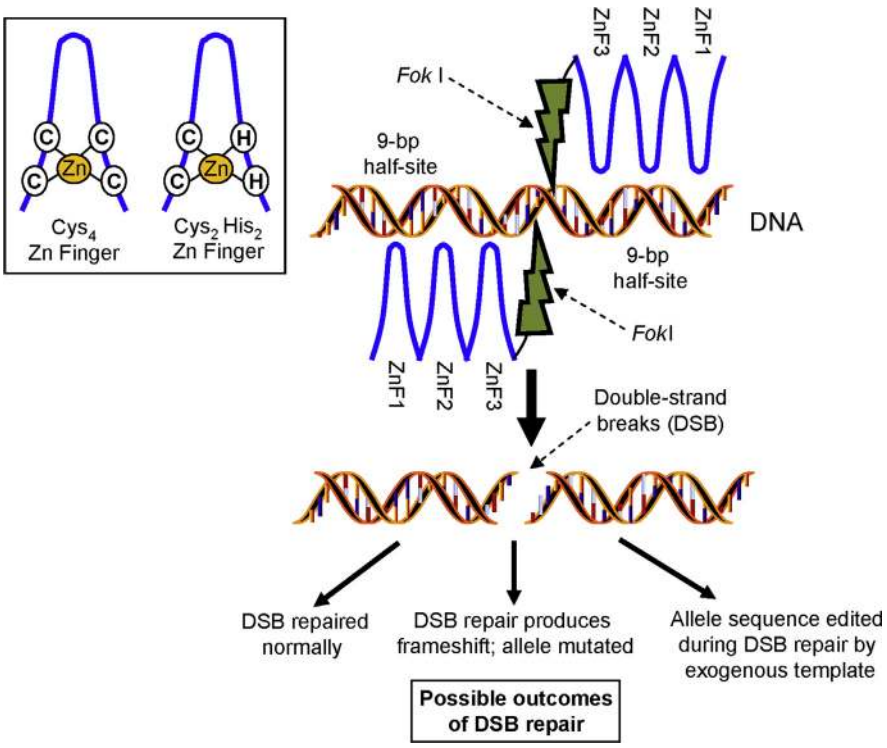


FIGURE 3.8 **Gene and genome manipulation using Zn-finger nuclease.** The figure shows a pair of ZFNs bound to their target site. Three Zn-finger domains are marked ZnF1, 2, and 3. Each three-finger array binds to a 9-bp half-site and is associated with a *Fok*I nuclease domain. A ZFN pair cleaves its target site within the variable-length spacer sequence between the half-sites. There are three possible outcomes of the DSB repair. The inset shows two types of Zn-finger motifs, a Cys_4 and a Cys_2His_2 motif.

DSBs using homology-directed repair (HDR) or non-homologous end-joining (NHEJ) pathways, and these repair pathways can be utilized to edit the genome. For example, by providing template (homologous) donor DNA along with ZFNs for HDR, information encoded on the introduced template can be used to repair the DSB, and in that process some nucleotides can be changed (gene editing including correction), or it is even possible to add a new gene at the site of the break. The NHEJ repair pathway ligates the two broken ends, with occasional small insertions or deletions at the site of the break, resulting in frame-shift and disruption of the target gene (Figure 3.8). Thus, the genome-editing function of ZFNs is based on the introduction of site-specific DNA DSBs into the locus of interest. By fusing *Fok*I to different types of Zn fingers that recognize different trinucleotide sequences, the ZFNs can be targeted to different parts of the genome for desired genome editing. ZFN technology has been successfully used to manipulate the genomes of many plant and animal species.

One of the major achievements of ZFN technology has been the generation of gene knockout models in species other than mice, which was not possible using the standard gene-targeting technique. By microinjection of ZFNs designed to target an integrated reporter and two endogenous rat genes, immunoglobulin M (*IgM*) and *Rab38*, in a one-cell rat embryo, successful gene targeting was reported. A high frequency of animals had 25 to 100% disruption at the target loci and these mutations were faithfully and efficiently transmitted through the germline. Transcription-activator-like effector nuclease (TALEN) technology is similar to ZFN technology. The main difference is in the DNA-targeting protein, which is the TAL effector (TALE) protein. The TALE protein can be fused to *Fok*I to generate the TALEN. Unlike ZFN and TALEN that are protein-guided genome editing tools, CRISPR-Cas system is a RNA-guided genome editing tool. CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, and Cas is CRISPR-associated nuclease. Target recognition by Cas nuclease requires a ''seed sequence'' within CRISPR RNA (crRNA) that acts as a guide to Cas. Thus, almost any DNA sequence can be targeted by redesigning the crRNA seed sequence. In prokaryotes, the CRISPR-Cas system acts as RNA interference (RNAi, discussed in the following section) based immune system to defend against invading viral DNA because the short crRNAs that guide the recognition of targets for degradation are produced by the processing of a long transcript.[18]

RNA interference (RNAi) is another way of **knocking down** (instead of **knocking out**) genome expression and studying the phenotype. In *Caenorhabditis elegans*, the effect of silencing gene expression on a large scale has been studied by multiple groups, who were able to study about a third of the predicted genes. Using a reusable RNAi library of 16,757 bacterial clones, Kamath et al.[19] were able to knock down the expression of about 86% of the 19,427 predicted genes. Each bacterial strain in the library was capable of expressing dsRNA

designed to correspond to a single gene. Mutant phenotypes for 1722 genes were identified; about two-thirds of these were not previously associated with a phenotype. Such genome-wide RNAi analysis has also been accomplished in *Drosophila*.[20] The authors applied an RNAi screen of 19,470 dsRNAs in cultured cells to characterize the function of nearly 91% of predicted *Drosophila* genes in cell growth and viability. Interestingly, the authors found 438 dsRNAs that identified essential genes, among which 80% lacked mutant alleles.

3.8 SPECIAL TOPIC: OPTICAL MAPPING OF DNA

Michael L. Kotewicz, Ph.D., Office of Applied Research and Safety Assessment, CFSAN, FDA

3.8.1 Introduction

In chromosomes, which range from 1–6 million bp in bacteria to 100 million bp in humans, what graphic software tools allow one to locate and distinguish details as small as single nucleotide polymorphisms, mid-sized chromosomal changes (10,000–200,000 bp), and inversions across millions of base pairs? No graphic tool, to date, performs ideally at both these extremes. One software tool well suited for the fine-scale mapping of nucleotides and detailed chromosome alignments is **Mauve**.[21] Mauve and the updated **progressiveMauve** are extremely powerful desktop graphic tools for aligning chromosomes and defining both homologous genome segments and single-nucleotide differences. At the opposite scale, the graphic software in **MapSolver™** was designed to work with optical maps of chromosome restriction fragments and *in silico* sequence-based maps of reference bacterial chromosomes. MapSolver's strengths are its easy graphic ability to ramp up and down thousands and millions of base pairs and to detail differences in aligned optical maps and reference *in silico* chromosome maps.

Optical maps are physical maps assembled from overlapping restriction-fragment maps of long chromosomal pieces, and they represent a sample of the sequence across the complete chromosome. For each restriction fragment, the cut site at the beginning of the fragment and the cut site at the end score the presence of these sequence pairs; for example, a *Bam*HI map scores GGATCC pair sets in the chromosome as well as measuring the nucleotide distance between those sequence pairs. The map could be considered a digital chromosome. Within the limits of fragment size measurements, 1–2%, where sets of fragments in a new isolate's optical map align to fragments from a reference sequenced genome, there is a direct correlation of the map fragments with the reference sequences and genes in those fragments. The alignment scores represent the strength of the correlation of map and sequence, where the limit of detection for differences such as insertions and deletions is 1–5 kb in the optical map. The optical mapping software optimally presents a simple graphic, best suited to detect, measure, and display chromosome differences from about 5000 to millions of bp. Differences created by events such as close-proximity multiple prophage insertions can span 300,000 bp and complex multiple inversions can span several million bp. In contrast, Mauve is like a street map, detailed to seeing single nucleotide addresses, and just as one would not use a street map to find continents on a globe, Mauve is not quite as well suited for rapidly determining and viewing these larger chromosomal differences; something that optical maps do extremely well. What optical maps lose in terms of resolution and nucleotide detail, they make up for in ease of use and perspective. It is worth testing a set of alignments in both software packages and comparing the advantages and limitations of each for examining chromosome differences (Figure 3.9). Mauve gives a sequence-based segmental view of compared chromosomes, while MapSolver™ gives a difference-based alignment of restriction fragments. For maps, the sequence information is correlated, albeit indirectly, with sequences in aligned reference fragments.

3.8.2 Optical Maps

Optical maps are physical maps generated from long chromosomal DNA preparations attached and restriction digested on surfaces. For a number of reasons—including G/C content, average fragment size generated for a given genome, and overall number of cuts—optical maps are usually generated using six-base-cutter restriction enzymes, such as *Bam*HI (GGATCC) or *Nco*I (CCATGG), although there is some flexibility in enzyme of choice. In addition to displaying the physical DNA maps, MapSolver™ software is used to generate reference *in silico* maps from sequence data. These annotated reference genomes are used to define the differences found in comparative alignments with optical maps.

There is an additional use for MapSolver™: higher resolution mini-maps, usually generated on shorter DNA sequences ranging from 5000 to 1 million bp using more-frequently cutting restriction enzymes, such as four-base-recognition enzymes. These mini-maps are useful in several regards. One is for comparative genomic studies determining the structures of chromosomal variations. The other is for the rapid display of sequencing misassemblies. Initially, mini-maps were conceived as allowing a more detailed map to be constructed by sub-cutting sites within larger fragments of *in silico*

FIGURE 3.9 **The alignment of the *in silico* optical maps of two related strains of *E. coli* O157:H7: TW14359 from the 2006 US spinach-associated outbreak, and Sakai, the Japanese outbreak associated with sprouts.** (A) Two pairs of aligned maps using MapSolver™, the non-aligned regions of the chromosomes are white, aligned regions are green; in the lower aligned pair, regions of interest have been "painted" from the sequence-based annotations. Prophages are yellow/orange, prophages carrying the Shiga toxin genes are red, and pathogenicity islands are blue. (B) Mauve alignment of the same two sequenced chromosomes, where similarly colored sections reflect sequence matches; note white streaks within colored boxes, indicating short unaligned sequences within larger aligned sequence blocks.

*Bam*HI (GGATCC) maps with *Sau*3AI (GATC) fragmentation. Dr. David Lacher has refined mini-mapping in our laboratory. He noted that *Sau*3AI produces a much more heterogeneous mixture of large and small fragments, and that other four-base cutters such as *Hha*I (GCGC) and even other six-base cutters such as *Hpa*I (GTTAAC) provide a more evenly distributed, higher density set of fragments in these *in silico* mini-maps, especially for *E. coli*. For example, the six *Bam*HI fragments for the 112-kb TW14359 *yehV* prophage region produce a *Sau*3AI mini-map with 344 fragments; *Hha*I produces a much more homogenous set of 725 fragments that yields better coverage of differences. The *Hha*I mini-map of the *yehV* region of TW14359 and Sakai (Figure 3.10) shows the detail of two 1.3-kb insertion/deletions (indels) in the left flanking chromosomal DNA outside the prophages. The mini-map clearly shows two distinctive differences within the two *yehV* prophages, but in addition the mini-map details another 1.3-kb indel, a 12.6-kb region containing Shiga toxin genes in Sakai, and a quite different, unaligned 14.5-kb set of fragments, hence different sequence, in TW14359. The remaining 28 mini-map fragments (7.0 kb) are homologous in the two prophages, delineating the variant Shiga toxin region within otherwise homologous regions.

Optical mapping is a also a corroborative technology for sequencing; it is independent of amplification technologies, and importantly, mistakes in DNA assemblies are readily identified, notably across ribosomal RNA and repeated conserved regions of multiple prophages.[22] It is also a complementary and refining technology for traditional low-resolution pulsed-field gel electrophoresis (PFGE) analysis, the gold standard for bacterial epidemiological identification.[23] A contiguous 600-fragment map locates chromosomal markers, and it greatly exceeds the 40-fragment resolution of PFGE. Most importantly, the optical maps define the contiguous relationships of all the fragments, while PFGE gives no direct band correlation with chromosomal position. Optical mapping accurately identifies both large fragments not resolved by PFGE and small fragments not detected by PFGE (Figure 3.11).

Optical maps are fundamentally shorthand representations of the sequences of chromosomes generated by mapping restriction-enzyme cut sites; they are reflections of whole-chromosome sequences. For a typical bacterial chromosome of 4−6 Mbp, six-base-recognition restriction enzymes such as *Bam*HI (GGATCC) or *Nco*I (CCATGG)—for *Escherichia coli* and *Salmonella enterica* isolates—generate a map with 400−600 contiguous restriction fragments. Changes in genome sequences ablate or create cut sites, creating restriction fragment length polymorphisms (RFLPs). More importantly, differences in chromosomes between related strains generate changes in the sizes and distribution of fragments that light up in aligned maps. Optical mapping allows the rapid construction of ordered restriction fragment maps for chromosomes that can be as small as 150−200-kb bacterial plasmids, but optical maps are optimally suited for detecting differences in chromosomes of bacteria which range from 1−10 million bp. Overall, the 5-Mbp chromosomes of bacteria can be sized to within 10−20 kb, an accuracy of about 0.1 to 0.3%.[24] Whereas single nucleotide polymorphisms are

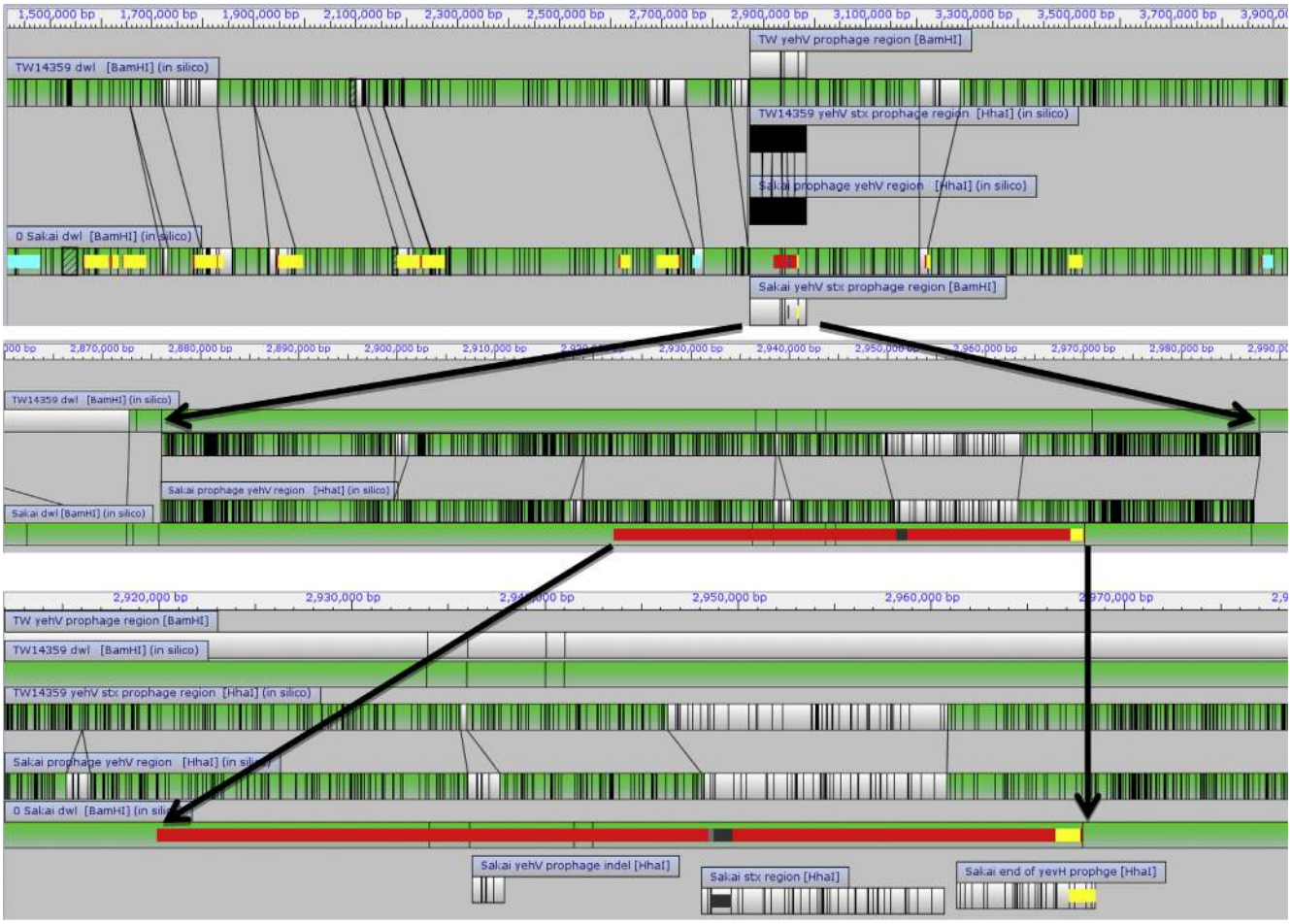


FIGURE 3.10 **Mini-maps: six-base cutter *Bam*HI (GGATCC) versus four-base cutter *Hha*I (GCGC).** Three successively enlarged MapSolver™ views of the *yehV* prophages.

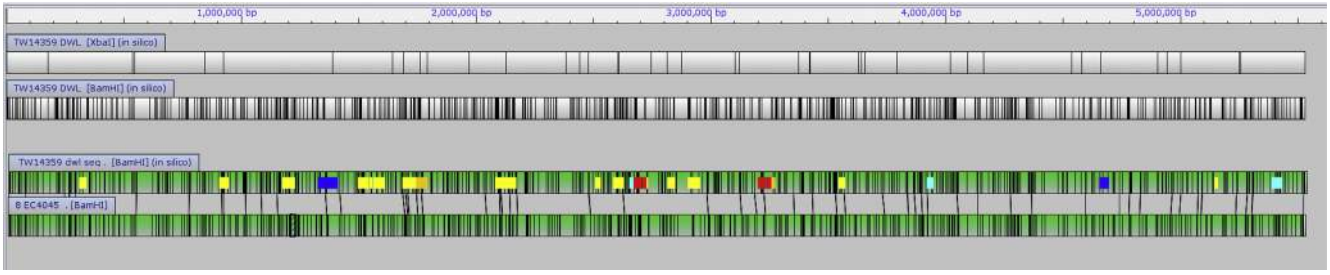


FIGURE 3.11 **Optical limit of detection.** Upper two unaligned maps: *Xba*I (42 fragments) versus *Bam*HI (642 fragments) *in silico* TW14359 maps; lower two maps: aligned painted *in silico* (642 fragments) versus optical map (529 fragments) of spinach-outbreak strain, isolates TW14359 and EC4045. A total of 113 fragment differences are in small fragments, 21 to 1000 bp, at the optical limit of detection.

crucial for differentiating highly clonal *Salmonella* isolates, *Escherichia coli* strains, particularly pathogens such as *E. coli* O157:H7 isolates, differ by prophages and insertions and deletions.[25]

There are two other related technologies for determining the structure of chromosomes with comparable mid to long molecule resolution, one involving fluidic separation of large DNA molecules from Pathogenetix, Woburn, MA, and the other involving nanochannel fluidic chips that spread out confined native long genome fragments labeled at restriction-enzyme-nicked sites with fluorescent tags, from BioNano, San Diego, CA. This discussion is focused on optical mapping using hardware (the Argus mapping station) and software (MapSolver™)

for comparative genomics, from OpGen, Gaithersburg, MD.

There are a number of other nucleotide-based software packages for looking at long regions of DNA molecules, including DNASTAR's Lasergene and a number of retired (2007) Genetics Computer Group (GCG) available within the European Molecular Biology **Open** Software Suite, an open-source software analysis package at EMBOSS (http://helix.nih.gov/Applications/emboss.html). Software packages from next-generation sequencing companies are continually upgrading and although designed and useful for examining sequence contigs (consensus regions of DNA derived from sets of overlapping DNA segments) and assemblies and although not necessarily optimized for comparative genomics, they are moving in that direction.

3.8.3 Overview; Making an Optical Map

Optical maps have been generated for a wide range of bacterial species involved in industrial microbiology, clinical illnesses, and food-borne bacterial outbreaks, as well as for larger chromosomes from fungal and mammalian sources. For a bacterium, the optical map of its chromosome is generated by growing up cells from an isolate or a set of isolates and gently lysing them to release high-molecular-weight DNA (Figure 3.12A). The DNA molecules are loaded into carefully designed microfluidic channels (Figure 3.12B, in this case 40 channels in a 2×2 cm area. DNA molecules attach by charge interactions with the derivatized glass surface and distribute as long linear individual molecules onto the surface (Figure 3.12C).

The attached molecules are digested with an appropriate restriction enzyme and the DNA is stained with

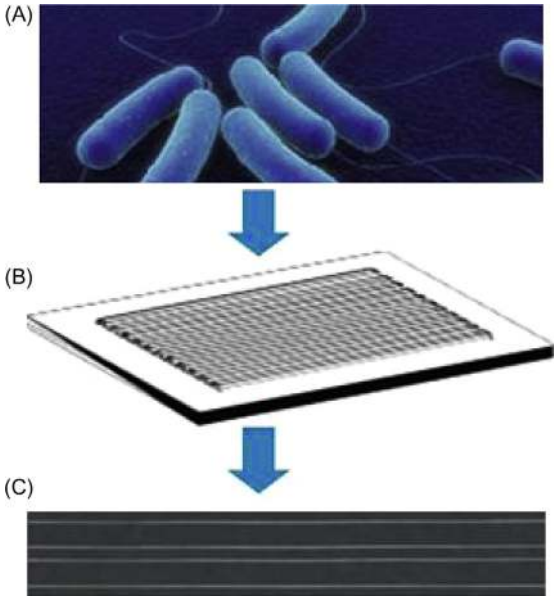


FIGURE 3.12 **Preparation of high-molecular-weight DNA.** (A) Bacterial cells prior to lysis; (B) forty-microfluidic-chamber device on coverslip; (C) DNA in one channel attached to derivatized glass surface.

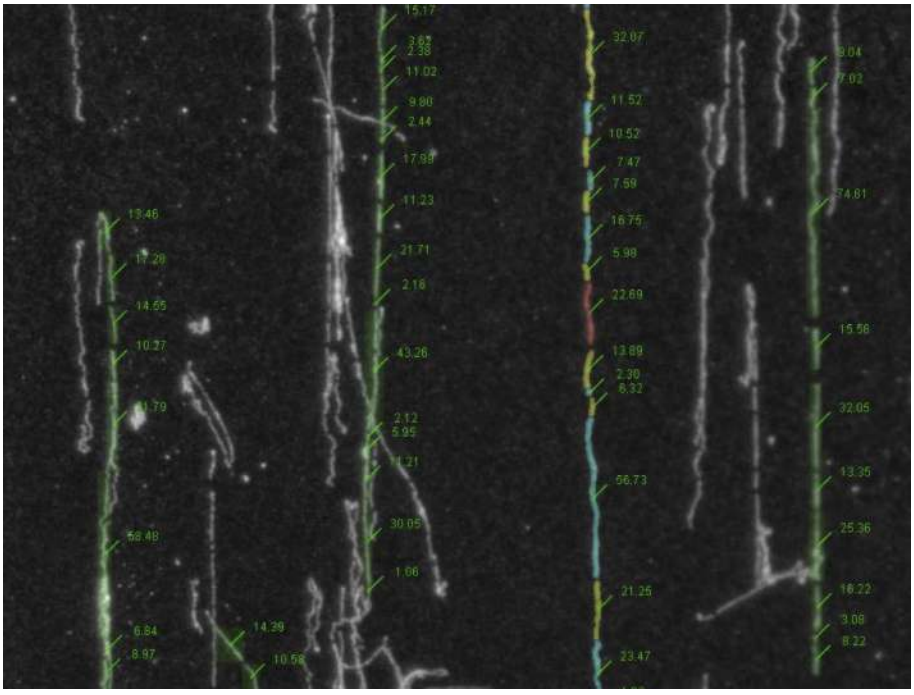


FIGURE 3.13 **Restriction-digested DNA attached to the cover slip surface as seen under the Argus microscope and assembly platform.** The image is from an assembly data set; molecular weights of fragments are indicated. The multicolored strand to the right of the figure center line is a molecule from the assembled map, for examination of details. Note the extent of linearity or wiggle in each restriction fragment and the gap sizes. These are some of the quality control parameters used to judge data sets.

the fluorescent dye JOJO-1. The salt conditions of the wash after staining cause the DNA to constrict a slight amount such that a small measurable gap is created at the cut sites, but the restriction fragments remain attached to the surface. Automated software is used to measure the sizes and positions of contiguous restriction fragments along thousands of chromosome-fragment molecules. Depending on the size of the genome of the organism being mapped, molecules are collected, each containing 10 to 100 contiguous restriction fragments. For example, 2000 to 50,000 molecules are usually collected for analysis for a 1 to 6-million-bp bacterial chromosome. The attached, digested DNA fragments range from 250 to 400 kb; some are as large as 1.5 Mbp. **The limit of detection of fragments is about 500 bp** (Figure 3.13).

The data from these thousands of molecules are assembled into complete genomic maps by overlapping same-sized fragment runs, similar to the assembly of overlapping DNA sequencing runs (Figure 3.14A). In these assemblies, the minimum coverage for each fragment is 30× (Figure 3.14B and C). The completed assemblies are usually oriented to a defined start reference or origin and scaled to a reference sequence.

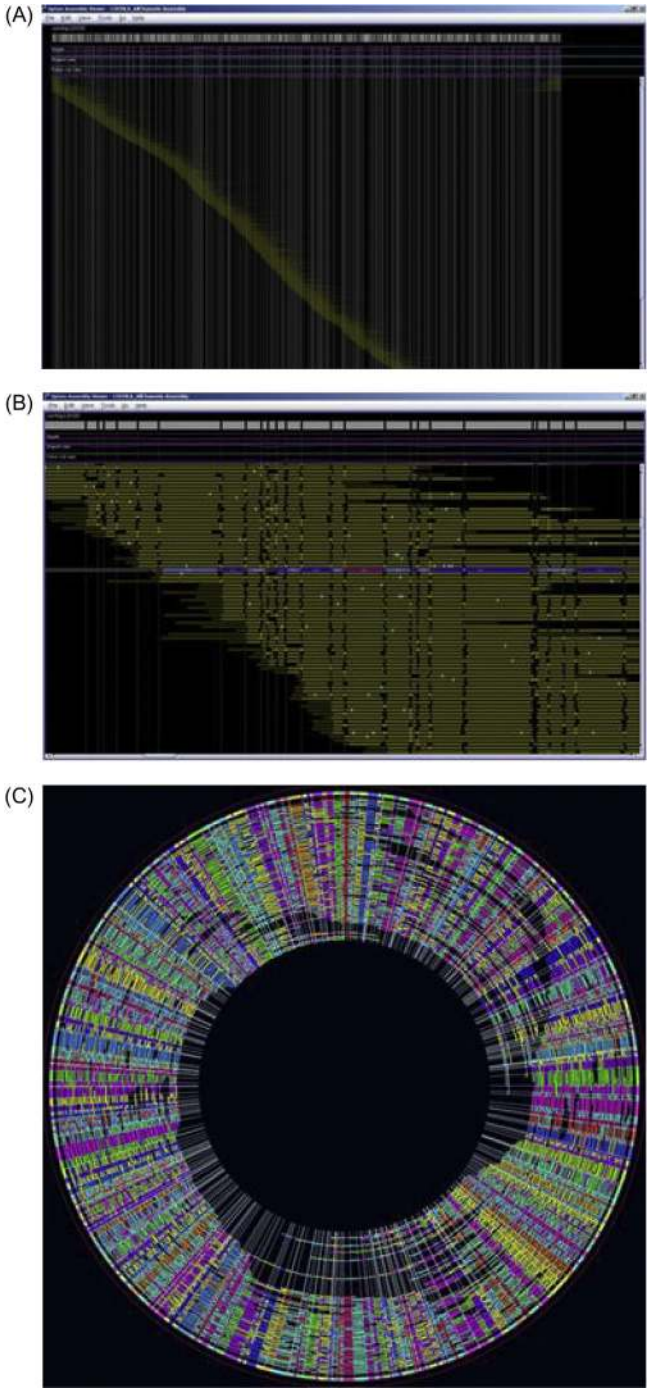


FIGURE 3.14 **Assembly of collected molecules.** (A) In a typical matching of an alignment of 1500 to 50,000 molecules, overlapping restriction fragments grow the chromosome until ends cease growth, or for circular chromosomes, until overlap to previous fragment sets occurs. (B) An enlargement of the overlapping molecule assembly. (C) A graphic representation of the like-colored fragments assembling, in this case into a circular chromosome. In all cases, a criterion of a minimum 30 molecules representation for each restriction fragment is set. More often, there are hundreds of fragments present for many assemblies, adding to the statistical reliability of fragment-size determinations.

3.8.4 Conclusions

Optical mapping provides information on the genome that cannot be obtained from PFGE profiles and a perspective very different from comparing whole-chromosome sequences. Optical mapping is a powerful tool for studying structural genomics because it provides a bird's eye view of chromosomal morphology and architecture. Consequently, optical mapping can be used to visualize and compare different genomes, such as genomes of related species/strains, as well as genomes of pathogenic and nonpathogenic strains within a bacterial species. Optical mapping can also be used to study the same genome in different states.

Since some of the first publications in 1993, optical mapping has been developed and extended from sizing restriction fragments on bacteriophage lambda and bacterial artificial chromosome (BAC) clones (48,500 to 150,000 bp), to scaffolding larger chromosomes such as those in *Candida albicans* (8 chromosomes, 16 Mbp),[26] *Plasmodium falciparum* (14 chromosomes, 23.3 Mbp),[27] rice (24 chromosomes, 389 Mbp),[28] maize (20 chromosomes, 2300 Mbp),[29] mouse (40 chromosomes, 2500 Mbp),[30] humans (46 chromosomes, 3000 Mbp),[31] and most recently the goat genome (60 chromosomes, 2900 Mbp).[32,33]

With its mid-range resolution and graphic flexibilities, optical mapping is ideal for the examination of whole

chromosomes extending from viruses to humans, for independently validating sequence assemblies, for scaffolding higher-order 10–100-Mbp chromosome sequence contigs, and for rapidly detecting differences between the chromosomes of outbreak strains of bacteria.[34–36]

References

1. Sanger F, et al. *Proc Natl Acad Sci USA* 1977;**74**:5463–7.
2. Ronaghi M, et al. *Anal Biochem* 1996;**242**:84–9.
3. Wheeler DA, et al. *Nature* 2008;**452**:872–6.
4. Loman NJ, et al. *Nat Rev Microbiol* 2012;**10**:599–606.
5. 454 Life Sciences Corporation. *How is genome sequencing done?* Available online at: <http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf>.
6. Mardis ER. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
7. Illumina. *Hist Illumina Seq* 2013.. Available online at: <http://www.illumina.com/technology/solexa_technology.ilmn>.
8. Applied Biosystems. *Applied biosytems SOLiD 4 system*; 2010. Available online at: <http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_078637.pdf>.
9. Schadt EE, et al. *Hum Mol Genet* 2010;**19**(Review Issue 2):R227–40.
10. Pacific Biosciences. *SMRT Technol* 2013. Available online at: <http://www.pacificbiosciences.com/products/smrt-technology/>.
11. Choudhuri S. *J Biochem Mol Toxicol* 2004;**18**:171–9.
12. Choudhuri S. *Toxicol Mech Methods* 2006;**16**:137–59.
13. Mockler TC, et al. *Genomics* 2005;**85**:1–15.
14. Yazaki J, et al. *Curr Opin Plant Biol* 2007;**10**:1–9.
15. Bertone P, et al. *Science* 2004;**306**:2242–6.
16. Giaever G, et al. *Nature* 2002;**418**:387–91.
17. Skarnes WC, et al. *Nature* 2011;**474**:337–42.
18. Gaj T, et al. *Trends Biotechnol* 2013;**31**:397–405.
19. Kamath RS, et al. *Nature* 2003;**421**:231–7.
20. Boutros M, et al. *Science* 2004;**303**:832–5.
21. Darling AE, et al. *PLoS ONE* 2010;**5**:e11147.
22. Latreille P, et al. *BMC Genomics* 2007;**8**:321.
23. Ribot EM, et al. *Foodborne Pathog Dis* 2006;**3**:59–67.
24. Kotewicz ML, et al. *Microbiology* 2007;**153**:1720–33.
25. Kudva IT, et al. *J Bacteriol* 2002;**184**:1873–9.
26. van het Hoog M, et al. *Genome Biol* 2007;**8**:R52.
27. Riley MC, et al. *Malar J* 2011;**10**:252.
28. Zhou S, et al. *BMC Genomics* 2007;**8**:278.
29. Zhou S, et al. *PLoS Genet* 2009;**5**:e1000711.
30. Church DM, et al. *PLoS Biol* 2009;**7**:e1000112.
31. Teague B, et al. *Proc Natl Acad Sci USA* 2010;**107**:10848–53.
32. Dong Y, et al. *Nat Biotechnol* 2012;**31**:135–41.
33. Mak HC. *Nat Biotechnol* 2013;**31**:123.
34. Zhou S, et al. *J Bacteriol* 2004;**186**:7773–82.
35. Chen Q, et al. *Microbiology* 2006;**152**:1041–54.
36. Kotewicz ML, et al. *Microbiology* 2008;**154**:3518–28.

CHAPTER

4

The Beginning of Bioinformatics*

OUTLINE

4.1 MARGARET DAYHOFF, RICHARD ECK, ROBERT LEDLEY, AND THE BEGINNING OF BIOINFOMATICS

Although bioinformatics is one of the buzzwords in the post-genomic era, it is by no means a completely new discipline. The beginning of the pioneering work by Margaret Dayhoff, Richard Eck, and Robert Ledley in computer-aided analysis of protein data goes back to the period around 1960. Dayhoff, Eck, and Ledley capitalized on their experience and training in computing, mathematics, and life sciences in collecting and organizing protein sequences, sequence analysis, and studies of protein evolution.[1,2,3] Their work could be regarded as the direct ancestor of modern bioinformatics. In 1965, Dayhoff, Eck, and a couple of colleagues compiled the first **Atlas of Protein Sequence and Structure**, which had ~50 sequences known at the time. The second volume was published in 1966 and had a little over 100 sequences. This compilation of protein sequence and structure information was the predecessor of the current gene and protein databases that form the backbone of contemporary bioinformatics. In subsequent years, as more and more protein sequences were reported, the Atlas grew in size and popularity under the leadership of Dayhoff. Eventually, this database became The **Protein Information Resource (PIR) database**, now maintained at Georgetown University.

Margaret Dayhoff was a professor at Georgetown University Medical Center. As an independent researcher, Dayhoff brought her background of mathematics, chemistry, and computing to address problems in biology, particularly protein chemistry, and became the pioneer in the application of mathematics and computational methods to biochemistry. One of her most important contributions was developing, together with Richard Eck, the single-letter code for amino acids that is used by all protein analysis tools. She developed a computer algorithm for protein-sequence alignment, which was (correctly) thought to reveal their evolutionary history.

Richard Eck studied chemical engineering and plant biology. In 1961, Eck published a paper in *Nature* in which he compared all the sequences of hemoglobin variants, and other proteins such as insulin, from different species. He realized that the information on amino-acid sequences could be organized in different ways in order to produce specific patterns. He also identified numerous amino-acid substitutions in proteins and noted that the pattern of substitutions was not random. In a conference in 1964, Eck presented a **cryptogrammic** method to trace the evolution of proteins. He suggested that, using this result, one could

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

calculate the degree of relatedness of each protein with reference to its ancestors, and draw a family tree in which the distances between the branches represented a quantitative measure of relatedness. Thus, Eck outlined the basis of reconstruction of a phylogenetic tree.

Robert Ledley, who studied theoretical physics and dentistry, envisioned an important application of computers to sequence analysis. He suggested that after the polypeptide chain is cut into many overlapping fragments, whose sequences could be determined by peptide sequencing, the fragment reassembly of partial sequences to obtain full sequences could be done using computers. Thus, Ledley suggested that computers could assist biochemists in their efforts to determine protein sequences. He invited Dayhoff to join the staff of National Bureau of Standards (NBRF; later the National Institute of Standards and Technology, or NIST) in 1960 to continue investigating this question. Dayhoff and Ledley wrote FORTRAN programs that could direct the assembly of partial peptide sequences in the right order in less than 5 minutes.

Both Dayhoff and Eck became involved in evolutionary studies of proteins while Ledley continued with his interest in the application of computers in biology. Dayhoff started playing an increasingly important role in protein-sequence analysis and continued to contribute to evolutionary biology based on her studies on protein sequences. She published the first reconstruction of a phylogenetic tree using a maximum parsimony method, discussed in Chapter 9. She also developed the first amino-acid substitution matrix for studying protein evolution, called the **PAM matrix**. PAM stands for **point accepted mutation** (also referred to as **percent accepted mutation**) because it represents accepted point mutation per 100 amino acid residues. A publication by Dayhoff in the popular science journal *The Scientific American*, entitled *Computer Analysis of Protein Evolution*,[4] can be regarded as one of the most important initial publications in bioinformatics and molecular phylogenetics. For her enormous pioneering contributions, Margaret Dayhoff is popularly regarded as the founder of modern bioinformatics.

4.2 DEFINITION OF BIOINFORMATICS

The term "bioinformatics" was coined by Paulien Hogeweg and Ben Hesper in 1978.[5] In a recent review article recapitulating the history of bioinformatics, Hogeweg stated that the term had been used by Hogeweg and Hesper since the beginning of the 1970s, but was formally coined in 1978 in an article written in Dutch. In the beginning, the term was used to mean the study of informatic processes in biotic systems. Bioinformatics is basically informatics as applied to biology—that is, computer-aided analysis of biological data. There are many definitions/descriptions of bioinformatics; some of these definitions make no distinction between bioinformatics and computational biology as a whole. Luscombe et al.[6] defined bioinformatics as follows:

> Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale.

Higgs and Attwood[7] provided two definitions of bioinformatics that are same in spirit but stated in two different ways:

> (1) Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes; and (2) bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high-throughput experiments.

Therefore, for molecular biologists, bioinformatics is the discipline of computer-aided analysis of information relating to genes, genomes, and their products. In other words, for all practical purposes, bioinformatics can be regarded as **computational molecular biology**, that uses computational techniques to study the structure, function, regulation, and interactive network of genes and proteins. The ultimate goal is to analyze and predict the structure, organization, function, regulation, and dynamics of the entire genome of an organism.

4.3 BIOINFORMATICS VERSUS COMPUTATIONAL BIOLOGY

Computational biology is an umbrella term that includes any subdiscipline in biology that uses computer-aided analysis, modeling, and prediction. Some examples include the modeling of predator–prey relationships in an ecosystem, the modeling and prediction of population and community dynamics in an ecosystem, quantitative structure–activity analysis and prediction of the biological effects of chemicals, prediction of metabolic fate of chemicals in vivo, and pharmacokinetic modeling of drugs and xenobiotics, etc. In contrast, bioinformatics can be regarded as computational molecular biology, as indicated above. Therefore, according to the definitions discussed in this book, computational biology is much broader in scope and bioinformatics is a part of it. Bioinformatics, like

other areas of computational biology, is essentially a multidisciplinary science because it uses techniques and concepts from a number of disciplines, such as molecular biology and biochemistry, computer science, statistics and mathematics, and informatics (information science).

4.4 GOALS OF BIOINFORMATIC ANALYSIS

The ultimate goal of bioinformatics is to be able to predict the biological processes in health and disease. In order to acquire such an ability, a thorough understanding of the biological processes is necessary. Therefore, the proximate goal of bioinformatics is to develop such an understanding through analysis and integration of the information obtained on genes and proteins, as well as to develop new tools and continuously improve the existing set of tools for diverse types of analyses. Bioinformatics also aims to develop tools that help in the management of and access to data and information, including improved search and retrieval capability of genomic data and information from various types of databases. Some examples of common bioinformatic tools and analyses that are continuously being improved and refined are: data capture and storage capability; the usability of databases; data analysis; nucleic acid and protein sequence analysis and sequence annotation; structural analysis of proteins and prediction of protein structure, including three-dimensional (3D) structure; protein domain prediction; gene prediction; analysis of functional studies; analysis of gene and protein networks; and phylogenetic analysis.

The analytical tools in bioinformatics are computer algorithms and statistics. Improvements in the capacity of existing tools and the development of new tools are both driven by the need for newer dimensions and greater speed of analysis, as well as the ability to handle an ever-increasing amount of data. However, the success and prediction accuracy of bioinformatic analysis ultimately depends on our knowledge of the biology of organisms. Therefore, as more data accumulate in the databases and more scientific information becomes available, the progress of science and its prognostic ability will require and hence dictate the development of new bioinformatic tools. Acquisition of more data and information, storage of all that information, expansion of databases, new strategies needed for analysis, and advances in computing power are all expected to facilitate the analysis of large volumes of data and discovery of new biological principles and insights from which unifying principles of life and its evolution can be discerned.

4.5 BIOINFORMATICS TECHNICAL TOOLBOX

Bioinformatic analysis requires data (such as sequence information), databases, and analysis tools. Databases are built from data obtained through wet laboratory experiments. Some of the original nucleotide- and protein-sequence databases were created more than 30 years ago. Subsequently, information from these original databases was utilized to create curated and more refined databases to meet specific research needs. With the advances in genomics, proteomics, and metabolomics, particularly with the development of disciplines like pharmacogenomics and toxicogenomics, the need for storage of and access to the newly created datasets has led to the development of further specialized databases. Through the collaboration of academic, corporate, and regulatory scientists, standards have been developed as to how to submit a specific type of data to the relevant databases. A more detailed discussion of various databases will be undertaken in Chapter 5.

The bioinformatics technical toolbox provides analysis tools (algorithms) and visualization techniques of the data generated through high-throughput experiments, such as high-throughput sequencing, microarray analysis, mass spectrometry, and other proteomic techniques. The analysis tools are computer based (software), and the development of newer tools is driven by various needs, such as an increased need for handling the huge body of data, faster analysis, expanded scope of the analysis, multiple simultaneous analyses, to name a few. A few examples of software-driven analysis that have tremendously facilitated bioinformatics research are:

- Analysis of nucleotide sequences
- Detection of single nucleotide polymorphisms (SNPs) and copy number variation (CNV)
- Understanding the sequence features and differences between coding and noncoding regions
- Alignment of nucleotide sequences
- Prediction of open reading frames (ORFs), restriction-enzyme cutting sites in DNA, various *cis*-acting regulatory DNA elements in the gene, and putative miRNA-encoding sequences in the genome
- Gene-expression analysis
- Designing probes and primers
- Analysis of protein sequences
- Alignment of amino-acid sequences
- Prediction of protein structure (including 3D structure), protein−protein interactions, post-translational modifications of proteins, hydrophilicity/hydrophobicity and potential

antigenicity of proteins, and various protein domains, such as transmembrane domains

Prediction of phylogenetic relationships among proteins.

In addition, gene-expression analysis information has led to the development of systems biology tools that can perform simulation, steady-state analysis, network identification, complex behavior analysis of the system, and various other tasks.

References

1. Strasser BJ. *J Hist Biol* 2010;**43**:623–60.
2. Lee J. *Prot Sci* 2007;**16**:1509–10.
3. Doolittle RF. *PLoS Comput Biol* 2010;**6**:e1000875.
4. Dayhoff MO. *Sci Am* 1969;**221**:86–95.
5. Hogeweg P. *PloS Comput Biol* 2011;**7**:e1002021.
6. Luscombe NM, et al. *Methods Inf Med* 2001;**40**:346–58.
7. Higgs PG, Attwood TK. *Bioinformatics and molecular evolution*. MA: Blackwell; 2005.

CHAPTER

5

Data, Databases, Data Format, Database Search, Data Retrieval Systems, and Genome Browsers*

OUTLINE

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

5.1 GENOMIC DATA

A publication by Mark Gerstein and colleagues dating as far back as 2001 was entitled, *Interrelating Different Types of Genomic Data, from Proteome to Secretome: 'Oming in on Function*.[1] This title captures the scope of different types of genomic data. In genomic parlance, the suffix "ome" means the entire collection of an entity. For example, a transcriptome is the entire collection of all RNA transcripts in a cell/tissue at a given time point. Although transcriptome includes all RNA molecules, such as mRNA, rRNA, tRNA, and other noncoding RNAs, it is mostly used in the context of mRNAs. Similarly, the proteome is the entire collection of all proteins, miRNome means the entire collection of all microRNAs (miRNAs) in a cell/tissue at a given time point, and interactome means the collection of all possible molecular interactions (or a subset of molecular interactions) in a cell. *Mapping interactomes represents a major effort in the study of the cellular regulatory networks.*

The bulk of the raw genomic data that were accumulating even before the beginning of human genome sequencing are the DNA-sequence data (gene and mRNA sequence, the latter in the form of the sense strand[a] of complementary DNA (cDNA)). The collection of sequence data exploded as a result of the sequencing of the human genome and the genomes of other species. With DNA sequencing becoming increasingly refined and cheaper, there has been a corresponding increase in the quantity and quality of DNA-sequence data. Keeping pace with the DNA-sequence data has grown the gene- and protein-expression data. Again, this has been facilitated by the availability of techniques to study gene and protein expression; foremost among these techniques is the microarray, which has revolutionized the study of global gene expression. Such study of global gene expression profiling—that is, the study of transcriptomes—is called **transcriptomics**.

In addition to the sequence and expression data, there are other kinds of data that are genomic data in a broader sense, such as genome-wide monoallelic expression data, proteome data, metabolome data, protein−protein interaction data, protein structural data, protein−DNA interaction data, gene and protein network data, and small noncoding RNA (ncRNA) data. The latest addition to this list is probably genome-wide epigenetic modification data.

Collectively, all these data are expected to help us understand the structure, function, and interaction of cells with one another as well as with the environment. Interaction data should also shed light on the modular organization of the cell.

5.2 SEQUENCE DATA FORMATS

At the core of all genomic data are the sequence data. A sequence data format is a specific layout or arrangement of text characters, symbols, keywords, and description that identify a sequence and contain information about its various attributes. Sequence data file formats are American Standard Code for Information Interchange (ASCII) text files. A typical ASCII file includes text, numbers, and simple signs (such as @, #, $, parenthesis signs, etc.) that a computer can read and are printable; it has no special formatting, such as bold, italics, or underscoring. However, most modern ASCII-based formats support many additional characters.

Currently, many sequence formats exist; some are more common than others. Most databases that store sequence data, and various analysis packages that need sequence input for analysis, have developed their own formats for storing the data, as well as specific data-input formats for analysis.

A widely used input sequence format for the purpose of analysis is the FASTA format. A different input sequence format is required by the PHYLIP for phylogenetic analysis; these are discussed below.

5.2.1 FASTA Format

FASTA (pronounced fast "A") stands for "fast all". Many sequence-analysis programs, such as many sequence-alignment programs, need the data to be entered in FASTA format. The minimum amount of input information required in a typical FASTA format is as follows: the first line is the definition (or description) line that starts with the ">" sign, which is a crucial element in FASTA format. Analysis programs that need the sequence data input in FASTA format will fail to read the sequence if the ">" sign is not included. The ">" sign is followed by a definition (identifier) of the sequence. There should be no space between the ">" sign and the first letter of the definition line. FASTA format can allow more information on the definition line, as shown in the example below. The lines of the text should preferably contain less

[a]Out of the two strands in a gene or cDNA, the sequence and polarity (5′→3′) of one strand is the same as that of mRNA (except for the fact that DNA has "T" and mRNA has "U"). This strand is called the sense strand/coding strand/plus (+) strand. In a gene, the sense strand is NOT transcribed. The transcribed strand is called the template strand/antisense strand/noncoding strand/minus (−) strand. The term "sense" means that the sequence of codons can be obtained from it; hence, the sequence of encoded amino acids can be predicted from it. In the database, the sequence of the DNA sense strand is submitted.

than 80 characters. A sequence in FASTA format can be written with or without gaps.

The following are examples of FASTA sequence format (actual sequence truncated[b]).

Example 1:

```
>Mouse Oatp-5 protein
MGEPGKRVGI HRVRCFAKIK VFLLALIWAY ISKILSGVYM
... ... ...
```

Example 2:

```
>Mouse Oatp-5 mRNA
atccattcac tgactaacac aaggacaagt ttggagtgat
... ... ...
```

Example 3:

```
>gi|12619376|gb|AF213260.1| Mus musculus
kidney-specific organic anion transporting
polypeptide 5 mRNA, complete cds
atccattcac tgactaacac aaggacaagt ttggagtgat
... ... ...
```

Example 3 has both the GI (GeneInfo identifier) and the GenBank accession number in the FASTA format.

Note that although the sequence states mRNA it does not have any "U" but has "T" instead. This is because it is the sequence of the sense strand of cDNA. This is how sequences are submitted to the nucleotide databases.

5.2.2 PHYLIP Format

PHYLIP stands for "phylogeny inference package." It was developed by Dr Joe Felsenstein of The University of Washington, Seattle, in the mid-1980s. PHYLIP is a phylogenetic analysis package that can carry out many different analyses, such as parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees[c]. Data types that can be handled include DNA and protein sequences, gene frequencies, restriction sites, distance matrices. The simplest version of the PHYLIP input file format for methods like parsimony, compatibility, and maximum likelihood programs is shown below. The first line of the input file shows the number of species (in this example, four) and the number of characters (in this example, 16 nucleotides) in text format, separated by a space only. The information for each species starts with a 10-character species name. If the species name is not 10 characters long, then a space is introduced to make it 10-character equivalent. In the example, H. sapiens has a space before "sapiens," but other species names do not have any such space. DNA and protein sequence may start immediately after the species name and the sequence can be separated by a space, such as a space every 10 nucleotides.

```
 4 16
M.musculusggtcgtgcgc aggccc
R.norvegicatcacgctcc tagaac
H. sapiensaccacgccct ccacgt
P.troglodyacgcctcccc caagtc
```

5.3 CONVERSION OF SEQUENCE FORMATS USING READSEQ

In order to change a given sequence format to any one of the common sequence formats used in sequence analysis or phylogenetic analysis, the **Readseq** program can be used. It is a free web-based sequence file format conversion tool that reads the input sequence data and converts the input format to the format chosen by the user in a drop-down menu. A total of 19 different file formats are supported by Readseq. Some examples of common formats supported by Readseq are GENBANK, NBRF, EMBL, GCG, DNA Strider, FASTA, PHYLIP, PIR, MSF, and CLUSTAL. Readseq was developed by Dr Don Gilbert at Indiana University and is available at http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi. Various sites on the web maintain mirror sites of Readseq, such as those of the US National Center for Biotechnology Information (NCBI; http://www-bimas.cit.nih.gov/molbio/readseq/) and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/cgi-bin/readseq.cgi).

5.4 PRIMARY SEQUENCE DATABASES—GENBANK, EMBL-BANK, AND DDBJ

Primary sequence databases are archival in nature. They contain raw sequence data (experimental results) with some interpretation and explanation, but the data are not curated. There are also redundancies in the primary databases—that is, the same sequence might be submitted by different laboratories, sometimes under different names. A great majority of protein sequences in the primary databases are derived from computational translation of the open reading frame (ORF); hence they have not been experimentally verified for the most part. There are three primary databases that contain all the sequence data so far generated. These are **GenBank**, **EMBL** database, also called the **EMBL-Bank**, and **DDBJ** (DNA Databank of Japan).

[b]The details of the mouse Oatp-5 sequence along with the reference are shown later under sequence flatfile format.

[c]These are discussed in Chapter 9 in more detail.

GenBank, EMBL-Bank, and DDBJ are interconnected; so, data submitted to any one of these databases are shared by, and hence can be retrieved from, all three.

5.4.1 History

GenBank was created in 1979 at the Los Alamos National Laboratory and was called the Los Alamos Sequence Database. It was renamed GenBank in 1982 and became a public database. During 1989 to 1992, GenBank transitioned to the newly created NCBI, a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD. GenBank is built and distributed by the NCBI. NCBI began accepting direct submissions to GenBank in 1993. Since its creation, GenBank has grown at an exponential rate, doubling in size every 18 months.[2,3] The NCBI home page is http://www.ncbi.nlm.nih.gov/.

The EMBL was founded in July 1974 on the basis of an intergovernmental treaty of nine European countries plus Israel. It has grown in membership since then; Luxembourg became the twentieth member in 2007, and Australia joined as an associate member in 2008. The EMBL is located in Heidelberg, Germany. An outstation of EMBL is the European Bioinformatics Institute (EBI), located at Hinxton, near Cambridge, UK. The EMBL database as a central depository of nucleotide sequence was created in 1981 and was known as the EMBL Data Library. The EMBL Data Library moved to the EBI in 1993, and became the precursor to the current EMBL-Bank, which is also maintained at the EBI. The expression "EMBL-Bank" is not frequently used. In the literature, the EMBL-Bank is mostly referred to as EMBL nucleotide sequence database or EMBL database. In this book, the expression EMBL-Bank will be frequently used. The EMBL-Bank is now part of the European Nucleotide Archive (ENA), which consists of three main databases: the **Sequence Read Archive (SRA)**, the **Trace Archive** (these are discussed later), and the EMBL-Bank. The ENA is developed and maintained at the EMBL-EBI under the guidance of the International Nucleotide Sequence Database Consortium (INSDC; discussed below).[4–7] The EMBL-EBI home page is http://www.ebi.ac.uk/. Various databases and tools maintained by EMBL-EBI and made freely available for use can be accessed using EMBL Services at http://www.ebi.ac.uk/services.

DDBJ has been in operation since 1986 and it is maintained at the National Institutes of Genetics at Mishima, Japan. DDBJ is the sole nucleotide-sequence data bank in Asia. The DDBJ home page is http://www.ddbj.nig.ac.jp/. A few recent publications discuss many improvements and added features of DDBJ.[8–11]

The **INSDC** (http://www.insdc.org/), a collaborative consortium, was initiated between GenBank, EMBL (ENA), and DDBJ to connect these three databases. This collaboration created the International Nucleotide Sequence Database (INSD). For over 30 years, the INSDC has maintained the primary nucleotide-sequence database.[12] The INSDC advisory board is composed of members of each of the databases' advisory bodies. *The INSDC has a policy of providing free and unrestricted access to all the available data to scientists worldwide.*[13]

5.4.2 Sequence Submission to the Databases

During the early years of these databases, sequence data were obtained from the published literature and entered manually into the database. GenBank began accepting direct submissions in 1993. Sequence information can be submitted to the databases irrespective of publication of the information in a journal. However, any author reporting the cloning of a gene or an mRNA (as cDNA) in a publication needs to submit the sequence first to any one of the three primary databases, get an **accession number**, and provide that accession number with the publication.

5.4.2.1 Submission to NCBI/GenBank

Sequences can be submitted to the GenBank database using its web-based sequence submission tool called **BankIt**, which is available at http://www.ncbi.nlm.nih.gov/BankIt/oldbankit.html. Until several years ago, a gene sequence had to be submitted using BankIt one exon at a time, where each exon submission was given a unique accession number. Now, however, a set of sequences can be submitted at the same time. Therefore, one entire sequence containing exons and introns can be submitted by entering a proper identifier of each sequence segment during submission. This is all explained in BankIt submission help. Complex submissions containing long sequences, multiple annotations, gapped sequences, or phylogenetic and population studies should be submitted using the **Sequin** submission tool (http://www.ncbi.nlm.nih.gov/Sequin/). A single Sequin file should contain less than 10,000 sequences for maximum performance. Larger submissions should be made with **tbl2asn** (http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/). In contrast to BankIt, which is web based, both Sequin and tbl2asn are NCBI's stand-alone submission tools, and are available for download from the file transfer protocol (FTP) site for use on Mac, PC, and UNIX platforms. Therefore, the submitter can download Sequin or tbl2asn, work off-line to prepare the submission in the required format, and finally submit.

At the NCBI, in addition to GenBank, various other types of sequence data can be submitted to various other databases, such as the **Sequence Read Archive** (**SRA**; stores raw sequencing data from various

next-gen sequencing platforms), the **Trace Archive** (stores sequencing data from gel/capillary platforms such as Applied Biosystems ABI 3730), **dbSNP** (stores mutation data, such as single nucleotide polymorphisms, insertion/deletions, non-polymorphic variants etc.), **dbVar** (stores data on genomic structural variations), and **GEO** (stores MIAME-compliant gene-expression data; MIAME is discussed in a footnote later in the chapter). There are links to these databases from the NCBI website, at http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/. A 2013 publication provides updates on the database resources at the NCBI[14] and another article on GenBank discusses the improvements and many added features of GenBank.[3]

5.4.2.2 Submission to ENA/EMBL-Bank

Sequences can be submitted to EMBL-Bank using its web-based sequence submission tool called **Webin**. Webin allows submission of single and multiple sequences as well as very large numbers of sequences (bulk submissions). Webin link and directions are available at http://www.ebi.ac.uk/ena/about/embl_bank_submissions. In the past, the sequence length of a database record was limited to 350,000 bp. This restriction was lifted in June 2004; as of 2013, entries of any length are permitted in the database. An entire chromosome can now be represented in a single entry. Some genomes that were split in the past in order to comply with the 350,000-bp limit have now been updated into single entries.[15] As mentioned before, EMBL-Bank maintains the **Sequence Read Archive** (**SRA**) and **Trace Archive**.

5.4.2.3 Submission to DDBJ

The web page for sequence submission in DDBJ has recently undergone a complete makeover (http://www.ddbj.nig.ac.jp/faq/datasub-e.html). DDBJ recommends using the new web-based submission tool called the **Nucleotide Sequence Submission System** (**NSSS**; http://www.ddbj.nig.ac.jp/sub/websub-e.html). The NSSS has replaced **Sakura**, beginning November, 2012. Sakura was used for sequence submission for about 17 years (from 1995). However, if the sequences are very long or a large number of sequences are to be submitted at the same time, DDBJ recommends using its **Mass Submission System** (**MSS**), which is available at http://www.ddbj.nig.ac.jp/sub/mss_flow-e.html. Like the NCBI and EMBL-Bank, DDBJ also maintains a **Sequence Read Archive** (**SRA**) and **DDBJ Trace Archive** (**DTA**), which is a permanent repository of DNA sequence chromatograms (traces), base calls, and quality estimates for single-pass reads from various large-scale sequencing projects. Two publications discuss recent progress of the DDBJ.[9,11]

The SRA was established as a public repository for next-generation sequence data and is operated by the INSDC; partners include the NCBI, EMBL-EBI, and DDBJ. The SRA is accessible at http://www.ncbi.nlm.nih.gov/Traces/sra from the NCBI, at http://www.ebi.ac.uk/ena from the EBI, and at http://trace.ddbj.nig.ac.jp from DDBJ.[10,16]

5.4.3 Availability of the Submitted Sequence to the Public

During submission of a sequence, the submitter may choose to release the sequence information to the public at a later date (many months later than the actual date of submission to the database) by giving instruction during submission. This usually happens if there are multiple laboratories working on the same gene/protein, and the work of the scientist submitting the sequence is still not completed for publication (at the time the sequence information is submitted). If such a later release date is not chosen, the sequence is released as soon as the database staff is done with verifying the submission and related information.

5.4.4 Sequence Flatfile Format

During sequence submission, the submitter has to provide some relevant information about the sequence, such as the name of the mRNA/gene, the source, annotation, open reading frame, and putative translation product. All this information is displayed, along with the sequence, in a flatfile. The GenBank and DDBJ formats of a sequence flatfile are almost identical except for two fields: (1) GenBank entries contain GI numbers; each GI number is unique to a GenBank entry only; (2) DDBJ entries contain information about the total number of "A," "C," "G," and "T" in the sequence; GenBank entries do not have this. Like DDBJ, the EMBL-Bank entries also contain information about the total number of "A," "C," "G," and "T" in the sequence. The GI number (also written as "gi") stands for **GeneInfo Identifier** and was an early system used to access GenBank and related databases. The GI numbers are assigned consecutively to each sequence record processed by NCBI; a GI number of a sequence has no resemblance to the accession number of that sequence.[17] The EMBL-Bank format looks a little different, although the same information is contained in all. Each database maintains a detailed discussion about its flatfile format. The websites where the respective flatfile formats are discussed are as follows:

GenBank: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
DDBJ: http://www.ddbj.nig.ac.jp/sub/ref10-e.html
EMBl-Bank: ftp://ftp.ebi.ac.uk/pub/databases/embl/release/usrman.txt (EMBL-Bank User Manual).

Specific sequence information from GenBank can be retrieved from the **nucleotide** database if the accession number or GI number is known. If the accession number or GI number is not known, sequence information can still be retrieved from the nucleotide database using a combination of keywords, such as species name, sequence name, author's name (if known), etc. In this situation, many sequence information records may be retrieved, depending on the search terms used, and the search may have to be further narrowed to get the desired sequence. Gene and mRNA sequence records can also be obtained from the **Gene** database/portal.

Specific sequence information from the EMBL-Bank can be retrieved using **dbfetch**, as well as the **EMBL-SVA** (ENA Sequence Version Archive) if the accession number is known. If the accession number is not known, the **EB-eye** (**EBI**) search can be performed using keywords, such as a combination of species name, sequence name, etc. (figures indicated later).

Specific sequence information from DDBJ can be retrieved using the **getentry** retrieval system if the accession number is known. If the accession number is not known, sequence information can be retrieved using **ARSA** (All-round Retrieval of Sequence and Annotation), using a combination of keywords, as before. *Examples cited in the text will be mostly from NCBI/GenBank.*

5.4.4.1 GenBank Sequence Flatfile Format

Mus musculus kidney-specific organic anion transporting polypeptide 5 mRNA, complete cds

GenBank: AF213260.1

FASTA Graphics

```
LOCUS       AF213260               2798 bp    mRNA    linear   ROD 31-JAN-2001*
DEFINITION  Mus musculus kidney-specific organic anion transporting polypeptide
            5 mRNA, complete cds.
ACCESSION   AF213260
VERSION     AF213260.1  GI:12619376
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 2798)
  AUTHORS   Choudhuri,S., Ogura,K. and Klaassen,C.D.
  TITLE     Cloning, expression, and ontogeny of mouse organic
            anion-transporting polypeptide-5, a kidney-specific organic anion
            transporter
```

```
  JOURNAL   Biochem. Biophys. Res. Commun. 280 (1), 92-98 (2001)
   PUBMED   11162483
REFERENCE   2  (bases 1 to 2798)
  AUTHORS   Choudhuri,S., Ogura,K. and Klaassen,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-DEC-1999)Pharmacology, University of Kansas Medical
            Center, 3901 Rainbow Blvd., Kansas City, KS 66160, USA
FEATURES             Location/Qualifiers
     source          1..2798
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="BALB/c"
                     /db_xref="taxon:10090"
                     /tissue_type="kidney"
     CDS             179..2191
                     /note="Oatp5; transport protein"
                     /codon_start=1
                     /product="kidney-specific organic anion transporting
                     polypeptide 5"
                     /protein_id="AAG60350.1"
                     /db_xref="GI:12619377"
                     /translation="MGEPGKRVGIHRVRCFAKIKVFLLALIWAYISKILSGVYMSTML
                     TQLERQFNISTSIVGLINGSFEMGNLLVIVFVSYFGTKLHRPIMIGVGCAVMGLGCFI
                     ISLPHFLMGRYEYETTISPTSNLSSNSFLCVENRSQTLKPTQDPAECVKEIKSLMWIY
                     VLVGNIIRGIGETPIMPLGISYIEDFAKSENSPLYIGILEVGKMIGPILGYLMGPFCA
                     NIYVDTGSVNTDDLTITPTDTRWVGAWWIGFLVCAGVNVLTSIPFFFFPKTLPKEGLQ
                     DNGDGTENAKEEKHRDKAKEENQGIIKEFFLMMKNLFCNPIYMLCVLTSVLQVNGVAN
                     IVIYKPKYLEHHFGISTAKAVFLIGLYTTPSVSAGYLISGFIMKKLKITLKKAAIIAL
                     CLFMSECLLSLCNFMLTCDTTPIAGLTTSYEGIQQSFDMENKFLSDCNTRCNCLTKTW
                     DPVCGNNGLAYMSPCLAGCEKSVGTGANMVFQNCSCIRSSGNSSAVLGLCKKGPDCAN
```

```
                    KLQYFLIITVFCCFFYSLATIPGYMVFLRCMKSEEKSLGIGLQAFFMRLFAGIPAPIY
                    FGALIDRTCLHWGTLKCGEPGACRTYEVSSFRRLYLGLPAALRGSIILPSFFILRLIR
                    KLQIPGDTDSSEIELAETKPTEKESECTDMHKSSKVENDGELKTKL"
ORIGIN
        1 atccattcac tgactaacac aaggacaagt ttggagtgat ctgaactctg ggaagcctgt
       61 ggccagggaa gcctgcactg aggacagctg cttcctcagc tgctgtgtag actgagttcc
      121 atcaggcagt ggtaggactt tgaaagcaga gacatcctta aacaatcaga agaacaaaat
      181 gggagaacct gggaaaaggg ttggaatcca cagggtcagg tgctttgcca agatcaaggt
      241 gtttctgttg gcattaatat gggcatatat atccaaaata ctatcaggag tttacatgag
      301 tactatgctc acacaattag agagacaatt caatatttcc acatctatag ttggacttat
      361 caatgggagc tttgagatgg gtaacctttt ggtgattgta ttcgtgagtt attttggaac
      421 aaaactgcat agacctatca tgattggtgt tggttgtgca gttatgggcc tagggtgttt
      481 cataatatca ctacctcatt tcctcatggg cagatacgaa tatgaaacaa caatttcacc
      541 tacaagcaac ttgtcctcaa acagcttttt gtgtgtggaa aacagatccc agaccttaaa
      601 gccaacacaa gacccagcag agtgtgtgaa agaaattaaa tcattaatgt ggatatatgt
      661 actggtagga aacattatac gtggaattgg tgaaactccc atcatgcctt taggtatttc
      721 ctatatagaa gactttgcca aatcagaaaa ttctccttta tacattggaa ttttagaagt
      781 tgggaagatg attggcccaa tacttggata tttgatggga cctttctgtg caaacattta
      841 tgtagacaca gggtctgtga atacagatga cctgaccata actcccactg atacacgctg
      901 ggtcggtgct tggtggattg gctttttggt ctgtgcagga gtgaatgtcc tgaccagcat
      961 ccccttttttc ttctttccaa aaacactccc aaaggaagga ttacaggata atggggatgg
     1021 aactgaaaat gccaaagagg agaagcacag agacaaggcc aaggaggaaa accaaggaat
     1081 cattaaagaa ttcttcctta tgatgaagaa cctcttctgt aaccctattt acatgctttg
     1141 cgtccttaca agtgtgctcc aggtaaatgg agttgccaat attgtgattt acaagcctaa
     1201 atacctggaa catcattttg gaatctccac agcaaaggca gtcttcctca ttggtcttta
     1261 taccacacct tcagtatctg ctggatattt aattagtggt tttattatga agaagttgaa
     1321 gattactctc aagaaagctg caatcatagc actttgccta ttcatgtctg agtgcctttt
     1381 atccctttgt aactttatgc taacctgtga taccactcca attgccggct taactacctc
     1441 ttatgaagga attcagcagt cttttgatat ggagaataag tttctttctg actgcaacac
     1501 aaggtgtaac tgcttaacaa aaacatggga tccagtgtgt gggaacaatg gcctagcata
```

```
1561 catgtcacct tgccttgcag gctgtgaaaa gtctgttgga acaggagcca acatggtgtt

1621 tcaaaattgc agctgcattc ggtcatcagg aaactcatct gcagtcctgg ggctgtgtaa

1681 gaaaggccct gactgtgcta acaagcttca gtacttttta atcataacgg tattttgctg

1741 cttcttctac tcgttagcaa ccatacctgg gtacatggtt tttctgagat gtatgaagtc

1801 tgaagagaag tcacttggaa ttggattaca ggcatttttc atgagactat ttgctggtat

1861 tcctgcacct atttactttg gcgctttgat agacagaaca tgcttacatt ggggaactct

1921 gaaatgtggt gagccaggag catgcaggac ctatgaagtc agtagtttca ggcgcctcta

1981 tcttggattg cctgcagctc taagaggatc aatcattctt ccttcattct tcattctaag

2041 acttatcagg aaactccaaa tccctgggga cactgactct tcagaaattg aacttgcaga

2101 gacgaagccc acagagaagg aaagtgagtg cacagacatg cacaaaagtt ctaaggtcga

2161 gaacgatgga gaactgaaaa ctaagctgta atgaggtttc tactggccta tgcaaggcca

2221 cgaacagaat actcatttca tttcctttga atcataagag aaataatagg aaccctcatc

2281 tttaaggacc tcaaaagcta tttttctcat tataaaaata attactgata ttattttcag

2341 aacttcaggg tagcacttaa gattttccta gtgaagactt taatggtgac ccccaccctg

2401 gactttaaaa agccttcgtt ttcaaagagc attttctctt taaactcagt caaaggaaat

2461 gtgtgtttct tgcatatctt caagtagatt tcatttcact taatttcatt gaatttacat

2521 ttcaatattg gaggtaatta gagctgaaag tatgccttct ggttgtgtca tattgaaata

2581 aattgttcag attcatcctt tccatgtgca aggtgtctgc atgtgtcttt aactctttgg

2641 gagctgttat ctttcttttc tcattctaga cttttgatgc ttcagagatt agactctcac

2701 taatgtgtca tctcgtgttt tcaattccct ctttcattat tcatgtcaca tatttgatca

2761 ttttgtttag aactctgaca aatttaaaca ggttatta
```

*This is the GenBank flatfile of an original submission. The publication is indicated in the REFERENCE field of the submission.

In this example, the following information is provided by the data flatfile:

1. The first line, called the LOCUS line or LOCUS field, contains the locus name, the length of the sequence, and a three-letter word indicating the GenBank database division this sequence belongs to. In this example, ROD in the right-hand top corner indicates that the sequence is a rodent sequence[d]. The sequence was originally submitted to the database on 8 December, 1999 (highlighted). The date in the LOCUS field is the date of last modification. In this example, the sequence was last modified on 31 January, 2001. This modification date may be same as the release date, but there is no way to know that just by looking at the record.

[d]The GenBank sequence database has 18 divisions. ROD stands for the division that contains rodent sequences. This topic is discussed later in this chapter.

2. The sequence is mouse (*Mus musculus*) kidney-specific organic anion transporting polypeptide-5 (*Oatp5*) mRNA sequence. Oatp-5 is also known as Slc21a13 and Slco1a6. Although it is an mRNA sequence, note that there are no "U" residues; instead there are "T" residues in the sequence. This is because the sense strand sequence of the cDNA is submitted to the database as a convention. The sense strand has the same polarity (5′→3′) and the same sequence as the mRNA except for "T" in DNA and "U" in RNA.
3. This submission is version 1 of the original submission because the sequence has not been modified since it was first submitted. This is revealed by the version of the accession number (accession number is AF213260; first version is AF213260.1). It should be remembered that the reason why a version is replaced is not indicated in the flatfile. However, the date when a particular version is replaced by a newer version is indicated in the COMMENT field of the flatfile, along with the GI number of the replaced version. The GI number can be clicked to obtain the replaced version. This gives the user the opportunity to compare the different versions and identify the changes. This particular flatfile does not have the COMMENT field because there is no special note associated with this sequence. The original sequence may be modified by the submitter for various reasons. For example, resequencing of clones may reveal some error in the earlier sequencing; hence, the original sequence may need to be corrected. Sometimes, in the case of cDNA cloning using 5′ and 3′ rapid amplification of cDNA ends (RACE), the 5′- or the 3′-end of the clone may be incomplete, even though the ORF is complete. Subsequent mapping of the transcription start site often detects additional sequence that was missing from the 5′-end of the original sequence[e]. Reporting this additional sequence modifies the original submission. In this way, every time the original sequence is modified, the accession number remains the same, but the version number increases from dot 1 (.1) to dot 2 (.2) to dot 3 (.3), and so on. As already mentioned, the GI number (highlighted) is unique to the GenBank sequence flatfile; it is not found in EMBL-Bank or DDBJ sequence flatfiles.
4. The coding sequence (CDS), or the open reading frame (ORF), spans from base 179 to 2191. This means that the "A" of the ATG (translation start codon) is the 179th base and the second "A" of the TAA (translation stop codon) is the 2191st base.
5. The 5′- and 3′-untranslated region (UTR) sequences span bases 1−179 and bases 2192−2798, respectively. The sequence information does not contain any indication about the transcription start site (cap site) and thus the completeness of the 5′-UTR cannot be ascertained (although in this case the 5′-UTR is complete). If the 5′-UTR is known to be incomplete, this can be indicated by a "<1" sign (e.g. <1...100), meaning that the beginning of the 5′-UTR lies upstream of base 1 of the sequence. The completeness of the 3′-UTR can be verified by checking for the canonical poly(A) signal sequence "aataaa" or its variant "attaaa." The poly(A) signal sequence in an mRNA is usually located ~10−30 bases upstream of the polyadenylation site. In this example, the first "A" of the "aataaa" is the 2577th base, but the 3′-UTR is still longer than 2798 bases. This indicates that this mRNA may have alternatively polyadenylated forms; a shorter form that is polyadenylated 12 nt downstream from the first poly(A) signal,[18] and a longer form that is polyadenylated further downstream. The poly(A) signal sequence for this longer form is not present in the sequence, indicating that the present 3′-UTR is not complete. This is further supported by the RefSeq accession number NM_023718 (version NM_023718.3), which shows that the complete mouse Oatp-5 (Slco1a6) sequence is 2804 bases long and contains the second poly(A) signal sequence. Thus, the cited sequence here is shorter than the full-length sequence by only 6 bases. These extra 6 bases show the location of the second poly(A) signal sequence, which is "attaaa." In fact, in the cited example, the sequence is truncated right within the second poly(A) signal sequence.
6. The amino-acid (aa) sequence of the putative translation product (670 aa long) is also part of the submission. It contains the accession number of the protein database (AAG60350.1; highlighted).
7. There is information about the publication and the authors in the REFERENCE field.

[e]For certain applications, such as during the construction of a knockout construct, it is important to know the beginning of the transcription start site (hence the complete 5′-UTR) as well as the ORF, but it is not necessary to know the entire 3′-UTR.

5.4.4.2 *EMBL-Bank Sequence Flatfile Format*

(Same sequence as above.)

```
ID   AF213260; SV 1; linear; mRNA; STD; MUS; 2798 BP.
XX
AC   AF213260;
XX
DT   31-JAN-2001 (Rel. 66, Created)
DT   23-SEP-2008 (Rel. 97, Last updated, Version 2)
XX
DE   Mus musculus kidney-specific organic anion transporting polypeptide 5 mRNA,
DE   complete cds.
XX
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Mus; Mus.
XX
RN   [1]
RP   1-2798
RX   DOI; 10.1006/bbrc.2000.4072
RX   PUBMED; 11162483.
RA   Choudhuri S., Ogura K., Klaassen C.D.;
RT   "Cloning, expression, and ontogeny of mouse organic anion-transporting
RT   polypeptide-5, a kidney-specific organic anion transporter";
RL   Biochem. Biophys. Res. Commun. 280(1):92-98(2001).
XX
RN   [2]
RP   1-2798
```

```
RA Choudhuri S., Ogura K., Klaassen C.D.;
RT   ;
RL   Submitted (08-DEC-1999) to the INSDC.
RL   Pharmacology, University of Kansas Medical Center, 3901 Rainbow Blvd.,
RL   Kansas City, KS 66160, USA
XX
DR   Ensembl-Gn; ENSMUSG00000079262; Mus_musculus.
DR   Ensembl-Tr; ENSMUST00000111827; Mus_musculus.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..2798
FT                   /organism="Mus musculus"
FT                   /strain="BALB/c"
FT                   /mol_type="mRNA"
FT                   /tissue_type="kidney"
FT                   /db_xref="taxon:10090"
FT   CDS             179..2191
FT                   /codon_start=1
FT                   /product="kidney-specific organic anion transporting
FT                   polypeptide 5"
FT                   /note="Oatp5; transport protein"
FT                   /db_xref="GOA:Q99J94"
FT                   /db_xref="InterPro:IPR004156"
FT                   /db_xref="InterPro:IPR011497"
FT                   /db_xref="InterPro:IPR016196"
FT                   /db_xref="InterPro:IPR020846"
FT                   /db_xref="MGI:1351906"
```

```
FT                   /db_xref="UniProtKB/Swiss-Prot:Q99J94"
FT                   /protein_id="AAG60350.1"
FT                   /translation="MGEPGKRVGIHRVRCFAKIKVFLLALIWAYISKILSGVYMSTMLT
FT                   QLERQFNISTSIVGLINGSFEMGNLLVIVFVSYFG
FT                   LPHFLMGRYEYETTISPTSNLSSNSFLCVENRSQTLKPTQDPAECVKEIKSLMWIYVLV
FT                   GNIIRGIGETPIMPLGISYIEDFAKSENSPLYIGILEVGKMIGPILGYLMGPFCANIYV
FT                   DTGSVNTDDLTITPTDTRWVGAWWIGFLVCAGVNVLTSIPFFFFPKTLPKEGLQDNGDG
FT                   TENAKEEKHRDKAKEENQGIIKEFFLMMKNLFCNPIYMLCVLTSVLQVNGVANIVIYKP
FT                   KYLEHHFGISTAKAVFLIGLYTTPSVSAGYLISGFIMKKLKITLKKAAIIALCLFMSEC
FT                   LLSLCNFMLTCDTTPIAGLTTSYEGIQQSFDMENKFLSDCNTRCNCLTKTWDPVCGNNG
FT                   LAYMSPCLAGCEKSVGTGANMVFQNCSCIRSSGNSSAVLGLCKKGPDCANKLQYFLIIT
FT                   VFCCFFYSLATIPGYMVFLRCMKSEEKSLGIGLQAFFMRLFAGIPAPIYFGALIDRTCL
FT                   HWGTLKCGEPGACRTYEVSSFRRLYLGLPAALRGSIILPSFFILRLIRKLQIPGDTDSS
FT                   EIELAETKPTEKESECTDMHKSSKVENDGELKTKL"
XX
SQ   Sequence 2798 BP; 815 A; 544 C; 578 G; 861 T; 0 other;
     atccattcac tgactaacac aaggacaagt ttggagtgat ctgaactctg ggaagcctgt        60
     ggccagggaa gcctgcactg aggacagctg cttcctcagc tgctgtgtag actgagttcc       120
     atcaggcagt ggtaggactt tgaaagcaga gacatcctta aacaatcaga agaacaaaat       180
     gggagaacct gggaaaaggg ttggaatcca cagggtcagg tgctttgcca agatcaaggt       240
     gtttctgttg gcattaatat gggcatatat atccaaaata ctatcaggag tttacatgag       300
     tactatgctc acacaattag agagacaatt caatatttcc acatctatag ttggacttat       360
     caatgggagc tttgagatgg gtaacctttt ggtgattgta ttcgtgagtt attttggaac       420
     aaaactgcat agacctatca tgattggtgt tggttgtgca gttatgggcc tagggtgttt       480
     cataatatca ctacctcatt tcctcatggg cagatacgaa tatgaaacaa caatttcacc       540
     tacaagcaac ttgtcctcaa acagcttttt gtgtgtggaa aacagatccc agaccttaaa       600
     gccaacacaa gacccagcag agtgtgtgaa agaaattaaa tcattaatgt ggatatatgt       660
     actggtagga aacattatac gtggaattgg tgaaactccc atcatgcctt taggtatttc       720
     ctatatagaa gactttgcca aatcagaaaa ttctccttta tacattggaa ttttagaagt       780
     tgggaagatg attggcccaa tacttggata tttgatggga cctttctgtg caaacattta       840
     tgtagacaca gggtctgtga atacagatga cctgaccata actcccactg atacacgctg       900
```

```
ggtcggtgct tggtggattg gctttttggt ctgtgcagga gtgaatgtcc tgaccagcat       960

ccccttttc ttctttccaa aaacactccc aaaggaagga ttacaggata atggggatgg      1020

aactgaaaat gccaaagagg agaagcacag agacaaggcc aaggaggaaa accaaggaat      1080

cattaaagaa ttcttcctta tgatgaagaa cctcttctgt aaccctattt acatgctttg      1140

cgtccttaca agtgtgctcc aggtaaatgg agttgccaat attgtgattt acaagcctaa      1200

atacctggaa catcattttg gaatctccac agcaaaggca gtcttcctca ttggtcttta      1260

taccacacct tcagtatctg ctggatattt aattagtggt tttattatga agaagttgaa      1320

gattactctc aagaaagctg caatcatagc actttgccta ttcatgtctg agtgcctttt      1380

atccctttgt aactttatgc taacctgtga taccactcca attgccggct taactacctc      1440

ttatgaagga attcagcagt cttttgatat ggagaataag tttctttctg actgcaacac      1500

aaggtgtaac tgcttaacaa aaacatggga tccagtgtgt gggaacaatg gcctagcata      1560

catgtcacct tgccttgcag gctgtgaaaa gtctgttgga acaggagcca acatggtgtt      1620

tcaaaattgc agctgcattc ggtcatcagg aaactcatct gcagtcctgg ggctgtgtaa      1680

gaaaggccct gactgtgcta acaagcttca gtacttttta atcataacgg tattttgctg      1740

cttcttctac tcgttagcaa ccatacctgg gtacatggtt tttctgagat gtatgaagtc      1800

tgaagagaag tcacttggaa ttggattaca ggcatttttc atgagactat ttgctggtat      1860

tcctgcacct atttactttg gcgctttgat agacagaaca tgcttacatt ggggaactct      1920

gaaatgtggt gagccaggag catgcaggac ctatgaagtc agtagtttca ggcgcctcta      1980

tcttggattg cctgcagctc taagaggatc aatcattctt ccttcattct tcattctaag      2040

acttatcagg aaactccaaa tccctgggga cactgactct tcagaaattg aacttgcaga      2100

gacgaagccc acagagaagg aaagtgagtg cacagacatg cacaaaagtt ctaaggtcga      2160

gaacgatgga gaactgaaaa ctaagctgta atgaggtttc tactggccta tgcaaggcca      2220

cgaacagaat actcatttca tttcctttga atcataagag aaataatagg aaccctcatc      2280

tttaaggacc tcaaaagcta tttttctcat tataaaaata attactgata ttattttcag      2340

aacttcaggg tagcacttaa gattttccta gtgaagactt taatggtgac ccccaccctg      2400

gactttaaaa agccttcgtt ttcaaagagc attttctctt taaactcagt caaaggaaat      2460

gtgtgtttct tgcatatctt caagtagatt tcatttcact taatttcatt gaatttacat      2520

ttcaatattg gaggtaatta gagctgaaag tatgccttct ggttgtgtca tattgaaata      2580

aattgttcag attcatcctt tccatgtgca aggtgtctgc atgtgtcttt aactctttgg      2640

gagctgttat ctttcttttc tcattctaga cttttgatgc ttcagagatt agactctcac      2700

taatgtgtca tctcgtgttt tcaattccct ctttcattat tcatgtcaca tatttgatca      2760

ttttgtttag aactctgaca aatttaaaca ggttatta                              2798
```

Explanation for the two-letter abbreviations in EMBL-Bank flatfiles: ID, identification; SV, sequence version; AC, accession number; DT, date; DE, description; KW, keyword; OS, organism species; OC, organism classification; RN, reference number; RP, reference positions; RX, reference cross-reference; RA, reference author; RT, reference title; RL, reference location; DR, database cross-reference; CC, comments; FH, feature table header; FT, feature table data; SQ, sequence header; XX, spacer line.

As mentioned already, the EMBL-Bank and DDBJ sequence flatfile (DDBJ flatfile is not shown here) has the "A," "T," "G," and "C" content of the sequence listed (highlighted). The GenBank sequence flatfile does not contain this field. The EMBL flatfile maintains the sequence version number separately as SV, and does not tag it with the accession number. The date of the original submission as well as the last update of 23 September, 2008, creating version 2, are also highlighted.

5.4.5 Sequence Accession Numbers and Redundancy in Primary Databases

An accession number is a unique identifier for a sequence record, which applies to the complete record. It is usually a combination of a letter(s) and numbers. The databases GenBank, EMBL-Bank, and DDBJ all receive sequence submissions, assign accession numbers, and exchange data. Assignment of accession numbers is done following prior agreement within the INSDC collaboration. *When assigning accession numbers, each database uses certain accession prefix that it "owns." In other words, the prefix of an accession number indicates the database where the sequence information was originally submitted.* For example, AJ271682 and AF208545 are two different accession numbers of the same mRNA sequence. The mRNA (as cDNA) was cloned by two different laboratories. From the accession number prefix it is clear that AJ271682[19] (termed Oatp4) was submitted to EMBL-Bank, whereas AF208545[20] (termed rlst-1a) was submitted to GenBank. This mRNA is currently known by various names, such as Oatp4/rlst-1a/Oatp1b2/Slc21a10/Slco1b3. The accession number format for the nucleotide and protein sequence, as well as the details of the accession prefix used by different databases, can be found on the NCBI website[f].

Nucleotide: 1 letter + 5 numerals (e.g. J00750)
or 2 letters + 6 numerals (e.g. AF208545)
Protein: 3 letters + 5 numerals (e.g. AAG60350, CAB92299).

As indicated by the examples above, the sequence information of a specific gene/mRNA can be submitted by multiple authors in the primary databases because different groups may end up cloning the same mRNA and gene. Therefore, there is redundancy of sequence information in the primary databases. Although not frequent, some submitted sequences may also be contaminated with transposon sequence or unremoved vector sequence, adapter sequence, etc. Various sources of contamination of submitted sequence are discussed on the NCBI web page http://www.ncbi.nlm.nih.gov/VecScreen/contam.html. In order to help sequence submitters check their cloned sequence for possible contamination with vector sequences, the NCBI offers the VecScreen program (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) that checks the sequence against the UniVec vector sequence database. VecScreen also detects contamination with many of the adapters, linkers, and PCR primers commonly used in the most popular cDNA cloning strategies.

5.4.6 Divisions of the NCBI Primary Sequence Database

As stated above, **GenBank** is the NCBI primary sequence database, which is a collection of nucleotide and amino-acid sequences from many sources. This primary sequence database has been divided into many categories in order to organize the sequence information in many different ways to facilitate the search and use of a specific type of sequence information. For example, the **Entrez Nucleotide** database consists of three subdivisions: the expressed sequence tag database (**dbEST**), genome survey sequence database (**dbGSS**), and coreNucleotide database (all other nucleotides); a search in the coreNucleotide database returns results from all three. The EST (expressed sequence tag) database is a collection of short single-pass sequence reads of cDNAs (hence mRNA derived); the GSS (genome survey sequence) database is a collection of short single-pass sequence reads of genomic DNA; **HomoloGene** is a system or tool that retrieves homolog information in response to a query from completely sequenced eukaryotic genomes; the **HTG** (high-throughput genome) sequence database is a collection of both unfinished and finished high-throughput genome sequences produced by large-scale genome sequencing centers; the **SNP** (single nucleotide polymorphism) database is a database of various single nucleotide substitutions, short deletion-insertion polymorphisms (DIPs), retroposable element insertions, and microsatellite repeat variations (short tandem repeats or STRs), where each entry includes

[f]For detailed information on accession number and prefix, visit http://www.ncbi.nlm.nih.gov/Sequin/acc.html.

TABLE 5.1 Three-Letter Abbreviations of GenBank Divisions

1	PRI	Primate sequences
2	ROD	Rodent sequences
3	MAM	Other mammalian sequences
4	VRT	Other vertebrate sequences
5	INV	Invertebrate sequences
6	PLN	Plant, fungal, and algal sequences
7	BCT	Bacterial sequences
8	VRL	Viral sequences
9	PHG	Bacteriophage sequences
10	SYN	Synthetic sequences
11	UNA	Unannotated sequences
12	EST	Expressed sequence tag sequences
13	PAT	Patent sequences
14	STS	Sequence tagged sites sequences
15	GSS	Genome survey sequences
16	HTG	High-throughput genomic sequences
17	HTC	Unfinished high-throughput cDNA sequences
18	ENV	Environmental sampling sequences

the sequence surrounding the polymorphism, the occurrence frequency of the polymorphism (by population or individual), and the metadata, such as experimental method(s) and conditions[21]; the **RefSeq** (reference sequence) database is a collection of non-redundant, curated, and richly annotated sequences; the **STS** (sequence tagged sites) database is a collection of STSs (each STS occurs only once in the genome, hence is a unique sequence); the **UniGene** database is a collection of transcript sequences (ESTs, full-length mRNA sequences, alternatively spliced forms) that are derived from the same transcription locus, including pseudogenes, together with information on gene expression, protein similarities, etc.

The **GenBank** sequence database is also divided in a different way into 18 divisions. The GenBank division to which a record belongs is indicated with a three-letter abbreviation, as shown in Table 5.1.[22] The organismal divisions (such as PRI, ROD, MAM) are a convenient way to divide the larger sequence database into smaller segments for those who want to FTP[g] the database.

5.4.6.1 More on the Reference Sequence (RefSeq) Database

The Reference Sequence (RefSeq) database of the NCBI provides a solution to the redundancy and other potential errors in the primary databases. The RefSeq database is a collection of non-redundant, curated, and annotated sequences. RefSeq provides a single record for each natural biological molecule (DNA, RNA, or protein) for major organisms ranging from viruses to bacteria to eukaryotes. Each RefSeq sequence record is created by integrating all or a large fraction of the relevant available information into one non-redundant and richly annotated sequence. In other words, RefSeq is a synthesis of all the information obtained and integrated from multiple sources. Although the RefSeq database is non-redundant, the RefSeq collection does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, orthologs, paralogs, and alternative haplotypes.[23] A RefSeq flatfile looks like the regular GenBank flatfile shown above, except that it has a RefSeq accession number and a COMMENT section. The RefSeq flatfile lists all the sources from where information about the sequence has been obtained, and the COMMENT section cites the accession number(s) of the sequence record(s) used to derive the RefSeq sequence. The COMMENT section also indicates the status of the record—that is, whether the sequence information has been finalized and validated by NCBI review, as well as information about the protein product.

For example, as discussed above, the accession numbers AJ271682 and AF208545 represent the same mRNA molecule. Subsequent to its cloning, various other laboratories published on the function and expression of this gene as well. The information from 10 such published references was utilized to create a RefSeq sequence record about the rat (*Rattus norvegicus*) solute carrier organic anion transporter mRNA, with the RefSeq accession number NM_031650. Version 1 of the RefSeq record (NM_031650.1) identified it as *Slco1b2* mRNA, but version 2 (NM_031650.2) changed the nomenclature to *Slco1b3* mRNA. The NM_031650.1 and NM_031650.2 versions were not reviewed and curated by the NCBI; hence indicated as PROVISIONAL RefSeq in the COMMENT sections of these versions. The final NCBI review of this sequence record resulted in the validated RefSeq record with version 3 (NM_031650.3). Accordingly, the COMMENT section of version 3 states VALIDATED RefSeq. The COMMENT section cites the primary references used to derive the RefSeq sequence, and also shows other information about the sequence, such as function, transcript variants, etc., and states that the RefSeq record includes a subset of the publications that are available for this gene. The RefSeq record of rat *Slco1b3* full-length transcript (transcript variant 1) is shown below, up to the comment section (the sequence is not shown).

[g]FTP (file transfer protocol) is a standard protocol to transfer files from one location to another through the Internet.

RefSeq sequences have a different format of accession numbers for different entities compared to the accession number format in the primary databases; each accession number has a two-letter prefix and a multiple-number segment separated by an underscore sign. The two-letter prefix indicates the type of sequence. For example, NM_123456 indicates an mRNA sequence, NP_123456 indicates a protein sequence, and NC_123456 indicates a chromosome sequence. The key to RefSeq accession number prefixes is discussed in detail on the NCBI website (http://www.ncbi.nlm.nih.gov/refseq/ → Click "Accession" or directly at http://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly).

The following shows the RefSeq record of the full-length mRNA of rat *Slco1b3* (*Oatp4/rlst-1a/Oatp1b2/Slc21a10*) (the record is shown up to the COMMENT section; the rest is truncated; the fields discussed in the text are highlighted).

Rattus norvegicus solute carrier organic anion transporter family, member 1b3 (Slco1b3), transcript variant 1, mRNA

NCBI Reference Sequence: NM_031650.3

FASTA Graphics

```
LOCUS       NM_031650               3218 bp    mRNA    linear   ROD 25-FEB-2013
DEFINITION  Rattus norvegicus solute carrier organic anion transporter family,
            member 1b3 (Slco1b3), transcript variant 1, mRNA.
ACCESSION   NM_031650
VERSION     NM_031650.3  GI:396080334
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 3218)
  AUTHORS   Takashima,T., Hashizume,Y., Katayama,Y., Murai,M., Wada,Y.,
            Maeda,K., Sugiyama,Y. and Watanabe,Y.
  TITLE     The involvement of organic anion transporting polypeptide in the
            hepatic uptake of telmisartan in rats: PET studies with
            [(1)(1)C]telmisartan
  JOURNAL   Mol. Pharm. 8 (5), 1789-1798 (2011)
   PUBMED   21812443
```

REMARK GeneRIF: investigation of role of OATP1B3 in drug metabolism/distribution: Data indicate that hepatic uptake of telmisartan mainly consists of a saturable process mediated by OATP1B3.

REFERENCE 2 (bases 1 to 3218)

AUTHORS Richert,L., Tuschl,G., Abadie,C., Blanchard,N., Pekthong,D., Mantion,G., Weber,J.C. and Mueller,S.O.

TITLE Use of mRNA expression to detect the induction of drug metabolising enzymes in rat and human hepatocytes

JOURNAL Toxicol. Appl. Pharmacol. 235 (1), 86-96 (2009)

PUBMED 19118567

REFERENCE 3 (bases 1 to 3218)

AUTHORS Weiss,M., Hung,D.Y., Poenicke,K. and Roberts,M.S.

TITLE Kinetic analysis of saturable hepatic uptake of digoxin and its inhibition by rifampicin

JOURNAL Eur J Pharm Sci 34 (4-5), 345-350 (2008)

PUBMED 18573335

REFERENCE 4 (bases 1 to 3218)

AUTHORS Aoki,K., Nakajima,M., Hoshi,Y., Saso,N., Kato,S., Sugiyama,Y. and Sato,H.

TITLE Effect of aminoguanidine on lipopolysaccharide-induced changes in rat liver transporters and transcription factors

JOURNAL Biol. Pharm. Bull. 31 (3), 412-420 (2008)

PUBMED 18310902

REFERENCE 5 (bases 1 to 3218)

AUTHORS Donner,M.G., Schumacher,S., Warskulat,U., Heinemann,J. and Haussinger,D.

TITLE Obstructive cholestasis induces TNF periportal downregulation of Bsep and zonal regulation of Ntcp, -alpha- and IL-1 -mediated Oatp1a4, and Oatp1b2

JOURNAL Am. J. Physiol. Gastrointest. Liver Physiol. 293 (6), G1134-G1146 (2007)

PUBMED 17916651

REFERENCE 6 (bases 1 to 3218)

AUTHORS Cattori,V., van Montfoort,J.E., Stieger,B., Landmann,L., Meijer,D.K., Winterhalter,K.H., Meier,P.J. and Hagenbuch,B.

TITLE Localization of organic anion transporting polypeptide 4 (Oatp4) in rat liver and comparison of its substrate specificity with Oatp1, Oatp2 and Oatp3

JOURNAL Pflugers Arch. 443 (2), 188-195 (2001)

PUBMED 11713643

REFERENCE 7 (bases 1 to 3218)

AUTHORS Ismair,M.G., Stieger,B., Cattori,V., Hagenbuch,B., Fried,M., Meier,P.J. and Kullak-Ublick,G.A.

TITLE Hepatic uptake of cholecystokinin octapeptide by organic anion-transporting polypeptides OATP4 and OATP8 of rat and human liver

JOURNAL Gastroenterology 121 (5), 1185-1190 (2001)

PUBMED 11677211

REFERENCE 8 (bases 1 to 3218)

AUTHORS Choudhuri,S., Ogura,K. and Klaassen,C.D.

TITLE Cloning of the full-length coding sequence of rat liver-specific organic anion transporter-1 (rlst-1) and a splice variant and partial characterization of the rat lst-1 gene

JOURNAL Biochem. Biophys. Res. Commun. 274 (1), 79-86 (2000)

PUBMED 10903899

REFERENCE 9 (bases 1 to 3218)

AUTHORS Cattori,V., Hagenbuch,B., Hagenbuch,N., Stieger,B., Ha,R., Winterhalter,K.E. and Meier,P.J.

TITLE Identification of organic anion transporting polypeptide 4 (Oatp4) as a major full-length isoform of the liver-specific transporter-1 (rlst-1) in rat liver

REFERENCE 10 (bases 1 to 3218)

AUTHORS Kakyo,M., Unno,M., Tokui,T., Nakagomi,R., Nishio,T., Iwasashi,H., Nakai,D., Seki,M., Suzuki,M., Naitoh,T., Matsuno,S., Yawo,H. and Abe,T.

TITLE Molecular characterization and functional regulation of a novel rat liver-specific organic anion transporter rlst-1

JOURNAL Gastroenterology 117 (4), 770-775 (1999)

PUBMED 10500057

COMMENT VALIDATED REFSEQ: This record has undergone validation or preliminary review. The reference sequence was derived from AF208545.2 and AABR06034119.1.
On Jul 19, 2012 this sequence version replaced gi:284055291.

Summary: mediated uptake of a variety of organic anions including taurocholate, bromosulfophthalein and steroid conjugates [RGD, Feb 2006].

Transcript Variant: This variant (1) represents the longest transcript and encodes the longest isoform (1).

Sequence Note: This RefSeq record was created from transcript and genomic sequence data to make the sequence consistent with the reference genome assembly. The genomic coordinates used for the transcript record were based on transcript alignments.

Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications.

As indicated in the COMMENT section of the RefSeq record, one of the two primary records from which this RefSeq is derived has the accession number AF208545.2. This is version 2 of the original submission (REFERENCE #8). The other primary record, with the accession number AABR06034119.1, is a contribution from the Rat Genome Sequencing Consortium.

5.5 SECONDARY DATABASES

Secondary databases are curated, non-redundant databases that are derived from the primary (archival) databases. Multiple entries of the same sequence in primary databases are merged to create a single sequence in the secondary database with extensive annotation derived from all available information on the sequence. The sequence and all the information about it are manually curated. The final sequence flatfile has links to all the original entries about the sequence. For example, the NCBI **RefSeq database**[23] is a secondary database that is a collection of curated, non-redundant, well-annotated sequences including genomic DNA, transcripts, and proteins. In addition to providing a curated, non-redundant, well-annotated set of sequences, the RefSeq database also provides a lot of other information about these sequences, such as characterization, mutation, polymorphism analysis, expression studies, and comparative analyses. As indicated above, the RefSeq database, although non-redundant, does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, in addition to orthologs, paralogs, and alternative haplotypes.

5.5.1 An Example of a Non-Redundant, Curated Secondary Database of Proteins—The Swiss-Prot

One of the best non-redundant and curated secondary databases of proteins is **Swiss-Prot**. Swiss-Prot is now a part of the larger database system called the **Universal Protein Resource Knowledgebase** (**UniProtKB**), which was initiated in 2002 by the UniProt consortium. The UniProtKB consists of two parts: **UniProtKB/Swiss-Prot** (reviewed, manually annotated) and **UniProtKB/TrEMBL** (unreviewed, automatically annotated; TrEMBL = translated EMBL). UniProtKB/Swiss-Prot contains manually annotated records and information obtained from the literature and curator-evaluated computational analysis, whereas UniProtKB/TrEMBL contains computationally analyzed records that still need full manual annotation. The source of the protein sequences in UniProtKB can be multiple, such as translated coding sequence from EMBL-Bank/GenBank/DDBJ nucleotide-sequence databases, Protein Data Bank (PDB) database, Protein Information Resource (PIR) database, and sequences submitted directly to UniProtKB. Differences found between various sequencing reports are analyzed and fully described in the feature table, such as alternative splicing events and polymorphisms. Once in UniProtKB/Swiss-Prot, a protein entry is removed from UniProtKB/TrEMBL[h].

UniProt actually comprises four databases: UniProtKB, UniProt Reference Clusters (**UniRef**), UniProt Archive (**UniParc**), and UniProt Metagenomic and Environmental Sequences (**UniMES**). Of these, UniProtKB (Swiss-Prot and TrEMBL), UniParc, and UniRef are non-redundant databases (hence secondary databases).[24] However, the definition of "non-redundant" varies among these three databases. For UniProtKB/TrEMBL, non-redundancy means *one record for 100% identical full-length sequences in one species*; for UniProtKB/Swiss-Prot, non-redundancy means *one record per gene in one species*; for UniParc, non-redundancy means *one record for 100% identical sequences over the entire length, regardless of the species*; and for UniRef100, non-redundancy means *one record for 100% identical sequences, including fragments, regardless of the species*. In UniParc, each record is characterized by a unique identifier, or UPI. The format of the UniParc identifier is "UPI" followed by a combination of numbers and letters, to a total of 10. For example, identical ubiquitin sequences from various organisms can be found in UniParc record UPI00000006C4. For UniRef, there are three databases—UniRef100, UniRef90, and UniRef50; they merge sequences automatically across species. UniRef100 is non-redundant because identical sequences and subfragments are presented as a single entry.[25] A 2013 article provides updates on the activities at the UniProt resource.[26]

The Swiss-Prot database, which is widely used for sequence and other information on proteins, can be directly accessed at www.uniprot.org or it can be accessed through the **Expert Protein Analysis System** (**ExPASy;** http://www.expasy.org/). *The ExPASy is a resource portal of the Swiss Institute of Bioinformatics (SIB). ExPASy provides access to scientific databases as well as bioinformatic analysis tools*. From the ExPASy home page, the "**Resources A..Z**" link on the left can be clicked to go the alphabetically organized resource page and then the needed link, whether database or analytical tool, can be clicked for further analysis. A UniParc link is also available on this page.

[h] http://www.uniprot.org/

5.6 SOME EXAMPLES OF PUBLICLY AVAILABLE SECONDARY AND SPECIALIZED DATABASES

There are many secondary databases on nucleic acid and protein sequences, as well as on their various attributes, such as expression, structure, function, interactions, etc. In addition, there are also organism-specific databases, disease-oriented databases, toxicogenomic and toxicoproteomic databases, allergen databases, etc. Some of the publicly available databases are listed in Table 5.2.

In Table 5.2, only a few secondary and specialized databases that are publicly available have been mentioned. There are still many other specialized curated databases developed and maintained by various consortia or universities. All these databases could not be discussed because of space limitations.

5.6.1 A Special Note on Various NCBI Databases

It was indicated earlier in this chapter that most examples will be cited from the NCBI/GenBank. A wide

TABLE 5.2 Publicly Available Secondary and Specialized Databases

Database	Comments (with URLs)
Universal Protein Resource Knowledgebase (UniProtKB)	The UniProt Knowledgebase (UniProtKB) is the central repository for the collection of sequence and functional information on proteins with accurate, consistent, and rich annotation. UniProtKB is the product of UniProt, which is an international consortium between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR) at the Georgetown University Medical Center. In 2002, EBI, SIB, and PIR started collaboration to create a single high-quality database of protein sequence and function, by unifying the Swiss-Prot, TrEMBL, and PIR-PSD databases. Before this collaboration, EMBL-EBI maintained TrEMBL, SIB maintained Swiss-Prot, and PIR maintained the Protein Sequence Database (PIR-PSD). These data sets coexisted with different protein-sequence coverage and annotation priorities[26,27] (www.uniprot.org) UniProtKB has two sections: **UniProt/Swiss-Prot** and **UniProt/TrEMBL**. UniProt/Swiss-Prot contains sequences that are manually annotated, compared, and verified (curated) based on information from literature and curator-evaluated computational analysis. UniProt/TrEMBL (TrEMBL = translated EMBL) contains computationally annotated, unreviewed sequences. TrEMBL sequences are eventually manually curated to become part of Swiss-Prot and removed from TrEMBL Before becoming part of UniProt, PIR-PSD was the oldest annotated and curated protein-sequence database, established in 1984 as a successor to the original National Biomedical Research Foundation (NBRF) Protein Sequence Database. It was developed over a 20-year period by the late Margaret Dayhoff and published as the "Atlas of Protein Sequence and Structure" from 1965 to 1978. The link to PIR-PSD is http://pir.georgetown.edu/[28]
Worldwide Protein Data Bank (**wwPDB**)	Experimentally determined structures of proteins, and complex assemblies. wwPDB is a publicly available archive of macromolecular structural data[29] (http://www.wwpdb.org/)
Structural Classification of Proteins (**SCOP**) database	The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the PDB. Proteins are classified into families (clear evolutionarily relationship; this generally means that pairwise residue identities between the proteins are 30% and greater), superfamilies (probable common evolutionary origin), and folds (major structural similarity)[30] (http://scop.mrc-lmb.cam.ac.uk/scop/)
Class, Architecture, Topology, Homology (**CATH**) database	CATH is a manually curated classification of protein domain structures. Each protein is chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques, which include computational algorithms, empirical and statistical evidence, literature review, and expert analysis[31] (http://www.cathdb.info/)
PROSITE database	This consists of a large collection of biologically meaningful signature patterns or profiles. These signatures are not easily revealed by standard sequence alignment. Each signature can be linked to useful biological information on the protein family, domain, or functional site. Therefore, the database can be used to rapidly and reliably identify which known family of protein (if any) the new sequence belongs to. The PROSITE database uses two kinds of signatures, patterns and generalized profiles, to identify conserved regions[32] (http://prosite.expasy.org/)

(*Continued*)

TABLE 5.2 (Continued)

Database	Comments (with URLs)
PRINTS database	This is a compendium of protein fingerprints; a fingerprint is a group of conserved motifs used to characterize a protein family[33] (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php)
Protein Family (**Pfam**) database	Pfam is a comprehensive database of protein families; members of a family share significant similarity, thereby suggesting homology. Pfam allows the analysis of sequence data in order to search for related proteins in the database based on domains. Domains are regions of the protein, which in different combinations can determine the protein's function. Thus, proteins can be viewed as built from a specific combination of domains. Pfam contains two types of families: high-quality manually curated Pfam-A families and automatically generated Pfam-B families. Pfam uses multiple sequence alignments and hidden Markov models (HMM)[34] (http://www.sanger.ac.uk/resources/databases/pfam.html)
InterPro database	InterPro integrates various predictive protein signatures from diverse source repositories, such as Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, and TIGRFAMs. Protein signatures from various databases are integrated into InterPro manually. Curators combine signatures representing the same protein family, domain, or site into single database entries, and, where possible, trace biological relationships between the constituent signatures[35] (http://www.ebi.ac.uk/interpro/)
Biological General Repository for Interaction Datasets (**BioGRID**)	The BioGRID database is an online repository of interactions in which data are curated from both high-throughput data sets and individual focused studies, as derived from over 40,000 publications in the primary literature. The current compilation (as of July, 2013) has more than 700,000 raw protein and manually annotated genetic interactions from major model organisms. All BioGRID interaction records are directly mapped to experimental evidence in the supporting publication[36] (http://thebiogrid.org/)
Molecular Interaction database (**MINT**)	MINT is a public repository for protein−protein interactions reported in peer-reviewed journals. It focuses on experimentally verified protein−protein interactions mined from the scientific literature by expert curators. Currently it contains over 240,000 interaction data captured from over 4750 publications[37,38] (http://mint.bio.uniroma2.it/mint/)
Münich Information System for Protein Sequences (**MIPS**) database	The MIPS mammalian protein−protein interaction database is a resource of high-quality experimental protein-interaction data. The content is based on published experimental evidence that has been processed by human expert curators. MIPS also contains large-scale secondary data of protein similarities, currently containing 38 million non-redundant protein sequences[39,40] (http://mips.helmholtz-muenchen.de/proj/ppi/)
IntAct	IntAct is a freely available, open source molecular interaction database populated by data either curated from the literature or from direct data depositions. As of September 2011, IntAct contained approximately 275,000 curated binary interaction evidence records from over 5000 publications. The IntAct database also captures protein−small molecule (including phospholipids), protein−nucleic acid, and protein−gene locus interactions[41] (http://www.ebi.ac.uk/intact/)
Structural Database of Allergenic Proteins (**SDAP**)	SDAP is a web server that integrates a database of allergenic proteins with various computational tools that can assist structural biology studies related to allergens, including predicting the IgE-binding potential of food proteins. This database allows bioinformatic analysis as recommended by the *Codex Alimentarius* and UN Food and Agriculture Organization (FAO)/World Health Organization (WHO) Expert Committee on potential allergenicity of foods derived through modern biotechnology[a] (http://fermi.utmb.edu/SDAP/)
AllergenOnline/FARRP database (FARRP = Food Allergy Research and Resource Program at the University of Nebraska-Lincoln)	AllergenOnline provides access to a peer-reviewed allergen list and sequence searchable database intended for the identification of proteins, including food proteins, that may present a potential risk of allergenic cross-reactivity. The objective is to identify proteins that may require additional tests, such as serum IgE binding, basophil histamine release, or in vivo challenge to evaluate potential cross-reactivity (http://www.allergenonline.org/)
Allermatch database	The Allermatch database allows the comparison of a protein sequence with sequences of allergenic proteins in the database, in order to predict whether the protein being evaluated can be allergenic. This database allows bioinformatic analysis as recommended by the *Codex Alimentarius* and FAO/WHO Expert Committee on potential allergenicity of foods derived through modern biotechnology[42] (http://www.allermatch.org/)

(*Continued*)

TABLE 5.2 (Continued)

Database	Comments (with URLs)
Online Mendelian Inheritance in Man (**OMIM**) database	OMIM is a comprehensive compendium of human genes and genetic-disease-associated phenotypes. The full-text referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes[b] (http://www.ncbi.nlm.nih.gov/omim/ and http://omim.org/)
ArrayExpress database	A public database of microarray gene-expression data at the EBI. It accepts data generated by sequencing or array-based technologies and currently contains data from almost a million assays, from over 30,000 experiments. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.[43] ArrayExpress uses the minimum information about a microarray experiment (MIAME) annotation standard[c] (http://www.ebi.ac.uk/arrayexpress/)
Gene Expression Omnibus (**GEO**) database	The GEO is a public repository that archives and freely distributes MIAME-compliant microarray data, next-generation sequencing data, and other forms of high-throughput functional genomic data submitted by the scientific community. It is one of three international functional genomics public data repositories, alongside ArrayExpress at the EBI and the DDBJ Omics Archive[44,45] (http://www.ncbi.nlm.nih.gov/geo/)
ArrayTrack database	A public database of microarray gene-expression data at the US Food and Drug Administration. ArrayTrack provides an integrated solution for managing, analyzing, and interpreting microarray gene-expression data and experimental parameters associated with pharmacogenomics or toxicogenomics studies—that is, studies on the effects of drugs or other chemicals on gene expression. ArrayTrack supports MIAME-compliant data[46] (http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm)
Comparative Toxicogenomic database (**CTD**)	This is a public database of information built on curated data from the scientific literature about interactions between environmental chemicals and gene products and their relationships to diseases. As of 2013, CTD contains over 15 million toxicogenomic relationships. A user can look up specific literature-based information about genes, gene products, and toxicants of interest and their interactions[47] (http://ctdbase.org/)
Chemical Effects in Biological Systems (**CEBS**) database	The CEBS database has been developed by the National Center for Toxicogenomics within the National Institute for Environmental Health Sciences (NIEHS). CEBS integrates data obtained using 'omics technologies (transcriptomics, proteomics, metabolomics) as well as from traditional toxicology studies. Thus, CEBS combines the molecular genetic data with traditional clinical chemistry and histopathology data. This combination allows researchers to fully capture information on dose response, time response, and environmental-stress-induced gene expression. The database captures information from multiple species, such as humans, rats, mice, and *Caenorhabditis elegans*[48] (http://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm)
DrugMatrix database	DrugMatrix is a toxicogenomic and molecular toxicology database and informatics system developed by the National Toxicology Program (NTP). It contains data from standard toxicological experiments along with large-scale gene-expression data from various organs and tissues. DrugMatrix contains toxicogenomic profiles for 638 different compounds that include approved drugs, withdrawn drugs, and industrial and environmental toxicants[d] (https://ntp.niehs.nih.gov/drugmatrix/index.html)
FlyBase database	FlyBase is the leading database and web portal for genetic and genomic information focusing on *Drosophila melanogaster*, but also including data on other *Drosophila* species and related drosophilids. The current content of FlyBase comprises > 200,000 references, including > 87,000 research papers from > 2400 different journals, with publication dates ranging from the seventeenth century through to the present day[49,50] (http://flybase.org/)
NCBI databases	**Collection of various databases.** This is separately discussed below, in Section 5.6.1 (http://www.ncbi.nlm.nih.gov/)

[a] *Publications can be accessed at http://fermi.utmb.edu/SDAP/sdap_pub.html.*
[b] *OMIM is authored and edited at the Victor McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr Ada Hamosh. The official home page is www.omim.org.*
[c] *The minimum information about a microarray experiment (MIAME) is a microarray experimental data submission standard that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. The six most critical elements contributing towards MIAME are: (1) the raw data for each hybridization; (2) the final processed (normalized) data; (3) essential sample annotation, including experimental factors and their values (e.g. compound and dose in a dose–response experiment); (4) the experimental design, including sample data relationships (e.g. which raw data file relates to which sample, which hybridizations are technical, which are biological replicates); (5) sufficient annotation of the array (e.g. gene identifiers, genomic coordinates, oligonucleotide probe sequences, or reference commercial array catalog number); (6) the essential laboratory and data-processing protocols (e.g. what normalization method has been used to obtain the final processed data).*[51,52]
[d] *Publications can be accessed at https://ntp.niehs.nih.gov/drugmatrix/contributors.html.*

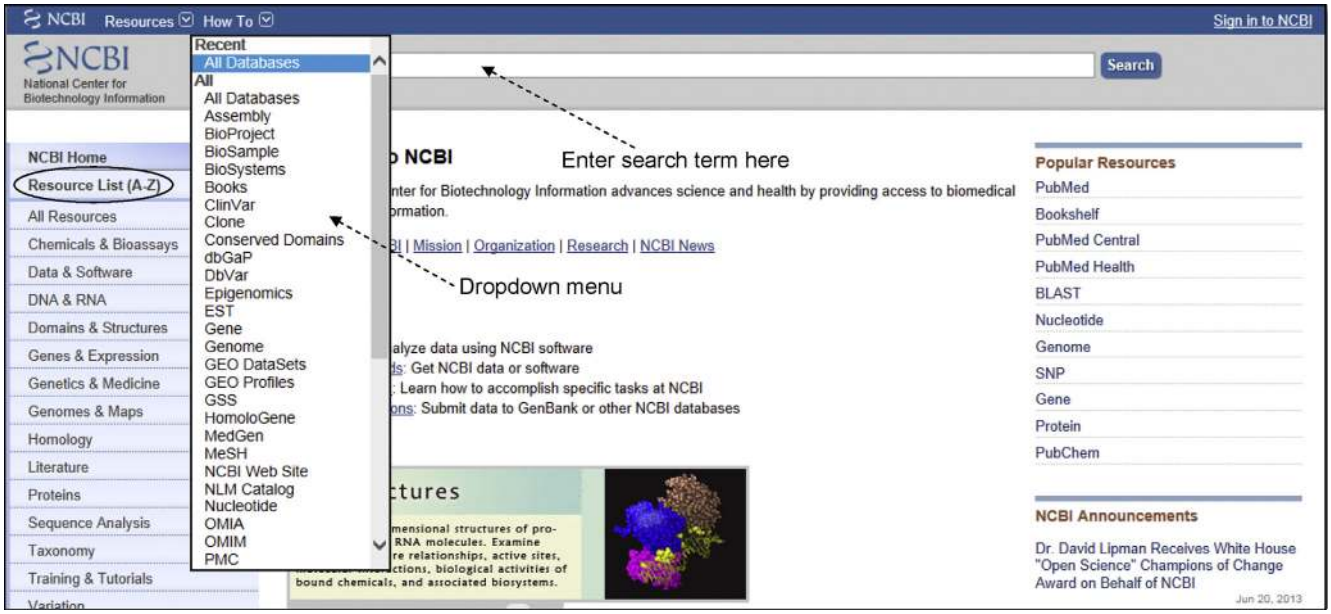


FIGURE 5.1 **Partial view of the NCBI home page (http://www.ncbi.nlm.nih.gov/; as of June, 2013).** A specific database can be selected from the drop-down menu and then the search term can be entered in the space shown. Hitting the "search" button returns the entries.

variety of high-quality resources, such as databases and tools, are made accessible to the public by the NCBI through a common retrieval system.[53,54] The databases are visible in the drop-down menu from the NCBI homepage. Some of the common databases are named below. Additionally, the link "**Resource List (A-Z)**" located at the left-hand top corner of the NCBI home page can be clicked to obtain links to all resources, including all the databases, browsers etc., organized alphabetically. Below the "Resource List (A-Z)", there is the link "**All Resources**." This link lists a specific class of resources under one tab; hence the "databases" tab lists all databases, "tools" tab lists all analysis tools, etc. (Figure 5.1).

Some of the widely used databases are **PubMed** (bibliographic database); **OMIM** (Online Mendelian Inheritance in Man; described above); the **Entrez Nucleotide database** (described above); the **Gene Expression Omnibus (GEO) database** (described above); the **Protein database** (curated sequences are in RefSeq); the **Genome database** (contains information on sequence, annotation, maps, chromosomes, and assemblies of all organisms whose genomes have been sequenced so far, and provides graphic display through the genomic browser Map Viewer); the **Structure database** (contains three-dimensional images of proteins); the **Gene database**[i] (contains information about individual genes from among the genomes represented in the RefSeq); the **Taxonomy database** (contains the names of all organisms that are represented by nucleotide or protein sequences); the **UniGene database** (contains non-redundant information on computationally identified transcripts from the same locus across species; described above); and the **Epigenomics database** (a relatively new database that provides epigenomic data in the context of biological sample information).

5.7 DATA RETRIEVAL

Data retrieval from different databases requires a search capability using a data retrieval system (tool). Some common data retrieval systems are **Entrez/GQuery**, **DBGET/LinkDB**, **Sequence Retrieval System** (**SRS**), and retrieval system from EMBL-EBI. Retrieval systems are capable of simultaneously searching multiple linked databases in response to a single search query and retrieve related data from multiple databases. *It is worth emphasizing at the outset that the appearance and functionality of various web-based resources are subject to frequent change. Therefore, various screenshots displayed here may change by the time this book is published. Nevertheless, knowing how to use the tools by following the screenshots presented in the book should still help the readers to understand and cope with the changes.*

[i]**Gene** is described as a searchable database of genes in the NCBI "Resource" section. However, Gene is also described as a portal that integrates gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data, using information from a wide range of resources, such as RefSeq maps, pathways, and genome- and locus-specific resources. From a user's perspective, **Gene** acts as a single-source specialized database containing information on specific genes across different species.

5.7.1 Search and Retrieval Using Entrez/GQuery

Entrez (GQuery, or global query; http://www.ncbi.nlm.nih.gov/sites/gquery) is a user-friendly, versatile, text-based search and retrieval system developed by the NCBI. It searches linked databases using a single word or combination of words entered as search term. Thus, Entrez provides a global query system and forms a web of connections with the databases (nodes in the web of connections). The search at the NCBI can be performed either using a specific database, or using Entrez across databases simultaneously.

Figure 5.1 shows the databases (partial list) that can be selected from the drop-down menu on the NCBI home page, and then the search term can be entered in the space shown. Hitting the "search" button will usually return a number of entries. Depending on the database selected for search and retrieval, the primary source of some of the retrieved entries may be other related but specialized databases. For example, the Nucleotide, RefSeq, EST, GSS, and Gene databases all have entries on the same nucleotide sequence or part thereof, under database-specific accession numbers and descriptors. Because all these databases are linked, selecting the Nucleotide database for searching a sequence will retrieve all entries related to the sequence from other related and specialized databases as well. However, selecting a specialized database will retrieve a smaller number of entries.

Alternatively, the user can access the Entrez home page and perform a search across all databases simultaneously by entering the search term in the space shown. Hitting "Search" will return the number of entries available in each database, which is displayed next to the database name. The Entrez home page has recently undergone a change in appearance. Figures 5.2A and 5.2B show a partial view of the Entrez home page. A screenshot of the Entrez home page captured in March 2013 is shown in Figure 5.2A, whereas a screenshot captured in June 2013 is shown in Figure 5.2B. These two screenshots are shown to underscore the fact that the appearance or versions of bioinformatic tools and database home pages are subject to change, although the utility pretty much remains the same and is mostly improved. The Entrez home page states GQuery (global query) now, and the order of database display has been reorganized in the new version. Both Figures 5.2A and 5.2B show only the top portion of the retrieved information that was obtained by performing a search using the search term "Mus musculus Slco1a6." Figures show the number of hits in various databases; PubMed has 2 and PubMed Central has 10 entries (as of June 2013), Nucleotide database has 10 entries (visible in Figure 5.2A but not in Figure 5.2B). Other databases not shown in the figure also have different numbers of entries. Clicking on the number or on the database name will return all the entries from that database. Without the data retrieval system, such simultaneous searching across multiple databases by entering the search term only once is not possible and individual databases have to be searched separately.

The simultaneous search capability and all-in-one display of results from multiple databases make the NCBI Entrez (GQuery) a user-friendly search and retrieval system for general users.

5.7.2 Search and Retrieval Using DBGET/LinkDB

DBGET/LinkDB (http://www.genome.jp/dbget/dbget_manual.html) is an integrated text-based search and retrieval system for major biological databases at GenomeNet. GenomeNet is the Japanese network of database and computational services for genome research and related biomedical research; it is operated by the Kyoto University Bioinformatics Center (http://www.bic.kyoto-u.ac.jp/). DBGET searches and extracts entries from a wide range of molecular biology databases, and LinkDB searches and computes links between entries in divergent databases. Databases being searched can exist in different servers, but from the user's point of view, they all exist in a single DBGET server.[55]

DBGET/LinkDB uses three basic commands for performing search and retrieval of database entries: **bfind**, **bget**, and **blink**. **bget** retrieves database entries based on a search combination (name:identifier), **bfind** retrieves database entries by keywords, whereas **blink** retrieves related entries in a given database as well as all databases.

5.7.3 Search and Retrieval Using Sequence Retrieval System

Examples of some publicly available Sequence Retrieval System (SRS) servers are http://www.embnet.sk:8080/srs81/; http://www.dkfz.de/srs/; http://iubio.bio.indiana.edu/srs/. There are many other such web-based servers, too. Figure 5.3 shows various services available from EMBL-EBI (http://www.ebi.ac.uk/services) that includes sequence retrieval functions as well. These can be accessed by clicking the "DNA & RNA" as well as "Proteins" links. A search in dbfetch (http://www.ebi.ac.uk/Tools/dbfetch/dbfetch/) requires the accession number, as shown in Figure 5.4. A search for multiple sequences can also be made by using multiple search terms and separating them using a comma.

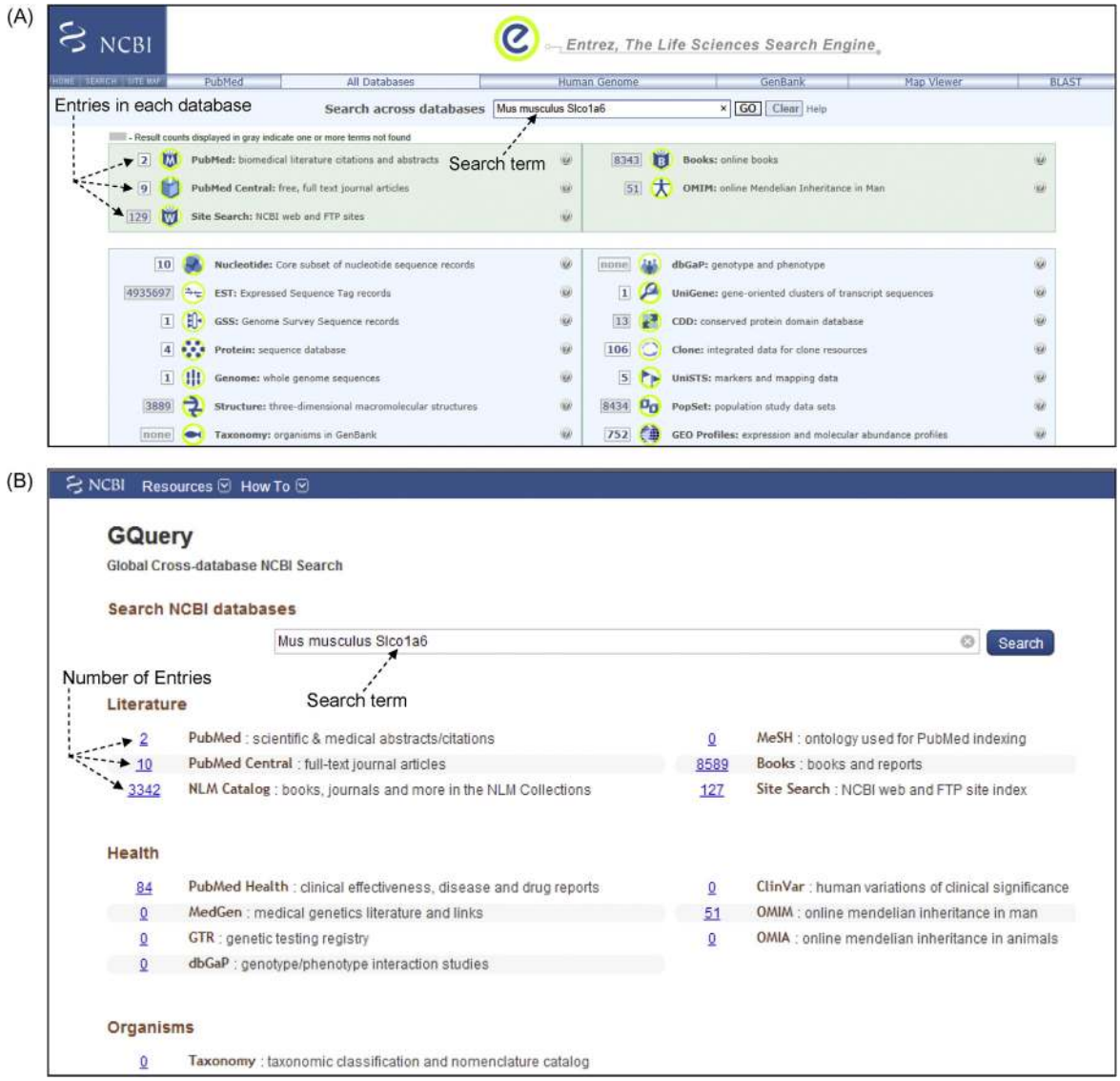


FIGURE 5.2 **Partial view of the Entrez home page at two different dates.** (A) A screenshot of the Entrez home page captured in March 2013. (B) A screenshot of the Entrez home page captured in June 2013. These two screenshots are shown to underscore the fact that the home page is subject to change, although the utility pretty much remains the same and is mostly improved. The Entrez home page states GQuery now. A user can perform a search across all the databases simultaneously by entering the search term in the space shown. Hitting "Search" will return the number of entries available in each database, displayed next to the database name. This may change with time as new information is added to various databases.

5.8 AN EXAMPLE OF RETRIEVAL OF MRNA/GENE INFORMATION

Information about an mRNA or gene[j] can be retrieved by selecting the "Nucleotide" (database) from the drop-down menu on the NCBI home page (Figure 5.1). The Nucleotide (database) provides a link to the grand collection of all nucleotide sequences from the primary as well as the specialized databases. A search using the mRNA or gene name in the Nucleotide databases retrieves many records, and depending on the search term the number of records may sometimes be too many to go through individually. The Nucleotide database can be searched in different ways to focus the search more

[j]The display of information output associated with any database is subject to change from time to time. This is because there is continuing effort to improve the information output and display features. Therefore, the graphic displays shown in the figures are not expected to remain the same all the time. Nevertheless, knowing how to harness and use the information should prepare readers to deal with any such changes.

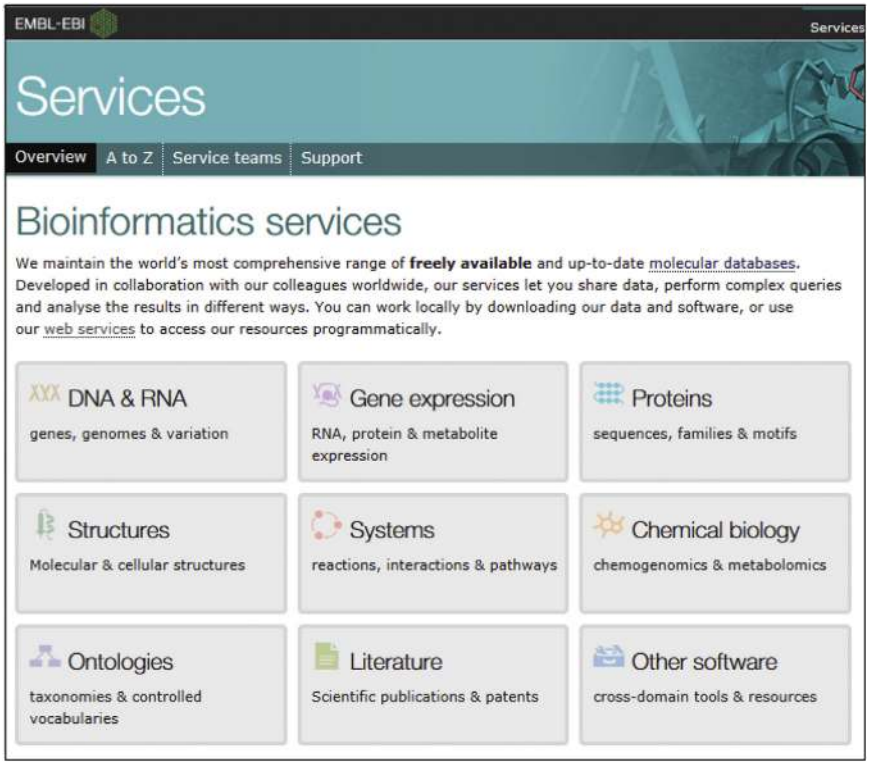


FIGURE 5.3 **Data Retrieval at EMBL-EBI.** Nucleotide sequence data can be retrieved by clicking the "DNA & RNA" link and accessing the ENA resource. Protein sequence data can be retrieved by clicking the "Protein" link and accessing the protein resource, such as UniProt. *(Source: EMBL-EBI, http://www.ebi.ac.uk/services).*

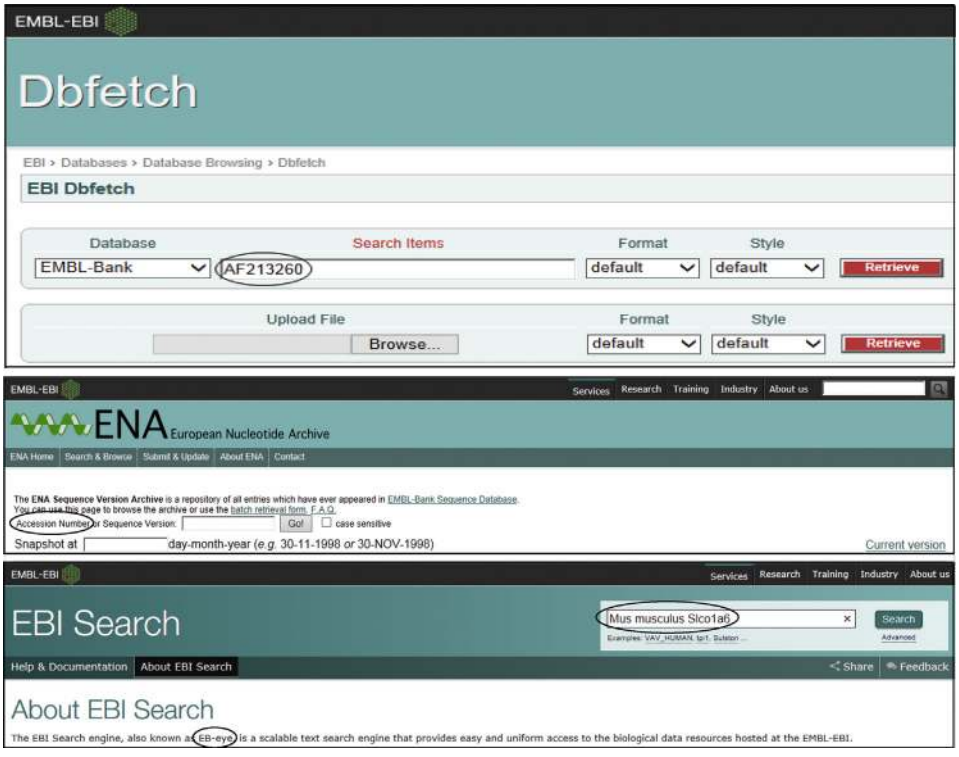


FIGURE 5.4 **Search and retrieval using dbfetch, ENA, and EB-eye.** Specific sequence information from the EMBL-Bank can be retrieved using **dbfetch** (upper panel), **ENA** (middle panel), and **EB-eye** (lower panel). These are partial screenshots.

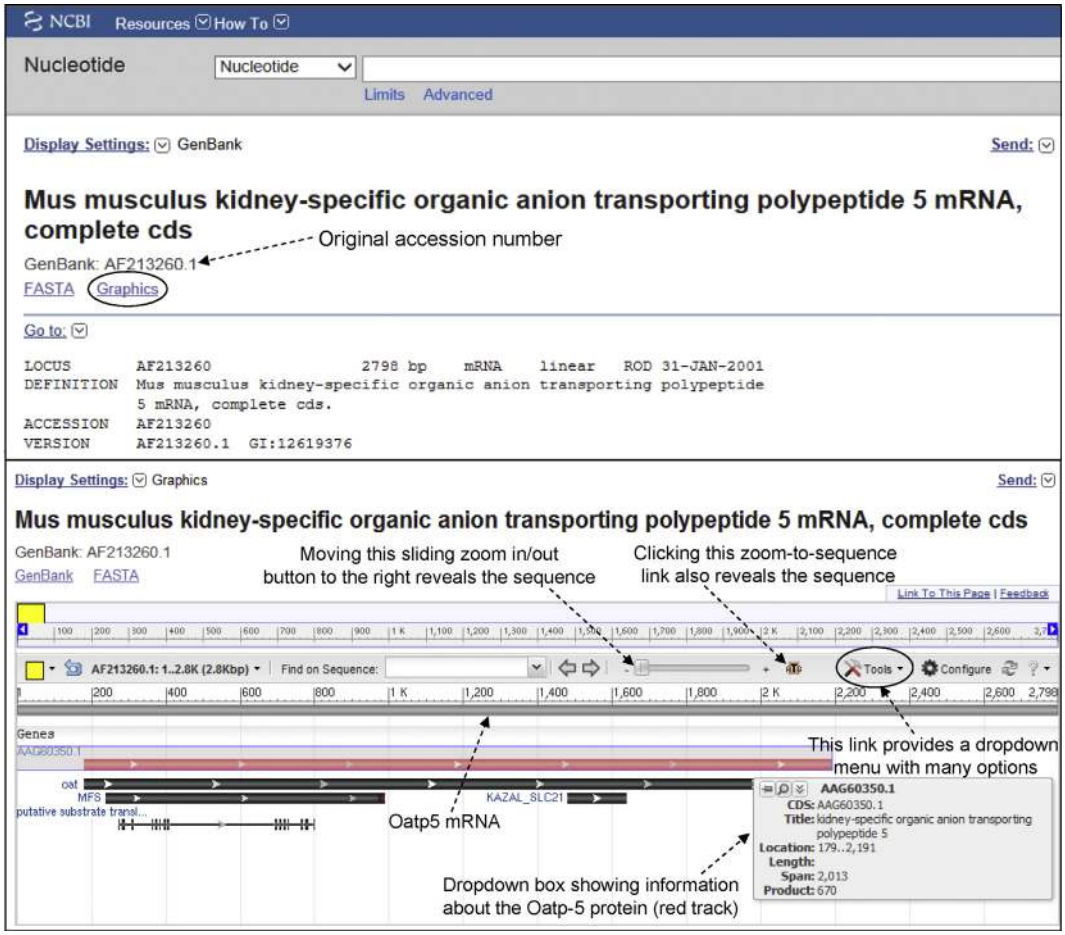


FIGURE 5.5 **GenBank information on mouse Oatp-5.** The upper panel shows the top portion of the GenBank record of the original submission of mouse Oatp-5 mRNA along with its accession number and the version. Below the accession number is the link to the graphics (circled). Clicking the graphics link will return the graphics of the mRNA and the protein shown in the lower panel. The lower panel also shows various links and tools in the Graphics page that can help visualize different aspects of the sequence as described in the text. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

narrowly, such as by utilizing the accession or GI number or even using the names of the authors of a submission. Of course, the user has to know this type of information. If the accession number or GI number of a sequence is known, the exact record can be directly retrieved. Currently, the GenBank nucleotide record provides a link to graphics of the sequence.

For example, Figure 5.5 (upper panel) shows the top portion of the GenBank record of the **original submission of mouse Oatp-5 mRNA**.[56] Mouse Oatp-5 was later given other names, such as Slc21a13 and Slco1a6, of which **Slco1a6 is the name used in all databases**. Slco1a6 stands for "solute carrier organic anion transporter (Slco) member 1a6." **In the text that follows, both the terms Oatp-5 and Slco1a6 will be used**. The flatfile of this original submission (accession: AF213260) has been shown before. Figure 5.5 upper panel shows the link to the graphics (circled). Clicking the graphics link will return the graphics of the mRNA and the protein and other relevant information shown in Figure 5.5 lower panel, Figure 5.6, and Figure 5.7, along with various links and tools that can help visualize different aspects of the sequence. The same graphical representation (and more) can also be retrieved by using the Gene database (discussed later). The red-colored track represents the mouse Oatp-5 protein. If the cursor is brought onto the track, a drop-down box appears that contains information about the red track; for example, the Oatp-5 coding sequence spans from base 179 to 2191, and the Oatp-5 protein contains 670 amino acids (Figure 5.5, lower panel). The figure shows a sliding zoom-in/out button; moving the button to the right first zooms in the figure and ultimately reveals the nucleotide sequence on the black track at the top, along with the corresponding amino-acid sequence on the red track. Alternatively the "zoom-to-sequence" link can be clicked to reveal the sequence. This automatically moves the sliding zoom-in/out button all the way to the right.

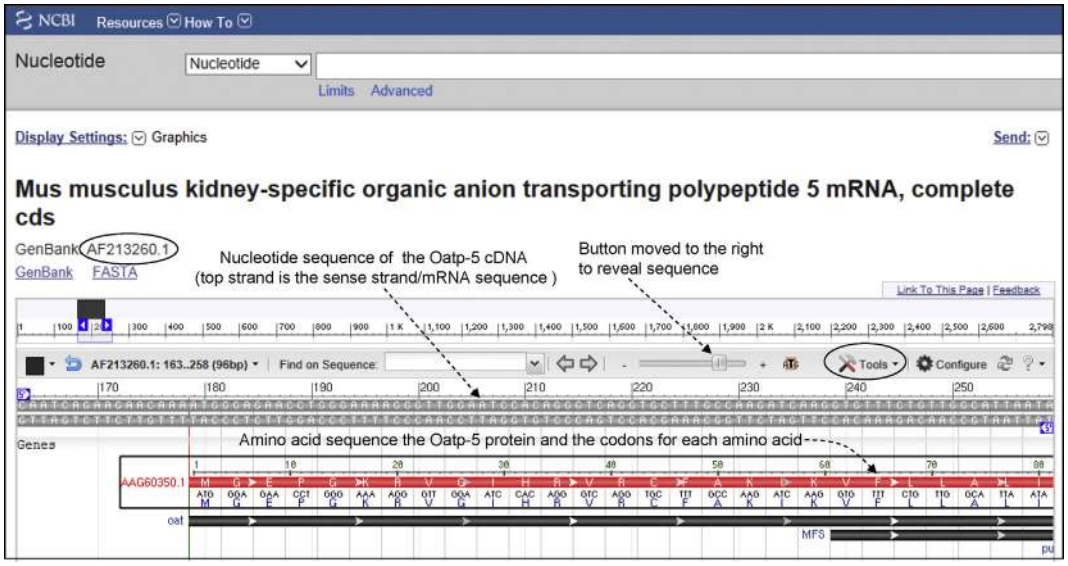


FIGURE 5.6 **The zoom-in state of the record shown in Figure 5.5 (lower panel), showing the sequence.** The figure shows the nucleotide sequence of Oatp-5 cDNA at the top, associated with the black track; and the amino-acid sequence of the Oatp-5 protein along with the codons for each amino acid, associated with the red track. The coding sequence begins from base 179, which is the "A" of "ATG." *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

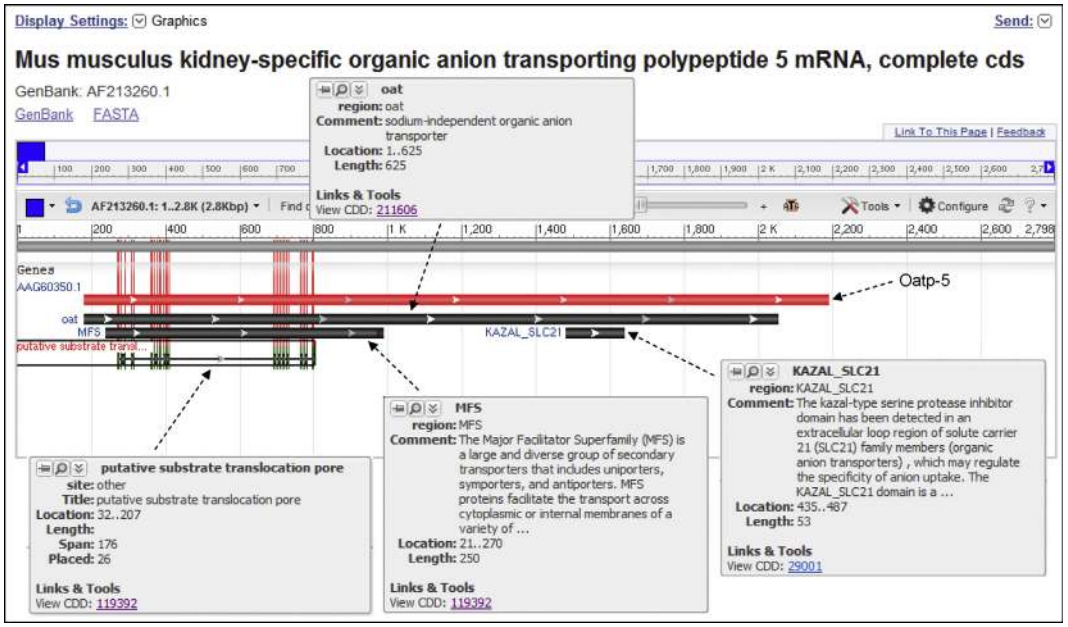


FIGURE 5.7 **A modified composite screenshot of the record shown in Figure 5.5 (lower panel).** The information on all the tracks in Figure 5.5 (lower panel) were separately captured and pasted to artificially create this figure. The figure shows the individual drop-down information boxes associated with each track. Note that it is not possible to obtain all the information drop-down boxes at the same time. This is because the cursor can be held only on one track at a time to obtain the drop-down information box.

The zoom-in state showing the sequence is shown in Figure 5.6 (partial sequence shown). It shows the nucleotide sequence of Oatp-5 cDNA at the top associated with the black track, and the amino-acid sequence of the Oatp-5 protein along with the codons for each amino acid associated with the red track. It is clear from Figure 5.6 that the coding sequence begins from base 179, which is the "A" of "ATG." Figure 5.7 is a modified composite figure (see the legend for Figure 5.7).

Compared to the the original submission (AF213260.1), the RefSeq record of Oatp-5 (called Slco1a6, with an accession number NM_023718 version 3) has more

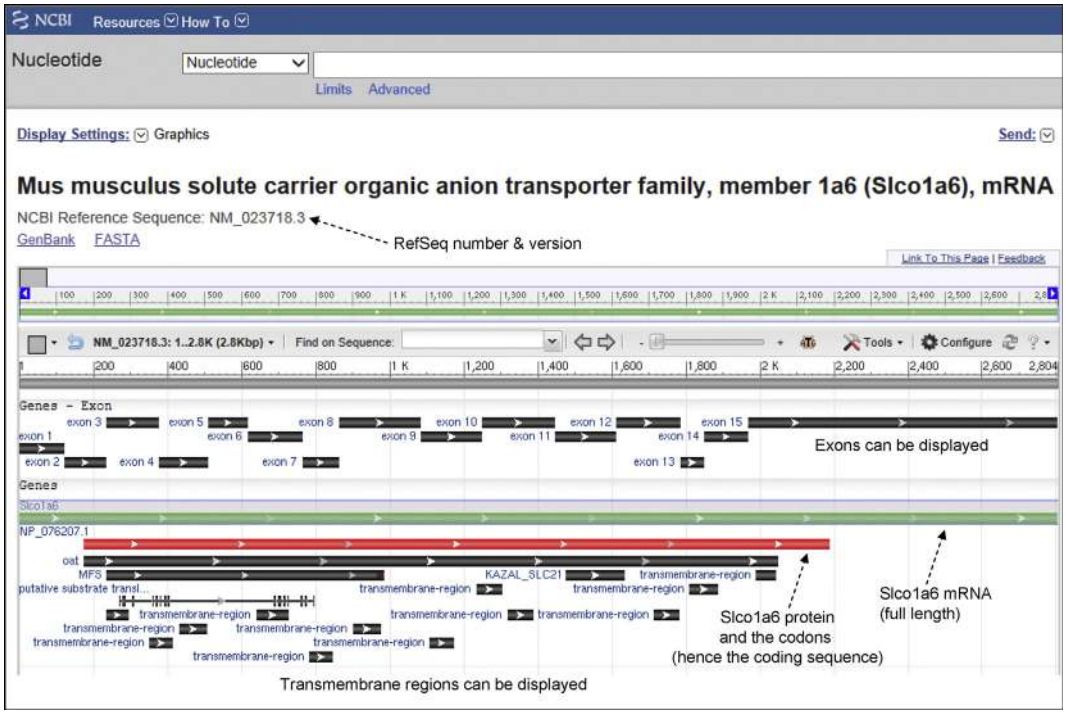


FIGURE 5.8 **The graphics of the RefSeq record for Oatp-5.** In the RefSeq record, Oatp-5 is identified as Slco1a6. The graphics of the RefSeq record show additional information that was not present in the original submission, such as information on the length and span of exons in mRNA, and the transmembrane regions in the protein. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

graphics available. Figure 5.8 shows the graphics of the RefSeq record, which identifies Oatp-5 as Slco1a6. The graphics of the RefSeq record show additional information that was not present in the original submission (Figures 5.5 and 5.6), such as information on the length and span of exons in mRNA and on transmembrane regions in the protein.

Figure 5.9 was created by first zooming in Figure 5.8 to reveal the sequence and then separately capturing and pasting the information about all the tracks to the screenshot; hence Figure 5.9 is an artificially created screenshot. As mentioned above, all the drop-down information boxes cannot be obtained at the same time; the cursor can be held on one track at a time so that the information about that track appears in the drop-down box. In these graphics, the green track represents the entire length (1...2804) of the *Slco1a6* (*Oatp5*) mRNA, and is associated with an information box. The red track represents the Slco1a6 protein along with the amino-acid codons; hence the red track also shows the coding sequence (base 175...2187). The graphics of the RefSeq record also displays information about all the exons. Figure 5.9 shows that exon 3, for example, is 142 bp long (235...376). Thus, base 235 through 376 of the *Slco1a6* mRNA is derived from exon 3 of the *Slco1a6* gene. Slco1a6 is a membrane transporter with more than 10 transmembrane regions (transmembrane domains or TMDs). Figure 5.9 shows that the first TMD of Slco1a6 is 20 amino acids long and spans from amino acid 21 to 40 (21...40). The UniProtKB/Swiss-Prot accession number of mouse Slco1a6 is Q99J94, and this is a curated entry; hence, the information has been validated.

Note that the original submission (AF213260.1) shows the coding sequence spanning from base 179 to 2191, but the RefSeq record (NM_023718.3) shows the coding sequence spanning from base 175 to 2187. This difference reflects an adjustment of four bases in the 5′-UTR of the RefSeq record compared to the original record. This was done during the creation and validation of the RefSeq record, which involved comparison with the Slco1a6 gene sequence record from the mouse reference genome.[57] Therefore, the information in the RefSeq record should be regarded as more accurate and up to date.

At the left-hand top corner of Figure 5.9, there is a link to "Display Settings"; next to it is "Graphics" (circled). The "Display Settings" is a drop-down menu that provides many options for viewing the sequence information. When the "Graphics" option is chosen, the information is displayed as graphics as in Figure 5.9 and other similar figures. Figure 5.10 shows information about the sequence in a different ("Revision History") format. Choosing the "Revision History" option from the "Display Settings" drop-down menu displays the entire history of revision of the sequence. Figure 5.10

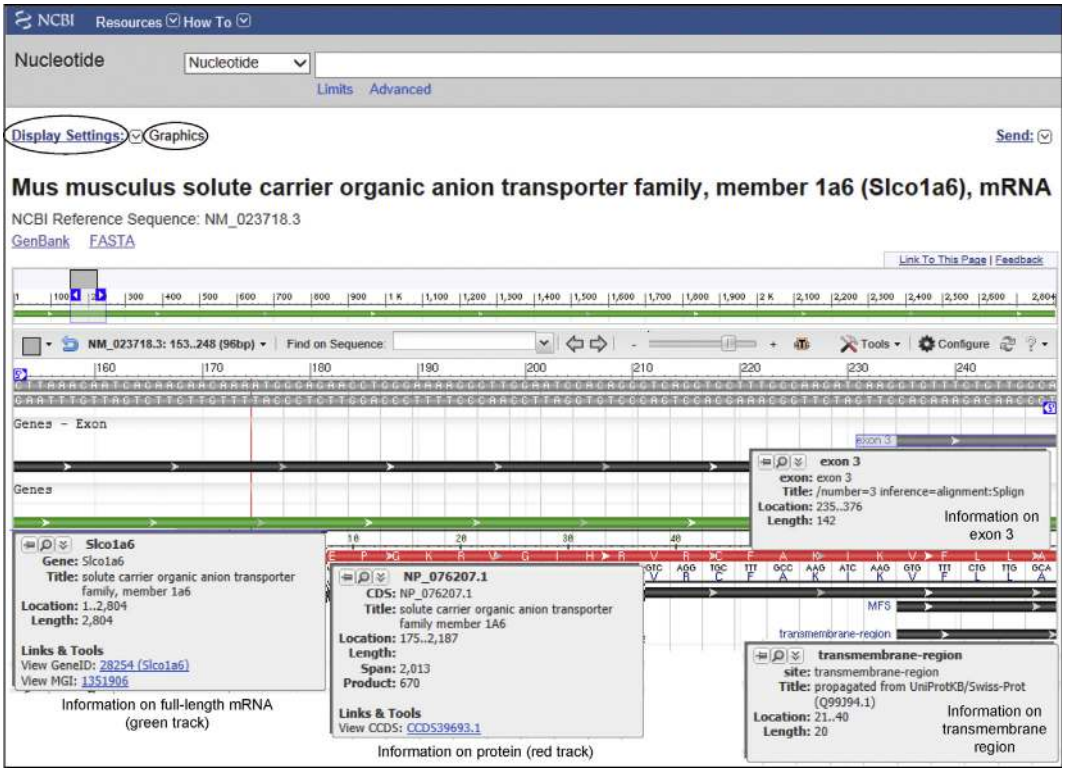


FIGURE 5.9 **A modified composite screenshot of the record shown in Figure 5.8 showing the individual drop-down information boxes associated with each track.** See text for details.

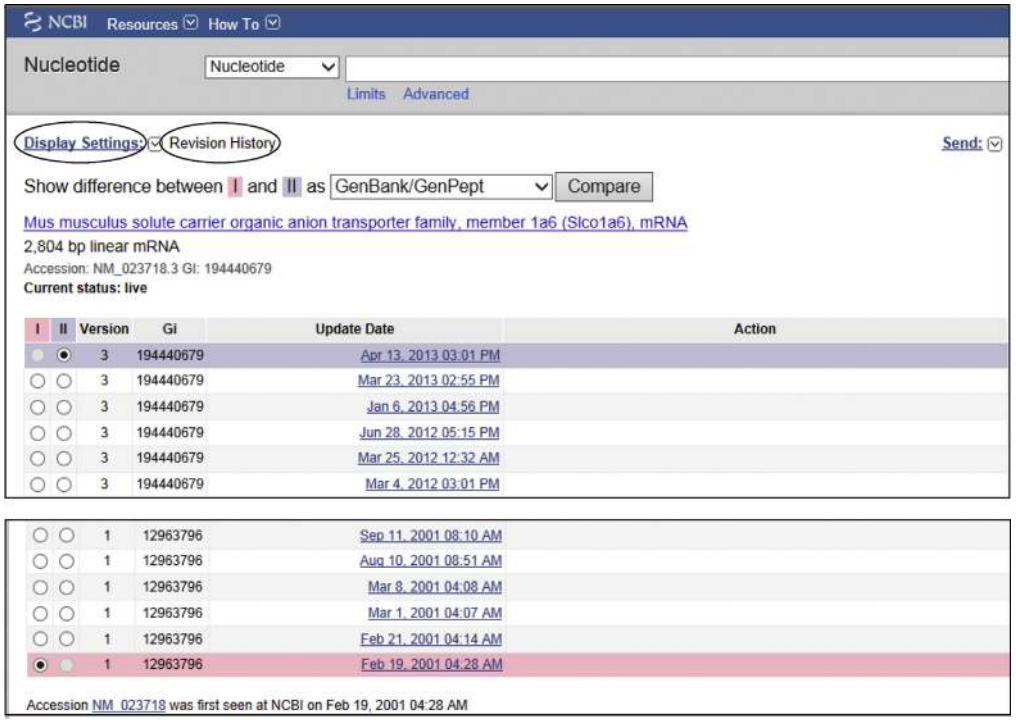


FIGURE 5.10 **The "Revision History" of Slco1a6.** The upper panel shows the upper part of the list and the lower panel shows the lower part of the list. By selecting two specific entries a comparison can be made to find out the revisions made in the sequence. The figure shows that the first and the last entry of the *Slco1a6* mRNA sequence have been selected for comparison. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

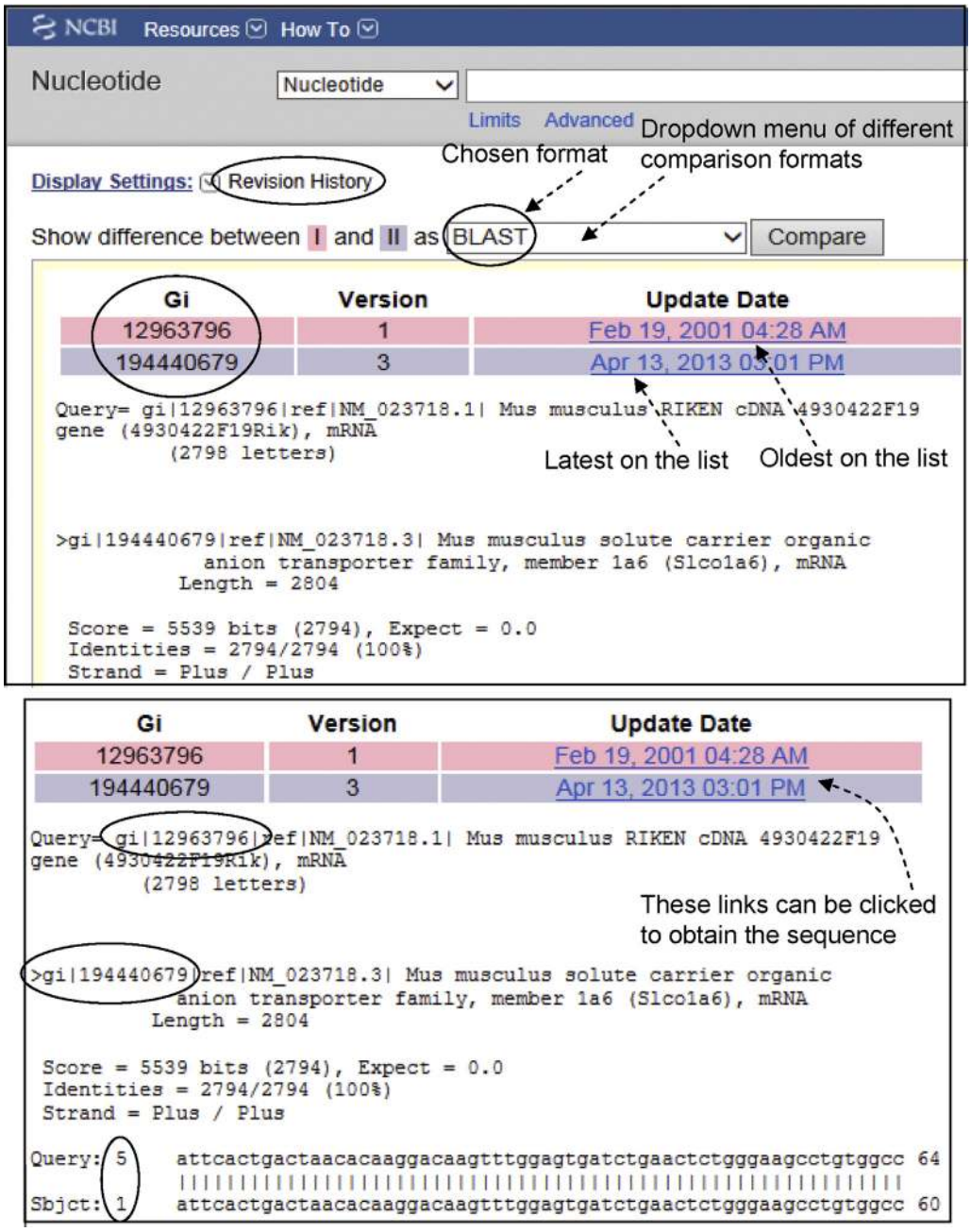


FIGURE 5.11 **Results of the comparison of the two versions of *Slco1a6* mRNAs selected in Figure 5.10.** The upper panel shows that the comparison format of the revision history from Figure 5.10 is BLAST pairwise alignment. The lower panel shows only the first 60 bases from the pairwise alignment. Base 1 of the Sbjct sequence starts aligning with base 5 of the Query sequence; this suggests that the original sequence entry (Query) with the GI number 12963796 had four extra bases at the beginning of the sequence that are not present in the latest entry (Sbjct) with the GI number 194440679. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

upper panel shows the upper part of the list and the lower panel shows the lower part of the list (the whole list is too long to display in one page). By selecting two specific entries, a comparison can be made to find out the revisions made in the sequence. Figure 5.10 shows that the first and the last entry of the *Slco1a6* mRNA sequence have been selected for comparison. Figure 5.11 shows the result of that comparison. Figure 5.11 upper panel shows that the comparison format chosen from the drop-down menu is BLAST pairwise alignment. The lower panel shows only the first 60 bases from the pairwise alignment. It shows that the alignment starts from base 5 of the original sequence entry (Query; GI number 12963796), indicating that the

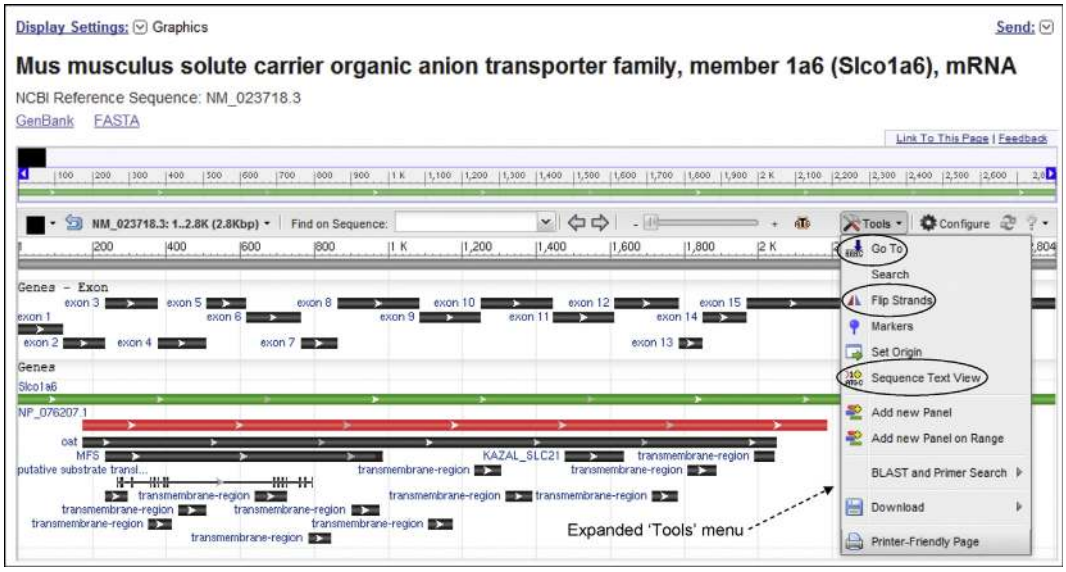


FIGURE 5.12 **The expanded "Tools" drop-down menu, showing its options.** See text for explanation. *(Source: http://www.ncbi.nlm.nih.gov/ →Nucleotide, information as of June 2013)*

original sequence entry had four extra bases (atcc) at the beginning of the sequence that are not present in the latest entry (Sbjct; GI number 194440679). Hence, base 1 of the Sbjct sequence starts aligning with base 5 of the Query sequence; the rest of the Query and Sbjct sequences are identical. These extra four bases (atcc) could have been a cloning/sequencing artifact in the original submission. This is why the original submission (AF213260.1) shows the coding sequence spanning from base 179 to 2191, but the RefSeq record (NM_023718.3) shows the coding sequence spanning from base 175 to 2187, reflecting an adjustment of four bases.

In the screenshots shown in Figures 5.5–5.9, there is a link to a "Tools" drop-down menu, which is shown expanded in Figure 5.12 to show the available options. Three such options are circled. The "Go To" option allows the user to go to a specific position in the sequence; the "Flip Strands" option allows the user to flip the polarity of the sequence; the "Sequence Text View" option allows the user to view the entire nucleotide sequence as well as the amino-acid sequence.

A search for Oatp-5/Slco1a6 can also be performed using the Gene database. Figure 5.13 shows the results of a query in the Gene database using the search term "Oatp-5" (circled in the figure) performed in June 2013. The search retrieved just two records, one for mouse, and one for rat. As indicated before, Oatp-5 is also known by two other names, Slco1a6 and Slc21a13. Each entry shows the official symbol, name, other aliases, other designations, chromosomal location, map position, and the RefSeq annotation information. For example, the second entry is mouse Oatp-5. Its official symbol is Slco1a6, other alias is Slc21a13, it is located on chromosome 6, it spans from nucleotide (nt) 142085768 to nt 142186149 on the reverse strand. Therefore, the mouse *Oatp5* gene is 100,382 bp long, and the Gene database ID is 28254, which can be used to retrieve the record directly from the Gene database.

If the mouse Slco1a6 result is clicked to open the detailed record, this record contains 10 information fields. These fields, shown in Figure 5.14, have been collapsed to fit the screen. Three fields will be discussed here: the "Summary" field, the "Genomic context" field, and the "Genomic regions, transcripts, and products" field. Other fields can be likewise expanded and explored for their information content.

The "**Summary**" field with its detailed information content is shown in Figure 5.15; the figure also shows the detailed information content of the "**Genomic context**" field. The "Summary" field shows that the official symbol Slco1a6 is provided by the Mouse Genome Informatics (MGI) group[k,58]. The *Slco1a6* gene has an ID MGI:1351906, which can be used to search for it in MGI databases. The link to MGI:1351906 can be clicked to obtain the Slco1a6 page of MGI (Figure 5.16). The inset in Figure 5.16 is actually located to the far right on the Slco1a6 page; it has been moved to fit the screenshot. The MGI Slco1a6 page shows its map

[k]MGI (http://www.informatics.jax.org/) is the international database resource that provides integrated genetic, genomic, and biological data for the laboratory mouse.

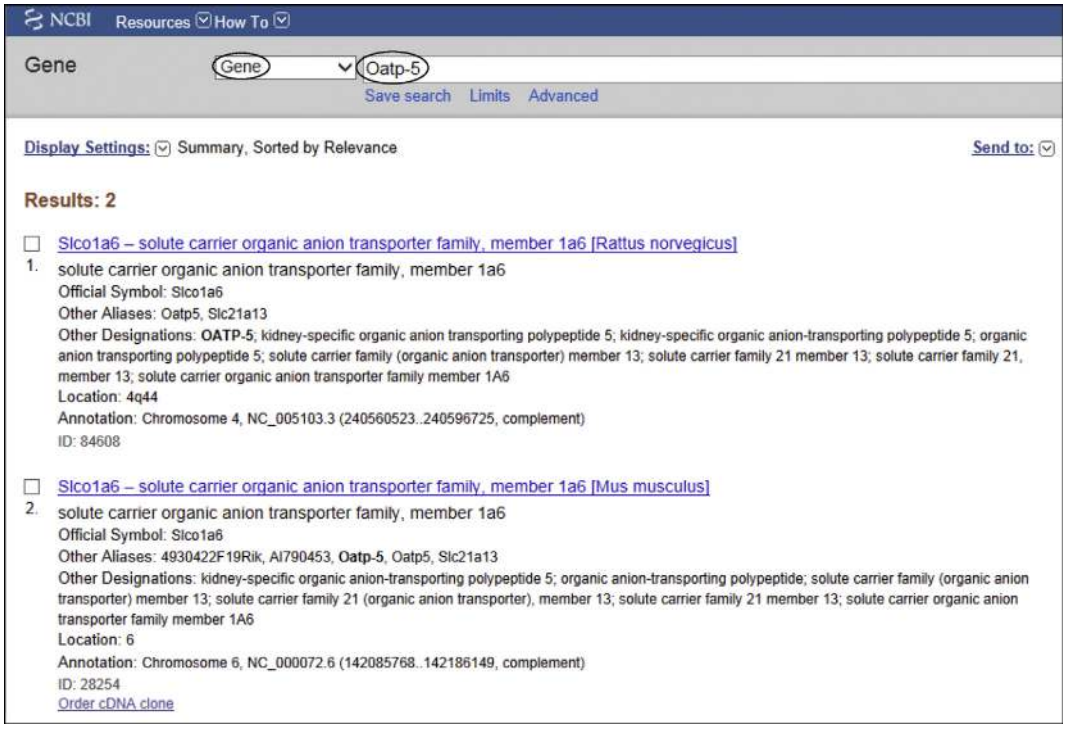


FIGURE 5.13 **The result of a query in the Gene database using the search term "Oatp-5" (circled).** See text for explanation. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

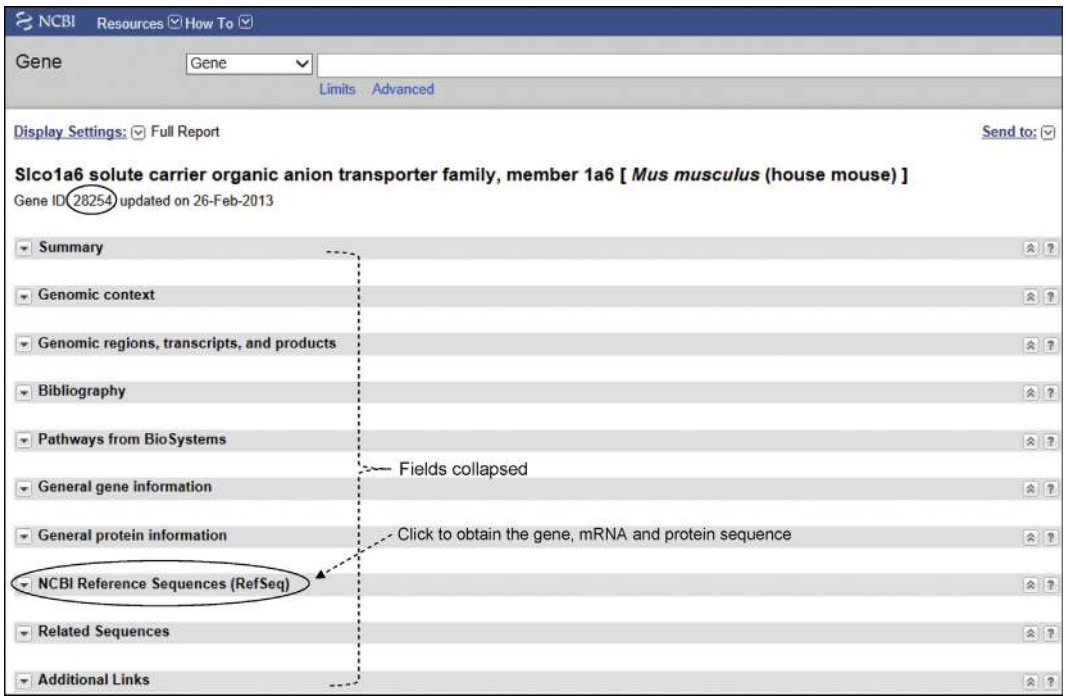


FIGURE 5.14 **The detailed record for the mouse Slco1a6 entry in Figure 5.13.** The detailed record shows 10 information fields. Each field can be clicked to expand. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

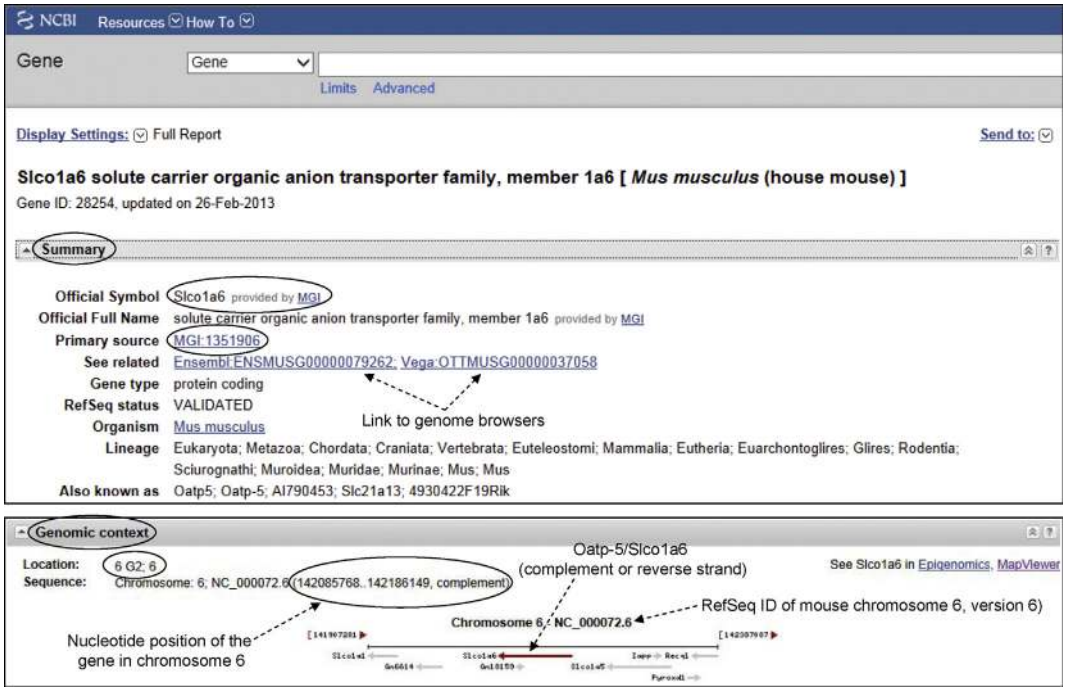


FIGURE 5.15 **The detailed information content of the "Summary" and "Genomic context" fields from the mouse Slco1a6 detailed record in Figure 5.14 after the fields are expanded.** The "Summary" field (upper panel) shows that the official symbol Slco1a6 is provided by the Mouse Genome Informatics (MGI) group. The *Slco1a6* gene has an ID MGI:1351906, which can be used to search for it in the MGI database. The "Genomic context" field (lower panel) shows the chromosomal and genomic location of the *Slco1a6* gene. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

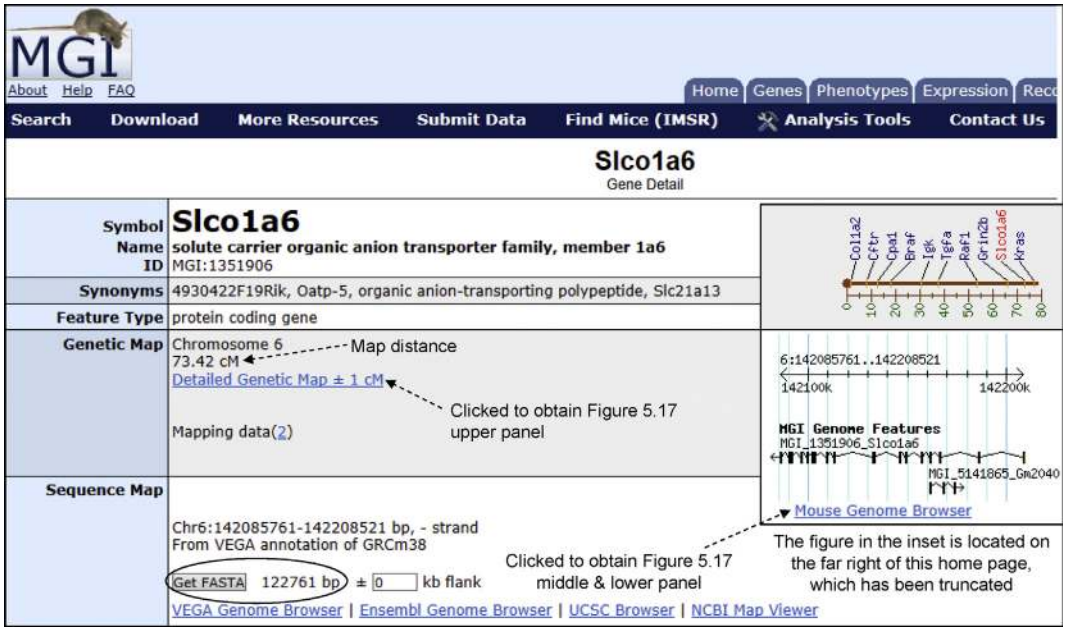


FIGURE 5.16 **Truncated screenshot of the MGI Slco1a6 page.** The figure in the inset is located to the far right on the actual Slco1a6 page. Because of the truncation of the Slco1a6 page to fit the figure, the inset has been copied and pasted close to the rest of the information. The page shows the genetic map position of the *Slco1a6* gene. The Slco1a6 page provides a lot of information and links to other information resources (see text). *(Source: http://www.informatics.jax.org/→MGI Slco1a6 page, information as of March 2013)*

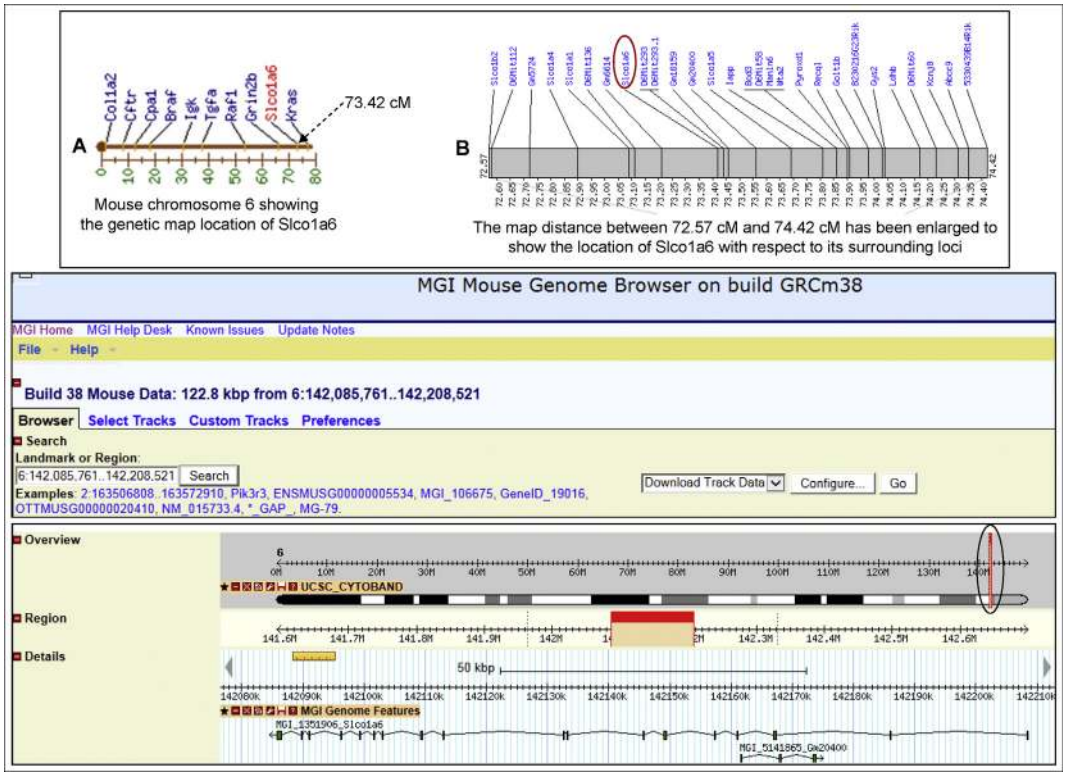


FIGURE 5.17 **Figure created by pasting three partial screenshots from the MGI pages on Slco1a6.** The upper panel was obtained by clicking the "Detailed Genetic Map ± 1 cM" link from Figure 5.16. It shows the chromosomal location of the *Slco1a6* locus in greater resolution with respect to the surrounding loci. The middle and lower panels were obtained by clicking the "Mouse Genome Browser" link shown in the inset in Figure 5.16. *(Source: http://www.informatics.jax.org/→MGI Slco1a6 page, information as of March 2013)*

position as 73.42 cM[l], which is with respect to position 0 at one end of the chromosome. Mouse chromosome 6 is an acrocentric chromosome—that is, the centromere is located almost at one end, creating an extremely short p arm and a very long q arm. The 0 position in the genetic map starts at one end of the chromosome near the centromere; so the *Slco1a6* gene with its genetic map position at 73.42 cM lies very close to the other end of chromosome 6 (Figure 5.17, upper panel). The MGI Slco1a6 page provides links to sequence map display on four genome browsers: VEGA, Ensembl, UCSC, and NCBI Map Viewer (Figure 5.16). However, the "Summary" field of the Gene database search record itself also provides links to the Ensembl and VEGA genome browsers (Figure 5.15). The "Sequence Map" field of the MGI Slco1a6 page also provides a "Get FASTA" link to the entire gene sequence in FASTA format from VEGA annotation of mouse genome build 38 (GRCm38[m]). Note that the total number of nucleotides is 122,761 bp (higher than 100,382 bp mentioned earlier; Figure 5.16, "Sequence Map" field, link circled). The Slco1a6 page has much more information (not shown here), that can be clicked and explored. Figure 5.17 is a composite figure that has been created by pasting three partial screenshots. The upper panel was obtained by clicking the "Detailed Genetic Map ± 1 cM" link from Figure 5.16. It shows the chromosomal location of the *Slco1a6* locus in greater resolution with respect to the surrounding loci. The middle and the lower panels were obtained by clicking the "Mouse Genome Browser" link (shown in the inset in Figure 5.16). Viewing sequence maps on genome browsers will be discussed later. Other links on the Slco1a6 page can be clicked to explore more information.

[l]1 centiMorgan (1 cM) = 1 map unit distance between two genes or genetic markers.

[m]GRC is an acronym for Genome Reference Consortium and m38 means the 38th version (build 38) of mouse genome sequence assembly. The GRC is responsible for assembling the human and mouse reference genomes, and in that process correct misrepresented loci and close remaining assembly gaps. The members of GRC include The Genome Center at Washington University, the Wellcome Trust Sanger Institute, the EBI, and the NCBI. The GRC website (http://www.genomereference.org) is available to view the progress of various projects, and communicate with the scientific community in general.

The "**Genomic context**" field with its detailed information content is shown in Figure 5.15, lower panel. The "Location" line on the left of the Genomic context field (Figure 5.15, lower panel) shows 6G2. This means that the *Oatp5/Slco1a6* gene maps to region G, band 2 of chromosome 6. Because mouse chromosomes are acrocentric (centromere almost at the end of the chromosome), creating an extremely short p arm and a very long q arm, sometimes the q arm is not mentioned. Therefore, the location can be expressed as both 6G2 and 6qG2. Below the location line is the "Sequence" line that shows "Chromosome: 6; NC_000072.6 (142085768…142186149, complement)." The NC_000072.6 is the RefSeq ID (accession number) for *Mus musculus* chromosome 6 (see Table 5.3), version 6; the "142085768…142186149" means that the *Oatp5/Slco1a6* gene spans from nt 142085768 to 142186149; hence, the gene is 100382 bp long. The "complement" means that the gene is located on the reverse strand of the chromosome[n]. Note that this nucleotide location span of the gene is based on the build 38 (GRCm38), which is the latest version of mouse genome sequence assembly as this section is being written. Below the location field, there is a diagram showing the chromosomal location of *Oatp5/Slco1a6* in relation to other closely linked genes, such as *Slco1a1*, and *Slco1a5*. The direction of the arrow is from right to left, indicating that the *Oatp5/Slco1a6* gene is on the reverse (minus) strand of the chromosome. In other words, the direction of transcription is from right to left.

Another direct way of obtaining the gene, mRNA, and protein sequences through the Gene database is the "NCBI Reference Sequence (RefSeq)" field. Figure 5.14 shows this field circled towards the bottom. Expanding this field provides links to the *Slco1a6* gene sequence in chromosome 6, *Slco1a6* mRNA, and Slco1a6 protein (with their respective RefSeq accession numbers). By clicking these links one can directly obtain the gene, mRNA, and protein sequences.

The "**Genomic regions, transcripts, and products**" field with its detailed information content is shown in Figure 5.18. The upper panel shows the gene (as a horizontal green line) with all the exons and introns, whereas the lower panel shows the sequence. The gene information is based on build 38 of the mouse genome assembly (GRCm38; circled); the field also shows the chromosome information (chromosome 6). If the "Graphics" link in the right-hand top corner (circled) is clicked, the chromosome 6 graphics page appears (Figure 5.19). The mRNA and protein sequences of Slco1a6 can be directly obtained by clicking the "Go to reference sequence details" link in the right-hand top corner (circled) (Figure 5.18).

The details of the exon and intron sequence information can be obtained by clicking "Display Settings" in the left-hand top corner and selecting "Gene Table" from the drop-down menu (Figure 5.20; circled; this

TABLE 5.3 RefSeq IDs (Accession Numbers) of Various Chromosomes in Human, Rat, and Mouse

	RefSeq ID of Chromosomes		
Chr #	***Homo sapiens***	***Rattus norvegicus***	***Mus musculus***
1	NC_000001	NC_005100	NC_000067
2	NC_000002	NC_005101	NC_000068
3	NC_000003	NC_005102	NC_000069
4	NC_000004	NC_005103	NC_000070
5	NC_000005	NC_005104	NC_000071
6	NC_000006	NC_005105	NC_000072
7	NC_000007	NC_005106	NC_000073
8	NC_000008	NC_005107	NC_000074
9	NC_000009	NC_005108	NC_000075
10	NC_000010	NC_005109	NC_000076
11	NC_000011	NC_005110	NC_000077
12	NC_000012	NC_005111	NC_000078
13	NC_000013	NC_005112	NC_000079
14	NC_000014	NC_005113	NC_000080
15	NC_000015	NC_005114	NC_000081
16	NC_000016	NC_005115	NC_000082
17	NC_000017	NC_005116	NC_000083
18	NC_000018	NC_005117	NC_000084
19	NC_000019	NC_005118	NC_000085
20	NC_000020	NC_005119	
21	NC_000021		
22	NC_000022		
X	NC_000023	NC_005120	NC_000086
Y	NC_000024		NC_000087

The version numbers are not shown here because they may change when a new assembly is reported

[n]Each chromosome (in an unduplicated state) is composed of one DNA molecule; hence two DNA strands. The DNA strand whose 5′-end is closer to the centromere is called the forward strand of the chromosome; the other strand is the reverse strand (or complement). Therefore, the direction from p→q arm of the chromosome is the same as the 5′→ 3′ direction of the forward strand. The sense strand (coding strand) of some genes resides in the forward strand whereas that of others resides in the reverse strand (complement) of the chromosome.

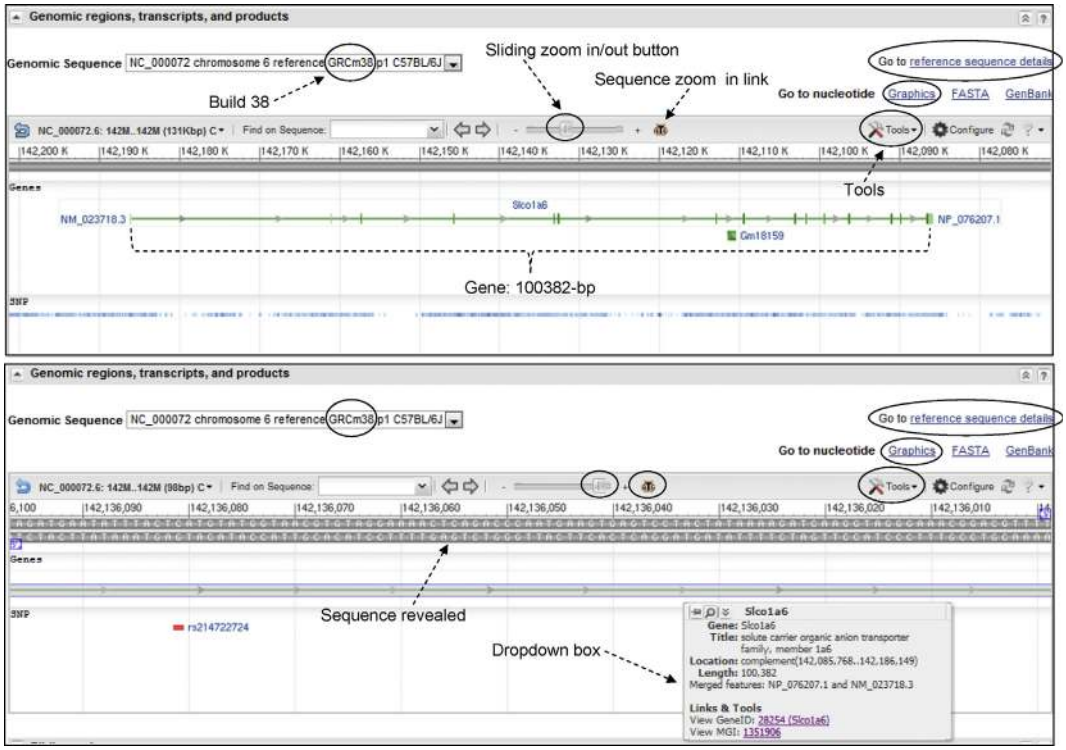


FIGURE 5.18 **The "Genomic regions, transcripts, and products" field from the mouse Slco1a6 detailed record in Figure 5.14 after the field is expanded.** Upper panel showing the gene with its exons and introns; lower panel showing the sequence. The gene information is based on build 38 of the mouse genome assembly (GRCm38). The RefSeq links to the mRNA and protein sequences of Slco1a6 can be directly obtained by clicking the "Go to reference sequence details" link in the right-hand top corner (circled). *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

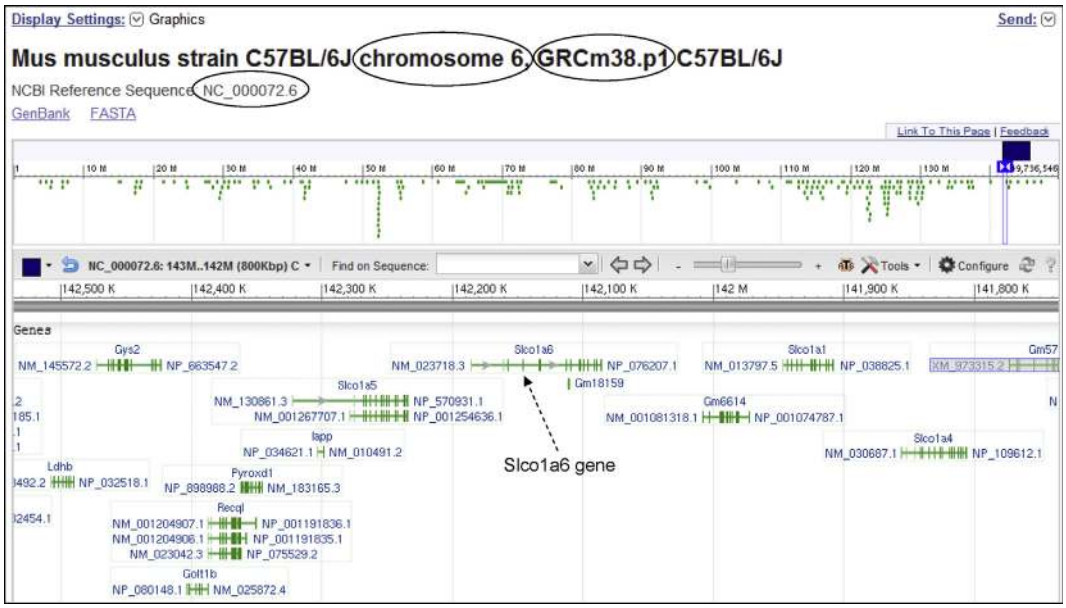


FIGURE 5.19 **The chromosome 6 graphics page, from the "Graphics" link in Figure 5.18.** The span of chromosome 6 shown is approximately 0.9×10^6 bp long, and it contains many genes, including many transporter genes. The vertical bars represent the exons. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

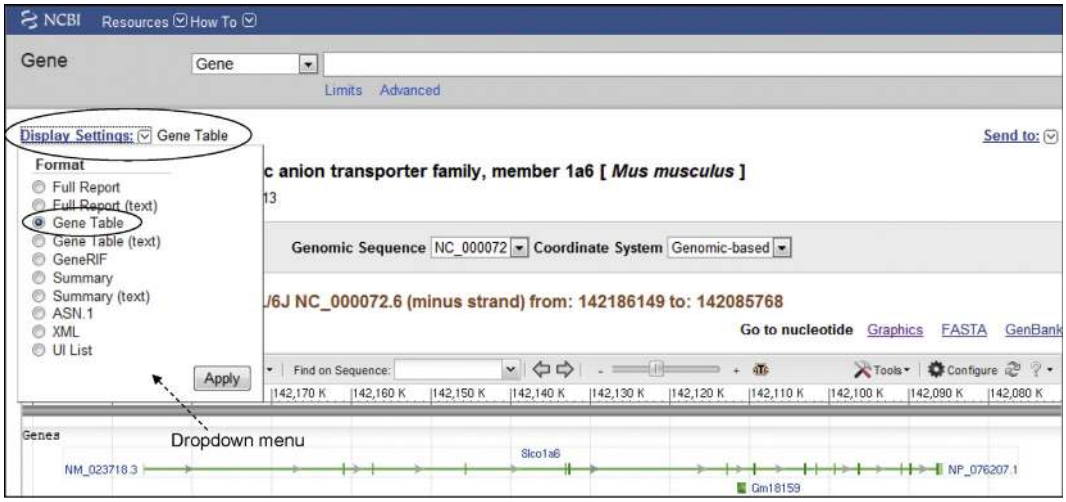


FIGURE 5.20 **Exon and intron sequence information for mouse Slco1a6.** Partial screenshot (upper part) of the details of the exon and intron sequence information that can be obtained by clicking the "Display Setting" in the left-hand top corner and selecting the "Gene Table" from the drop-down menu (circled). *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

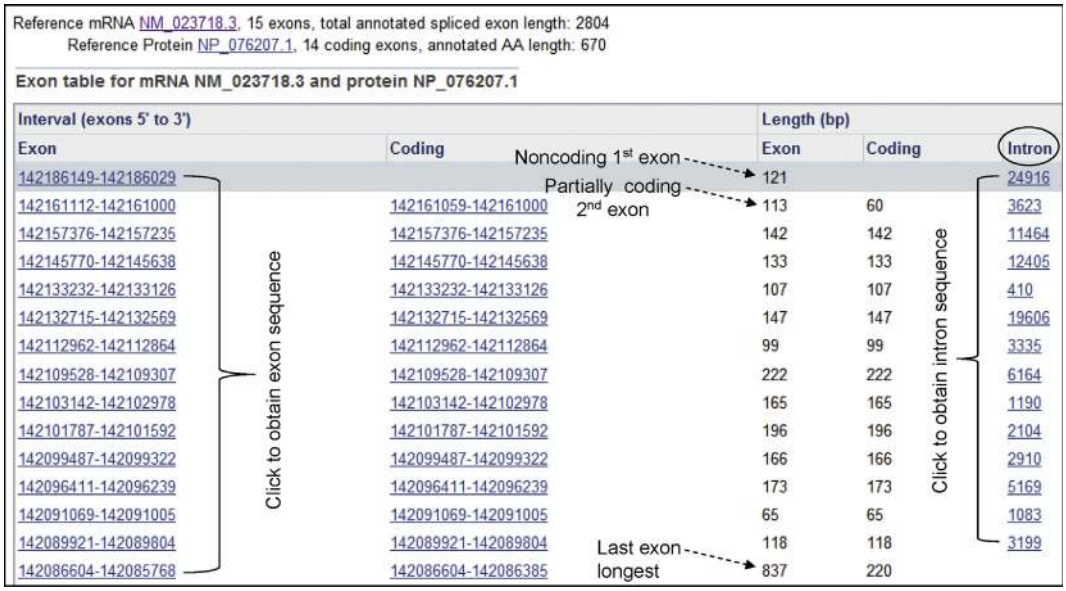


FIGURE 5.21 **Partial screenshot (lower part) of the details of the exon and intron sequence information (continuation of Figure 5.20).** Each exon or intron link can be clicked to obtain the exon or intron sequence, respectively.

figure is a partial screenshot showing the upper part of the display). The lower part of the display shows the details of the exon and intron sequence information (Figure 5.21). Each exon or intron link can be clicked to obtain the exon or intron sequence, respectively.

Below the "Genomic regions, transcripts, and products" field there is the "**Bibliography**" field (Figure 5.14). If this field is expanded by clicking, it shows a field called "**GeneRIFs: Gene References Into Functions**." The GeneRIF contains a link called "**Correction**," which provides an opportunity to the scientific community to update and add more relevant references in relation to the gene in question. This information can be submitted to the NCBI directly.

5.9 DATA VISUALIZATION IN GENOME BROWSERS

A genome browser[o] is a graphical interface for users to retrieve, browse, and analyze the sequence data of both known and predicted genes. Genome browsers stack annotation tracks underneath the genome coordinate positions. This allows graphic display of different types of information, such as gene density in a chromosome, distance between specific genes along the chromosome (which might shed some light on their possible coordinate regulation), map position of genes in specific cytogenetic bands, map position of a disease-related gene in a gene neighborhood, visualization of gene prediction, proteins, expression, variation, comparative analysis, etc. Therefore, annotated data are usually derived from multiple sources, including genomic databases. Each genome browser provides its own annotation of the assembled sequence independently. Information from many other databases can be overlaid on the annotated sequence in the display window. Genome assembly and annotation is a continuous and ongoing process. Therefore, when comparing the data output from different browsers, one should make sure that the comparison is being made based on the same genome-assembly version. On the browser "Gateway" page, the user selects the genome, gene name, etc. to initiate a search.

In addition to data visualization, genome browsers also aid in data retrieval and analysis, and data customization. As discussed above, genome browsers integrate various annotation data into a graphical view. Most of the existing genome browsers support search functions to locate genomic regions by coordinates, sequences, or keywords. Genome browsers also provide a customization platform for end-users to upload, create, and share their own annotation data.

In order to meet the challenge of handling and displaying genomic data, three genome browsers were initially created, soon after the working draft of the human genome was finished: the **NCBI Map Viewer**, the **Ensembl genome browser**, and the University of California Santa Cruz (**UCSC**) **Genome browser**. Subsequently, many other genome browsers have also been developed, some of which can be downloaded. One of these is the **VEGA genome browser**, which has been built on the Ensembl database. These four web-based genome browsers will be discussed here.

5.9.1 Ensembl Genome Browser

Ensembl[59] (www.ensembl.org/) is a collaborative project between the EMBL-EBI and the Sanger Center in the UK. It was started in 1999 with the goal to develop an annotation software system that could provide automated annotation of the human genome, and making the data available to scientists through the web. The development of the Ensembl browser is the result of this collaboration. With the sequencing of the genomes of so many other species, the scope of Ensembl has grown significantly; it now includes data on comparative genomics and regulation as well.

The figures based on the Ensembl browser are created using release 72 (Ensembl 72: June 2013, permanent link: http://Jun2013.archive.ensembl.org/index.html). Ensembl currently maintains all archives for at least two years. By the time this book is published, the release number will certainly have changed, and some details of the visual display features will have changed as well, although the overall display will likely remain similar. Therefore, the reader should still be able to use the browser function. Additionally, the reader can click "View in archive site" at the left-hand bottom corner of the Ensembl home page or use the permanent link cited above to access release 72 for comparison.

Figure 5.22 is a partial screenshot of the Ensembl home page. Entering the search term "Oatp-5" in the mouse database returns the results page shown in Figure 5.23. The upper panel of Figure 5.23 shows the number of records retrieved. If the "Gene" or "Transcript" link is clicked, a new window appears, shown in the lower panel of Figure 5.23. The lower panel shows that two important links in this page are "Gene ID" and "Location" (circled). Clicking "Gene ID" retrieves the gene information page shown in Figure 5.24 (upper panel). It shows the link to the gene (Location), the transcript (with all the known variants), and the protein. There is also a link to the consensus coding sequence (CCDS) database. The gene information page also contains a gene summary and displays (Figure 5.24, middle panel; partial view). Clicking the "Transcript ID" link of Slco1a6-001 returns the Transcript summary and display (Figure 5.24, lower panel). Clicking the link on the gene "Location" field retrieves the details of the gene in a new window. Figure 5.25 upper panel shows the location of *Slco1a6* on chromosome 6 (circled) and the detail of the region showing the surrounding loci of *Slco1a6*. Ensembl identifies the chromosomal location as 6G2 (not 6qG2). By

[o]The display of information output in any genome browser is subject to change. This is because there is continuing effort to improve browser function, versatility, and display features. In addition, genomic databases are continuously updated. Therefore, the graphic displays shown in the figures are not expected to remain the same over time. Nevertheless, knowing how to use the genome browser should prepare the reader to deal with any such changes. The information discussed in this section and shown in the various figures was obtained by accessing the Ensembl, UCSC, Map Viewer, and VEGA genome browsers in June 2013.

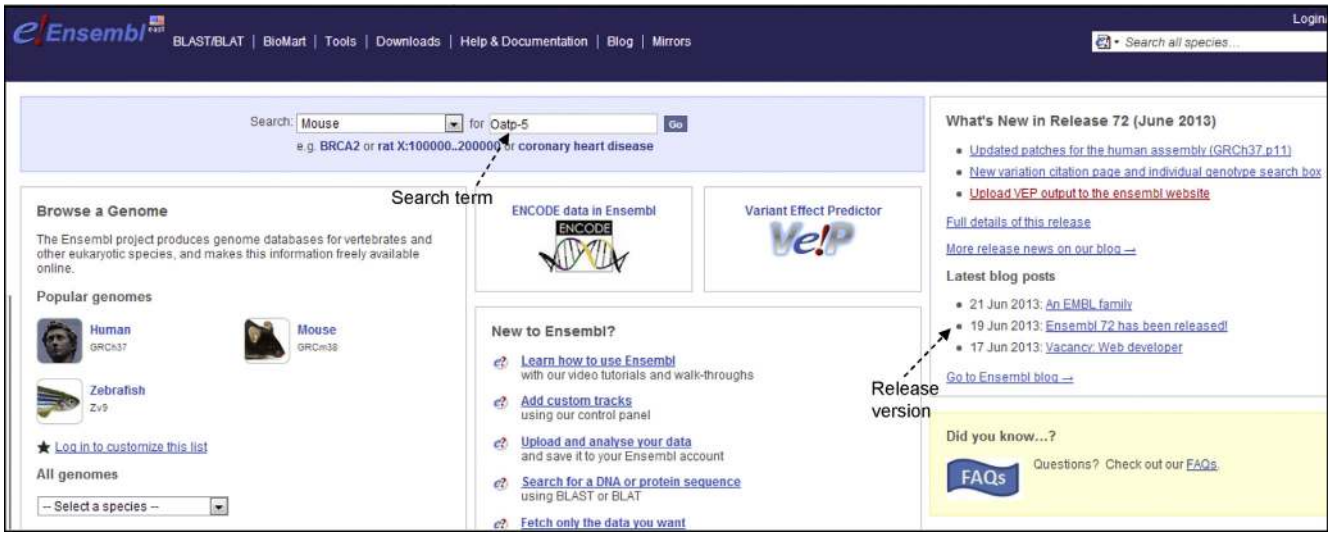


FIGURE 5.22 **Partial screenshot of the Ensembl home page.** Entering the search term "Oatp-5" in the mouse database returns the results page shown in Figure 5.23 upper panel. *(Source: www.ensembl.org/, Ensembl release 72—January 2013 with permanent link http://jan2013.archive.ensembl.org/index.html; information as of June 2013)*

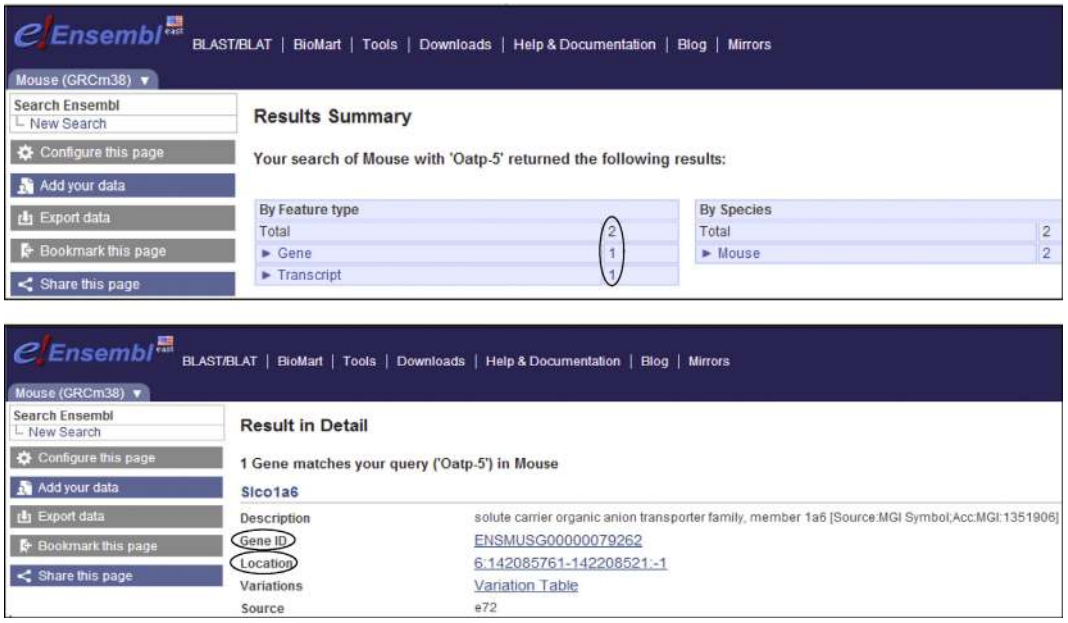


FIGURE 5.23 **Results of searching Ensembl for Oatp-5.** The upper panel shows the number of records retrieved by typing Oatp-5 as the search term. If the "Gene" or "Transcript" link is clicked, a new window appears (lower panel). *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

clicking Slco1a6, a drop-down box appears that contains more information. Figure 5.25 lower panel shows all four transcripts (splice variants) identified for *Slco1a6* as well as the CCDS annotated transcript. Similar drop-down boxes appear if the transcripts are clicked (not shown in the figure).

The user can play with various links to obtain more information and display about the gene, transcript, and protein. For example, the protein display is not shown here at all. Clicking the "Protein ID" link of Slco1a6-001 (Figure 5.24) displays the protein information, including the relative location of all the transmembrane helices.

Clicking the "**consensus coding sequence (CCDS)**" link of Slco1a6-001 (Figure 5.24) takes the user to the CCDS database home page (not shown). The CCDS project is a collaboration involving the EBI, NCBI,

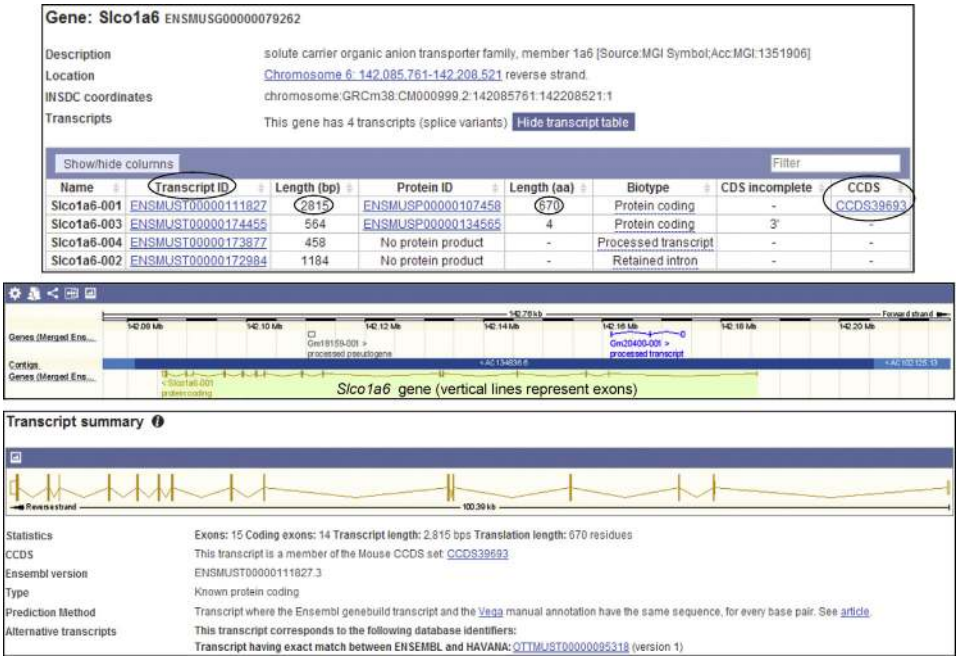


FIGURE 5.24 **Ensembl gene information page for Oatp-5.** Clicking "Gene ID" (Figure 5.23, lower panel) retrieves the gene information page (upper panel) with links to the gene location, the transcript (with all the known variants), and the protein, as well as the CCDS database. The gene information page displays the gene summary (middle panel; partial view). Clicking the "Transcript ID" link of Slco1a6-001 returns the transcript summary and display (lower panel). *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

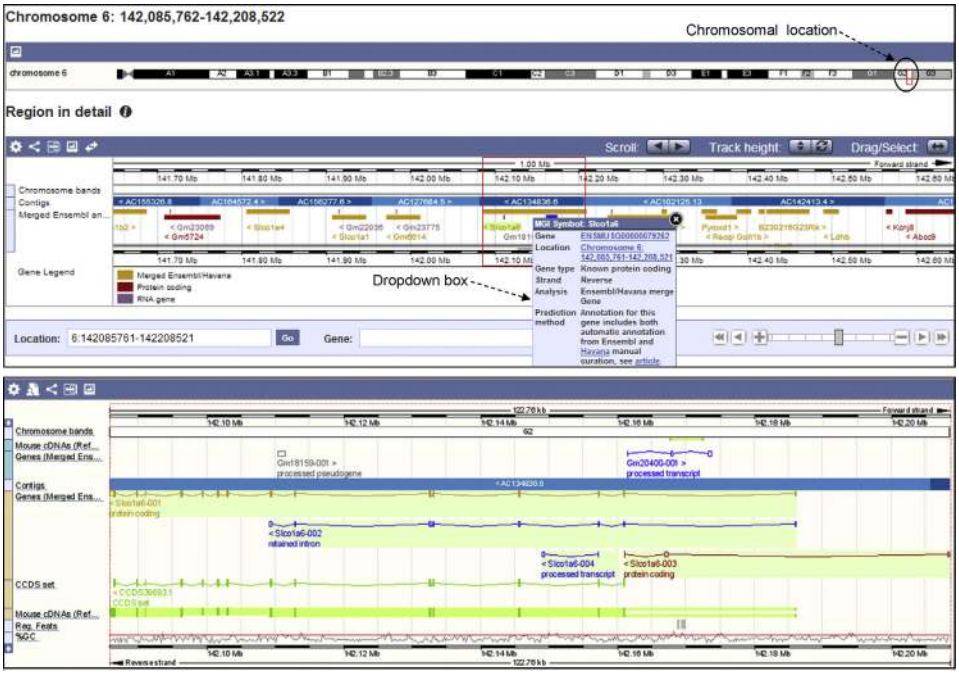


FIGURE 5.25 **Details of the gene information in Ensemble.** Clicking the link on the gene "Location" field (Figure 5.23, lower panel) retrieves the details of the gene. The upper panel shows the location of *Slco1a6* on chromosome 6 (circled) and the detail of the region showing the surrounding loci of *Slco1a6*. The lower panel shows all four transcripts (splice variants) identified for *Slco1a6* as well as the CCDS annotated transcript. *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

UCSC, and the Wellcome Trust Sanger Institute[60] (WTSI). The collaboration was developed in order to identify a core set of protein-coding regions that are consistently annotated on the reference mouse and human genomes. Mouse and human genomes were chosen because these genome sequences are now sufficiently stable. The long-term goal is to support convergence towards a standard set of gene annotations. CCDS assigns a CCDS ID to the annotated protein and these annotated proteins are represented on the NCBI Map Viewer, Ensembl, and UCSC genome browsers by links to the CCDS database. The CCDS ID of mouse Slco1a6 protein is 39693, version 1 (39693.1). The information in current CCDS (as of June 2013) is also based on mouse genome build 38. The CCDS has links to the NCBI, UCSC, Ensembl, and VEGA genome browsers, as well as a link to the NCBI database.

After a search is initiated in the Ensembl browser, a number of links appear in the left panel; of these, the "Add your data" link can be used to upload new data. Alternatively, on the Ensembl home page there are links to "add custom tracks" and "upload and analyze your data," as well as a link to Ensemble tutorials. These can be used to learn data retrieval, analysis, and customization, such as how to add or remove annotation tracks, and to upload and analyze users' own data. The Ensembl browser has detailed tutorials on these topics.

5.9.2 UCSC Genome Browser

The UCSC genome browser[61–63] (http://genome.ucsc.edu/) has been developed and maintained by the Genome Bioinformatics Group at the University of California at Santa Cruz (UCSC). It is a very widely used genome browser. It contains the reference sequence and working draft assemblies for a large collection of genomes. The browser zooms and scrolls over chromosomes showing annotation. Figure 5.26 shows a screenshot of the UCSC genome browser home page. The "Cite Us" link on the left panel lists all the publications associated with the development and updating of the UCSC genome browser (Figure 5.26; link circled). Clicking the "Genomes" or "Genome Browser" links (circled) takes the user to the "(Species) Genome Browser Gateway," from where the search can be launched. Figure 5.27 shows the Mouse Genome Browser Gateway. The gateway provides options for selecting the (organism) group, the species whose genome will be searched (the genome-assembly version is automatically selected as the latest one available), and the search term.

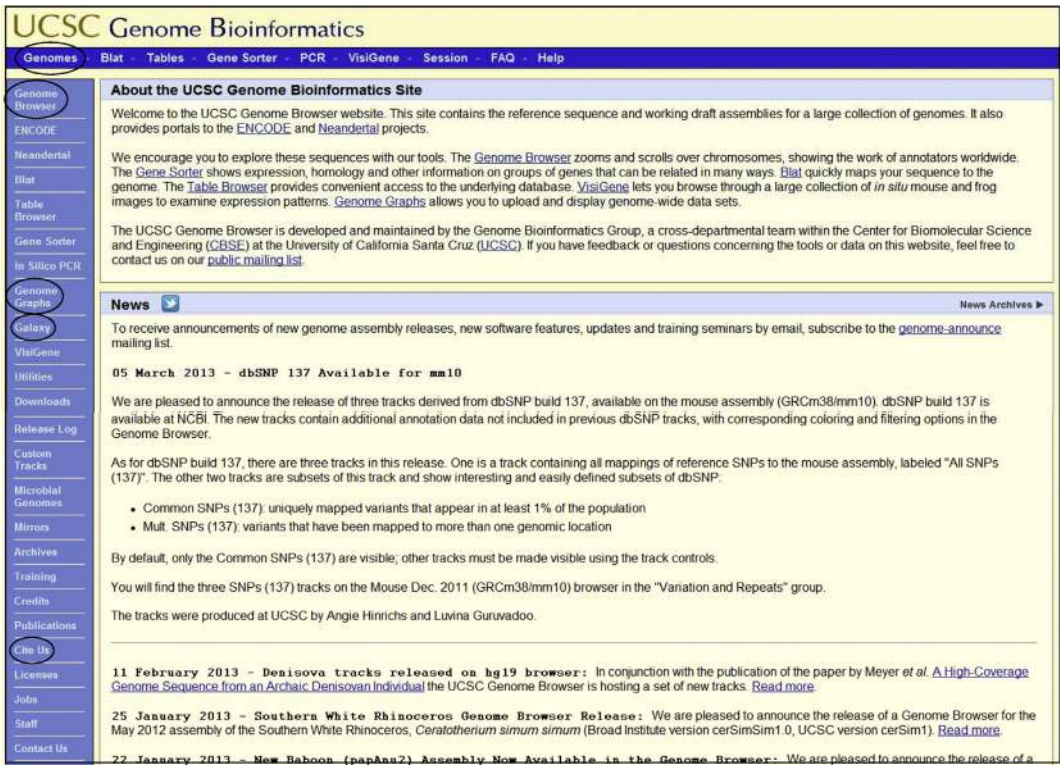


FIGURE 5.26 **Partial screenshot of the UCSC genome browser home page.** Since March 2013 when this screenshot was captured, Gibbon genome browser has been released (22 May 2013) and also the Ferret genome browser (12 June 2013). The UCSC genome browser home page as of June 2013 contains these update announcements. *(Source: http://genome.ucsc.edu/, information as of March 2013)*

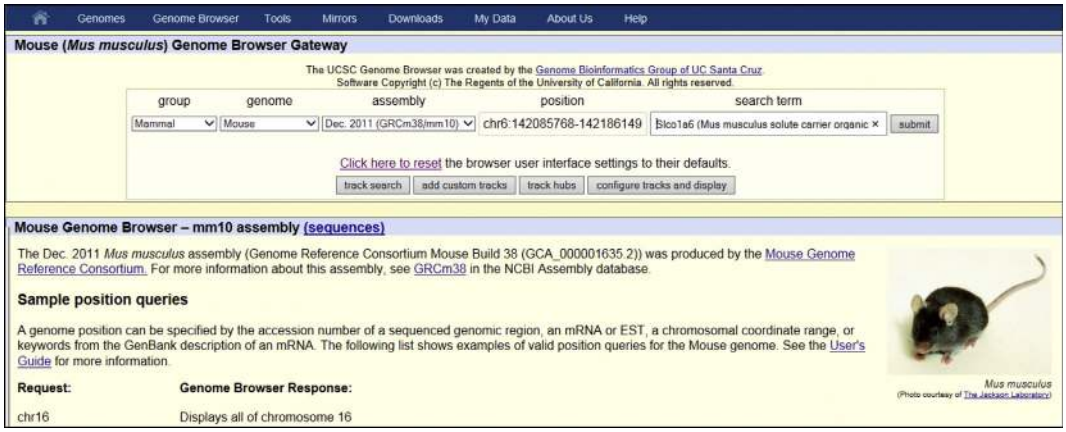


FIGURE 5.27 **The UCSC Mouse Genome Browser Gateway.** The search term used was Slco1a6. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

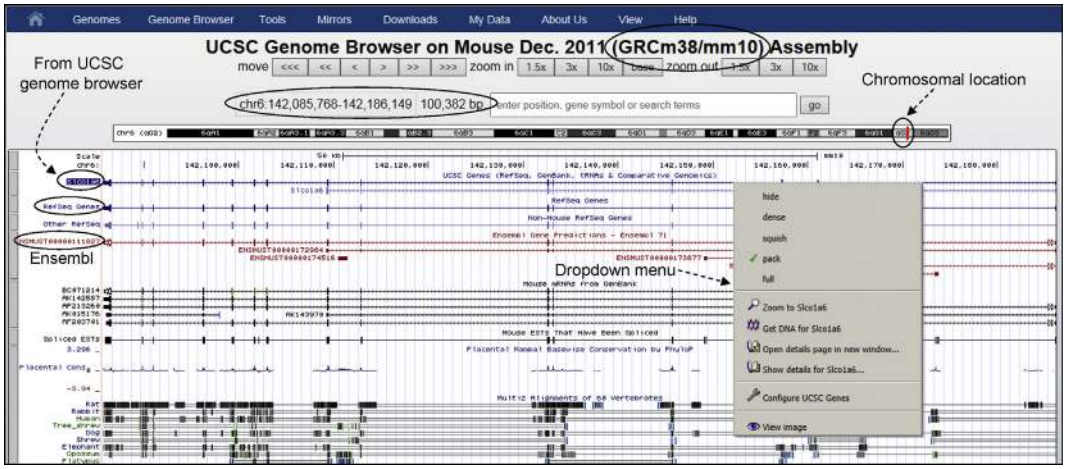


FIGURE 5.28 **UCSC Mouse Genome Browser record for Slco1a6.** Browser display of the Slco1a6 record from different sources (UCSC, RefSeq, Ensembl) represented as separate tracks. Right-clicking on any track produces a drop-down box that offers various options. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

Searching the UCSC genome browser for mouse "Slco1a6" retrieves information from multiple sources (Figure 5.28), such as the UCSC Gene (at the top, highlighted), RefSeq Gene, and Ensembl Gene resources. Right-clicking on any track produces a drop-down box that offers various options. Note that the chromosomal location is described as 6qG2 instead of 6G2. The page also shows the chromosomal location and the length of the gene as "chr6:142,085,768−142,186,149 100,382 bp" (circled). The *Slco1a6* gene organization and information from multiple sources is represented graphically: at the top (highlighted) is the "UCSC Genes" record (because it is the UCSC browser), next is the "RefSeq Genes" record, and the lower red line is the "Ensembl Genes" record. Note that the mouse genome build is noted as GRCm38/mm10. This is because mm10 is the UCSC version of GRCm38.

The UCSC genome browser also provides various other tools to retrieve genome-related data, such as Gene Sorter, BLAT, Table Browser, VisiGene, and Genome Graph. Each of these tools is useful in a unique way. For example, **Gene Sorter** shows the expression, homology, and other information on groups of related genes, **BLAT** (BLAST-like Alignment Tool) maps an input sequence to the genome, and **VisiGene** allows the user to browse through in situ images to examine the expression patterns. **Genome Graph** allows a user to upload and display genome-wide data sets. UCSC **Table Browser**[64] provides text-based access to a large collection of genome assemblies and annotation data stored in the genome browser database. Thus, it provides an alternative to the graphical-based genome browser. For example, Table Browser can be used to retrieve the data associated with a track in text format,

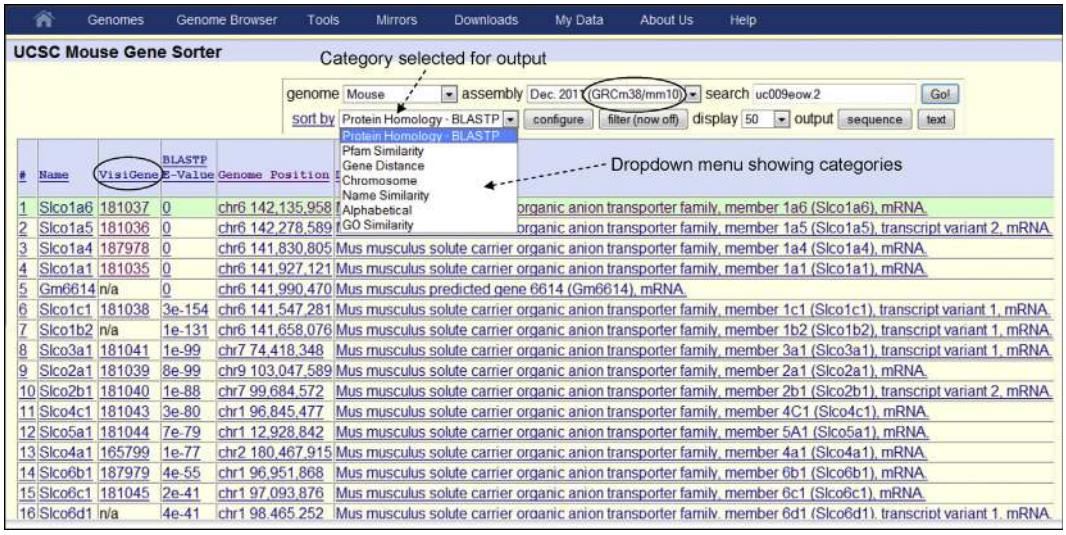


FIGURE 5.29 **Results of a search in Gene Sorter on mouse genome to find the proteins that are related to Slco1a6.** *(Source: http://genome.ucsc.edu/, information as of June 2013)*

to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. The discussion below will focus on Gene Sorter, BLAT, and VisiGene.

The **Gene Sorter**[P] program displays a table of genes that are related to one another. This relationship may be based on expression profiles, protein-level similarities, genomic proximity, etc. The categories by which relatedness is assessed are shown in the drop-down menu next to "sort by" link (Figure 5.29). The figure shows the results of a search in mouse genome to find the proteins that are related to Slco1a6. The search term selected was "Protein Homology – BLASTP," chosen from the drop-down menu. The search retrieved 15 other proteins that bear the closest relationship to Slco1a6 in terms of protein homologous relationship. The "Genome Position" column of the table shows the chromosomal location of these genes. The "VisiGene" column (circled) provides a link to the in situ images of the expression of the respective genes in mouse brain.

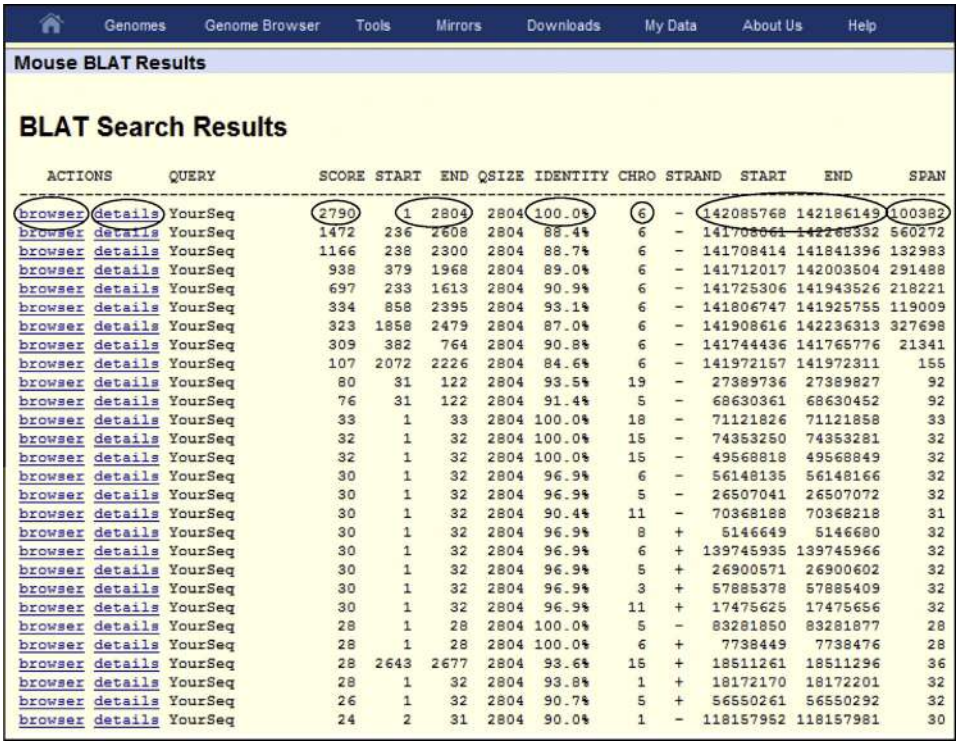
The **BLAT** (BLAST-like Alignment Tool) was written by Jim Kent at UCSC.[65] BLAT is used to map the input sequence to the genome—that is, to identify the location of a sequence in the genome. Therefore, BLAT works with the genomic context in memory, but it works by alignment-based similarity search. BLAT works for both DNA and proteins. For DNA, BLAT is designed to find sequences with ≥95% similarity with the input sequence, where the sequences are ideally 25 bases or more in length. For proteins, BLAT is designed to find sequences with ≥80% similarity with the input sequence, where the sequences are ideally 20 amino acids or more[Q].

BLAT is different from BLAST because, unlike BLAST, BLAT does not search the sequences from GenBank/EMBL-Bank/DDBJ; rather, BLAT uses an index derived from the genome assembly and it consists of all non-overlapping 11-mers except the heavily repeated sequences. For proteins, BLAT uses 4-mers.

Figure 5.30 shows the results of the BLAT analysis of the *Oatp5/Slco1a6* mRNA sequence. Various features of the best match, at the top, are circled. Clicking the "browser" link on the left shows a graphic display of the genomic location of the sequence in the browser. Clicking the "details" link shows the mapping of the input sequence in the mouse genome. Figure 5.31 shows that mouse *Oatp5/Slco1a6* mRNA sequence is derived from 15 exons of the *Oatp5/Slco1a6* gene. These 15 exons are listed on the left as "block 1" through "block 15." Clicking on any "block" link shows the location of the exon in the gene. The analysis also shows that the input sequence belongs to chromosome 6. The exon–intron sequences as well as the flanking sequences are also visible by scrolling up and down the sequence. Figure 5.32 is a composite figure that shows four exons ("blocks") mapped to mouse chromosome 6, showing the exon sequence and surrounding intron sequence, except for exon 1, which is flanked on the left-hand side (upstream) by the 5′-flanking sequence of the gene. The intronic splice

[P]The UCSC Gene Sorter was designed and implemented by Jim Kent, Fan Hsu, Donna Karolchik, David Haussler, and the UCSC Genome Bioinformatics Group (http://genome.ucsc.edu/cgi-bin/hgNear).

[Q]Source: http://genome.ucsc.edu/cgi-bin/hgBlat?command = start.



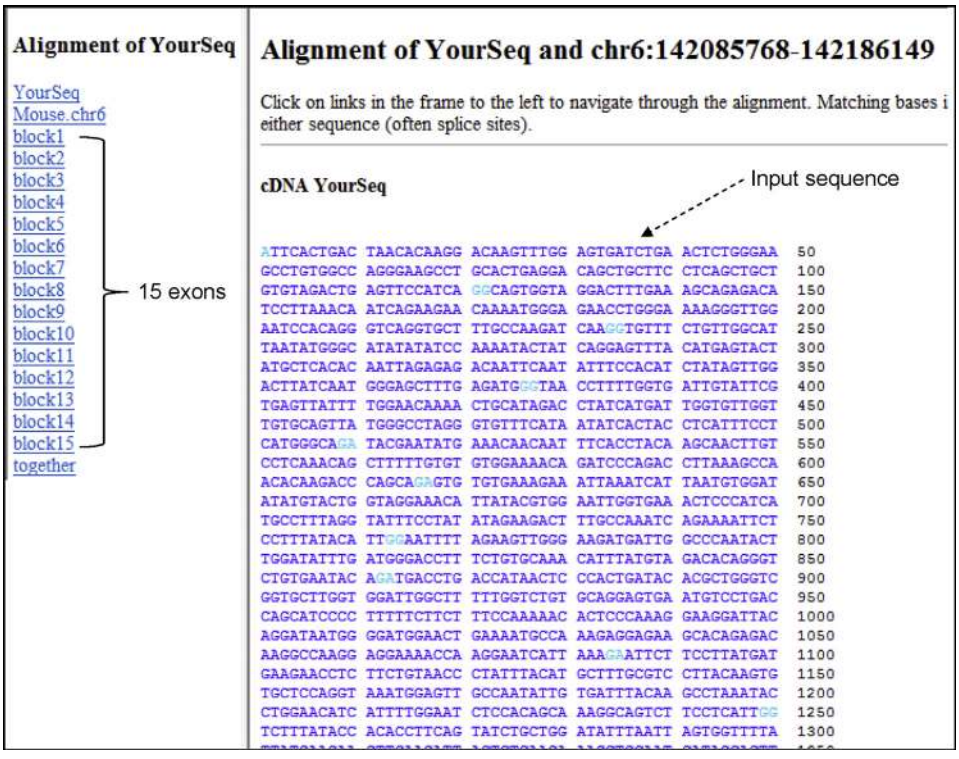
Genomes Genome Browser Tools Mirrors Downloads My Data About Us Help

Mouse BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	2790	1	2804	2804	100.0%	6	-	142085768	142186149	100382
browser details	YourSeq	1472	236	2608	2804	88.4%	6	-	141708061	142268332	560272
browser details	YourSeq	1166	238	2300	2804	88.7%	6	-	141708414	141841396	132983
browser details	YourSeq	938	379	1968	2804	89.0%	6	-	141712017	142003504	291488
browser details	YourSeq	697	233	1613	2804	90.9%	6	-	141725306	141943526	218221
browser details	YourSeq	334	858	2395	2804	93.1%	6	-	141806747	141925755	119009
browser details	YourSeq	323	1858	2479	2804	87.0%	6	-	141908616	142236313	327698
browser details	YourSeq	309	382	764	2804	90.8%	6	-	141744436	141765776	21341
browser details	YourSeq	107	2072	2226	2804	84.6%	6	-	141972157	141972311	155
browser details	YourSeq	80	31	122	2804	93.5%	19	-	27389736	27389827	92
browser details	YourSeq	76	31	122	2804	91.4%	5	-	68630361	68630452	92
browser details	YourSeq	33	1	33	2804	100.0%	18	-	71121826	71121858	33
browser details	YourSeq	32	1	32	2804	100.0%	15	-	74353250	74353281	32
browser details	YourSeq	32	1	32	2804	100.0%	15	-	49568818	49568849	32
browser details	YourSeq	30	1	32	2804	96.9%	6	-	56148135	56148166	32
browser details	YourSeq	30	1	32	2804	96.9%	5	-	26507041	26507072	32
browser details	YourSeq	30	1	32	2804	90.4%	11	-	70368188	70368218	31
browser details	YourSeq	30	1	32	2804	96.9%	8	+	5146649	5146680	32
browser details	YourSeq	30	1	32	2804	96.9%	6	+	139745935	139745966	32
browser details	YourSeq	30	1	32	2804	96.9%	5	+	26900571	26900602	32
browser details	YourSeq	30	1	32	2804	96.9%	3	+	57885378	57885409	32
browser details	YourSeq	30	1	32	2804	96.9%	11	+	17475625	17475656	32
browser details	YourSeq	28	1	28	2804	100.0%	5	-	83281850	83281877	28
browser details	YourSeq	28	1	28	2804	100.0%	6	+	7738449	7738476	28
browser details	YourSeq	28	2643	2677	2804	93.6%	15	+	18511261	18511296	36
browser details	YourSeq	28	1	32	2804	93.8%	1	+	18172170	18172201	32
browser details	YourSeq	26	1	32	2804	90.7%	5	+	56550261	56550292	32
browser details	YourSeq	24	2	31	2804	90.0%	1	-	118157952	118157981	30

FIGURE 5.30 **The results of BLAT analysis of the *Oatp5/Slco1a6* mRNA sequence.** The RefSeq sequence was used for the analysis. Clicking "browser" (circled) opens up the browser page shown in Figure 5.28. Clicking "details" (circled) opens up the record shown in Figure 5.31. *(Source: http://genome.ucsc.edu/, information as of June 2013)*



Alignment of YourSeq

YourSeq
Mouse.chr6
block1
block2
block3
block4
block5
block6
block7
block8
block9
block10
block11
block12
block13
block14
block15
together

15 exons

Alignment of YourSeq and chr6:142085768-142186149

Click on links in the frame to the left to navigate through the alignment. Matching bases i
either sequence (often splice sites).

cDNA YourSeq

Input sequence

```
ATTCACTGAC TAACACAAGG ACAAGTTTGG AGTGATCTGA ACTCTGGGAA   50
GCCTGTGGCC AGGGAAGCCT GCACTGAGGA CAGCTGCTTC CTCAGCTGCT  100
GTGTAGACTG AGTTCCATCA GGCAGTGGTA GGACTTTGAA AGCAGAGACA  150
TCCTTAAACA ATCAGAAGAA CAAAATGGGA GAACCTGGGA AAAGGGTTGG  200
AATCCACAGG GTCAGGTGCT TTGCCAAGAT CAAGGTGTTT CTGTTGGCAT  250
TAATATGGGC ATATATATCC AAAATACTAT CAGGAGTTTA CATGAGTACT  300
ATGCTCACAC AATTAGAGAG ACAATTCAAT ATTTCCACAT CTATAGTTGG  350
ACTTATCAAT GGGAGCTTTG AGATGGGTAA CCTTTTGGTG ATTGTATTCG  400
TGAGTTATTT TGGAACAAAA CTGCATAGAC CTATCATGAT TGGTGTTGGT  450
TGTGCAGTTA TGGGCCTAGG GTGTTTCATA ATATCACTAC CTCATTTCCT  500
CATGGGCAGA TACGAATATG AAACAACAAT TTCACCTACA AGCAACTTGT  550
CCTCAAACAG CTTTTTGTGT GTGGAAAACA GATCCCAGAC CTTAAAGCCA  600
ACACAAGACC CAGCAGAGTG TGTGAAAGAA ATTAAATCAT TAATGTGGAT  650
ATATGTACTG GTAGGAAACA TTATACGTGG AATTGGTGAA ACTCCCATCA  700
TGCCTTTAGG TATTTCCTAT ATAGAAGACT TTGCCAAATC AGAAAATTCT  750
CCTTTATACA TTGGAATTTT AGAAGTTGGG AAGATGATTG GCCCAATACT  800
TGGATATTTG ATGGGACCTT TCTGTGCAAA CATTTATGTA GACACAGGGT  850
CTGTGAATAC AGATGACCTG ACCATAACTC CCACTGATAC ACGCTGGGTC  900
GGTGCTTGGT GGATTGGCTT TTTGGTCTGT GCAGGAGTGA ATGTCCTGAC  950
CAGCATCCCC TTTTTCTTCT TTCCAAAAAC ACTCCCAAAG GAAGGATTAC 1000
AGGATAATGG GGATGGAACT GAAAATGCCA AAGAGGAGAA GCACAGAGAC 1050
AAGGCCAAGG AGGAAAACCA AGGAATCATT AAAGAATTCT TCCTTATGAT 1100
GAAGAACCTC TTCTGTAACC CTATTTACAT GCTTTGCGTC CTTACAAGTG 1150
TGCTCCAGGT AAATGGAGTT GCCAATATTG TGATTTACAA GCCTAAATAC 1200
CTGGAACATC ATTTTGGAAT CTCCACAGCA AAGGCAGTCT TCCTCATTGG 1250
TCTTTATACC ACACCTTCAG TATCTGCTGG ATATTTAATT AGTGGTTTTA 1300
```

FIGURE 5.31 **Mouse *Oatp5/Slco1a6* mRNA sequence is derived from 15 exons ("blocks") of the *Oatp5/Slco1a6* gene.** *(Source: http://genome.ucsc.edu/, information as of June 2013)*

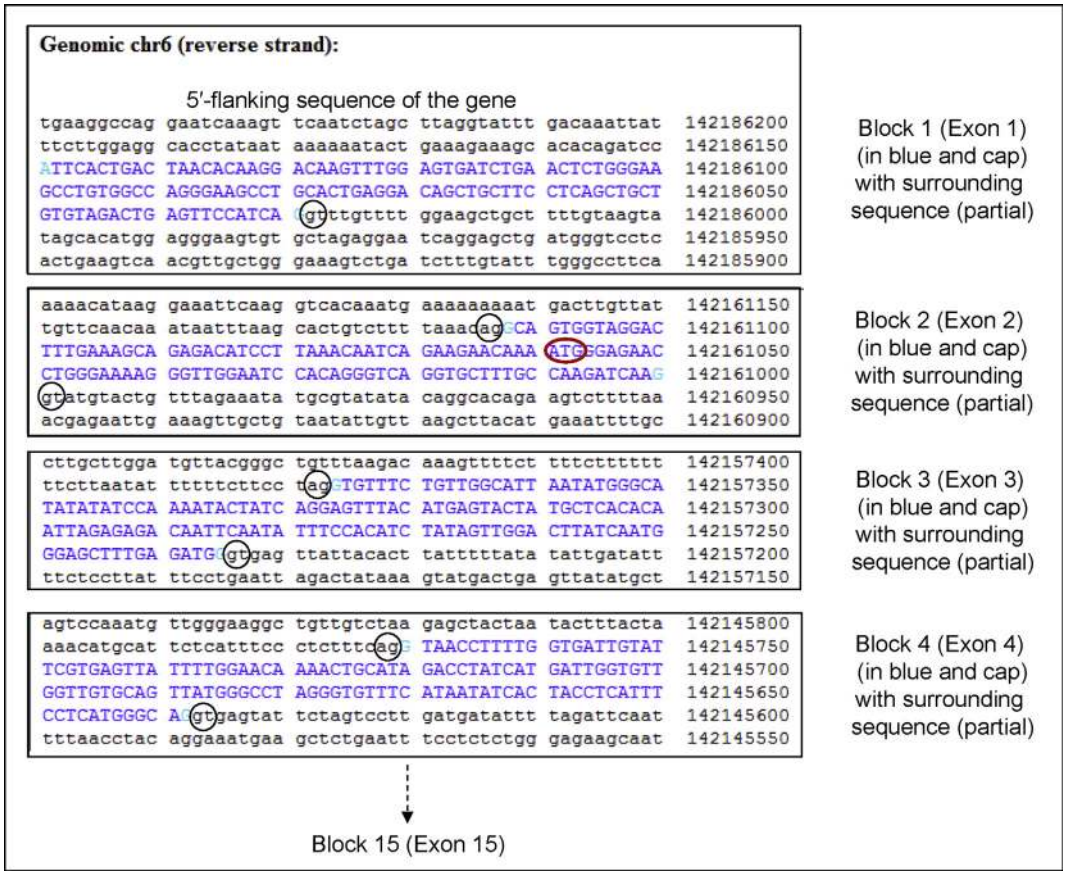


FIGURE 5.32 **A composite figure created to show four exons mapped to mouse chromosome 6.** Each exon sequence is shown in blue capital letters whereas the surrounding intron sequence (and 5′-flanking sequence for exon 1) is shown in black lowercase letters. The intronic splice donor and acceptor sites (gt. . .ag) are circled. The translation initiation codon ATG in exon 2 is also circled.

donor and acceptor sites (gt. . .ag) are circled. The translation initiation codon ATG in exon 2 is also circled. Thus, exon 1 is noncoding whereas exon 2 is partially coding. *Note that Figure 5.32 is not a true screenshot by itself but has been created by copying separate screenshots of BLAT display in order to show how BLAT maps the input sequence to the genome.*

The **VisiGene**[r] Image Browser is like a virtual microscope that provides in situ images. The search term is entered in the search box. Hitting the search button returns available images. Some search terms will return a number of images; others return a few or even only one, whereas still others return none. The source of the images is acknowledged on the image page. Figure 5.33 shows the VisiGene Image Browser page (partial view).

On the left panel of the UCSC genome browser, there is a link to "Genome Graphs," where data can be uploaded or imported into the database (Figure 5.26; link circled). The "Genome Graphs" tool can be used to display genome-wide data sets. The user can upload his/her own data for display by the tool. In order to display personal annotation tracks, the user has to format the data in one of the supported formats and upload the data into the Genome Browser using the "add custom tracks" button on the "Genome Browser Gateway" page (Figure 5.27). The UCSC genome browser has a detailed tutorial on this topic.

5.9.3 NCBI's Map Viewer

The genome browser of the NCBI is called Map Viewer. The current version of Map Viewer displays a chromosome as a vertical line. The direction of a plus strand in a vertical representation is from top to bottom, and that of the reverse or minus (complement) strand is from bottom to top. Map Viewer allows the visualization and search of an organism's complete genome and the chromosome maps, and retrieval of greater levels of detailed information, down to the sequence level, for a region of interest. Figure 5.34 shows the NCBI "Genome" home page with a link to

[r]VisiGene was written by Jim Kent and Galt Barber (http://genome.ucsc.edu/cgi-bin/hgVisiGene?command = start).

FIGURE 5.33 **Partial view of the VisiGene Image Browser page.** The image pages resulting from a search show the in situ image and acknowledge the source of the images. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

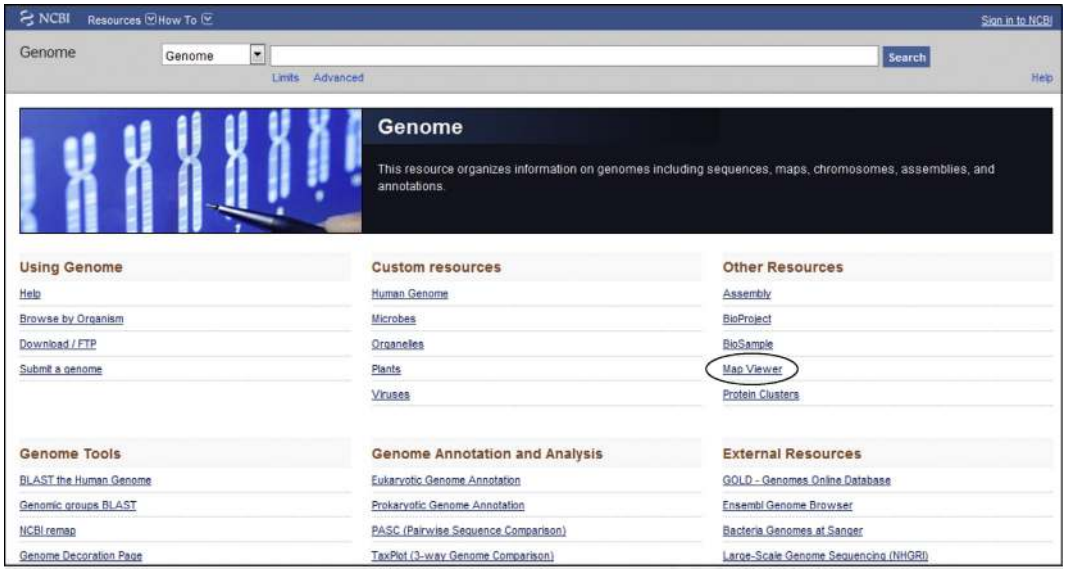


FIGURE 5.34 **NCBI "Genome" home page with a link to Map Viewer.** *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

Map Viewer (circled). Clicking the "Map Viewer" link opens the Map Viewer home page (Figure 5.35). The Map Viewer home page can be directly accessed at http://www.ncbi.nlm.nih.gov/mapview/.

The data display in genome browsers is subject to change and by the time this book is published, many of the figures presented here may not exactly match but will be helpful nonetheless.

A search with *Mus musculus* and Oatp-5 on the Map Viewer home page takes the user to the *Mus musculus* genome view, represented as 19 autosomes plus one X and one Y chromosome (Figure 5.36). The location of the gene (*Oatp5/Slco1a6*) is shown on chromosome 6 by a red mark. Below chromosome 6 there is "2" in red, indicating that the search term Oatp-5 retrieved 2 records shown below: one from the mouse reference genome and one from the Celera mouse genome assembly. If, instead, the search is performed using the search term Slco1a6, 102 records are retrieved (as of June 2013; not shown). Clicking chromosome 6 or

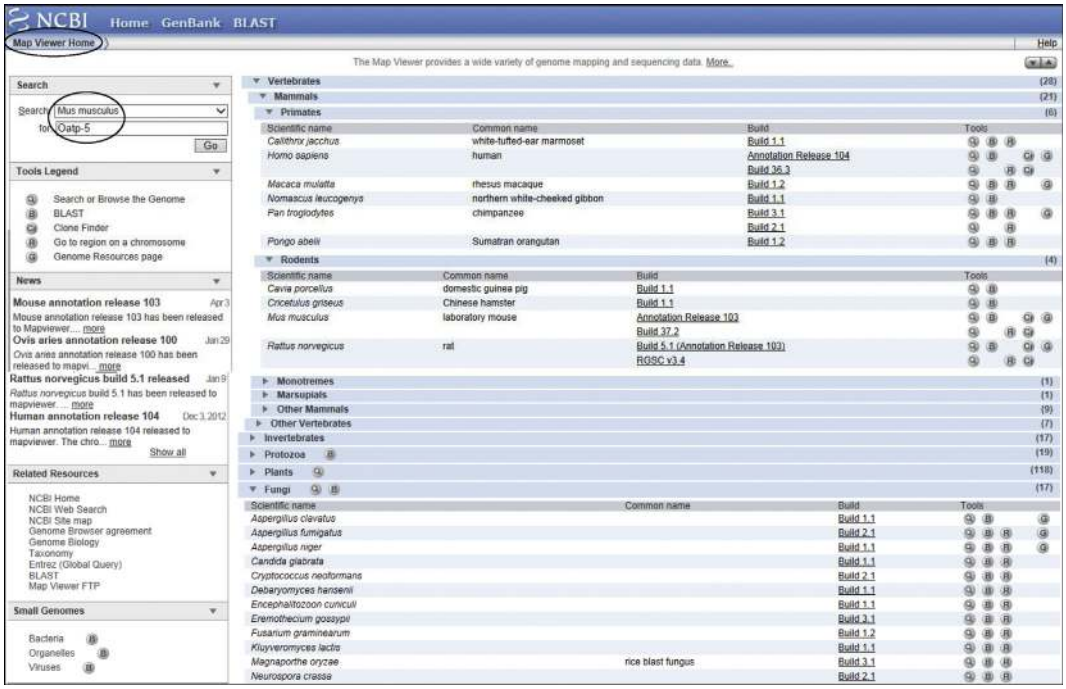


FIGURE 5.35 **Map Viewer home page.** *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June, 2013)*

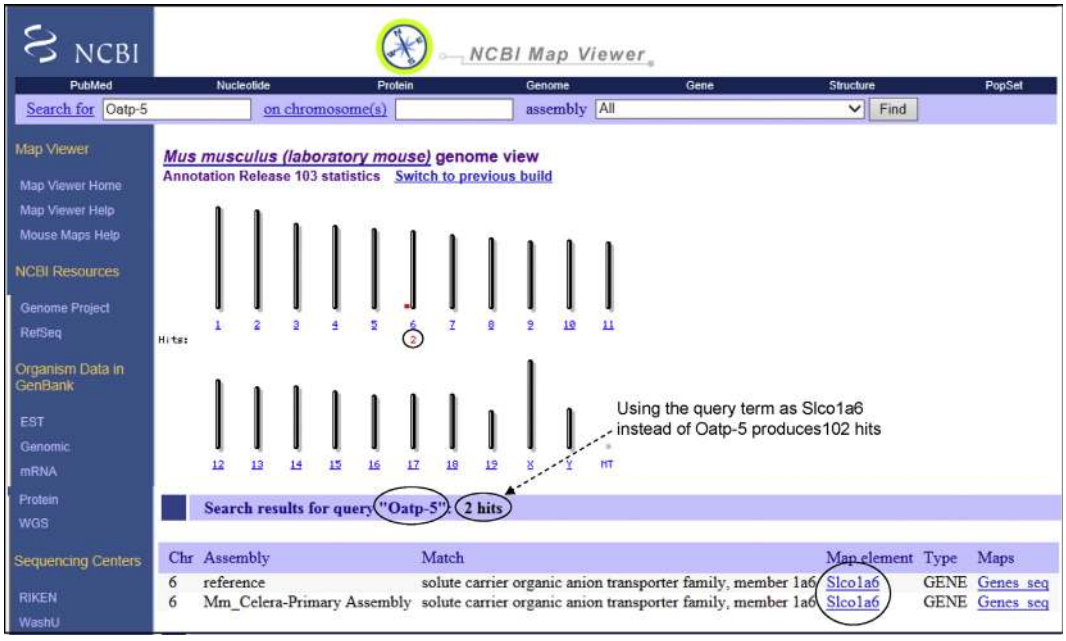


FIGURE 5.36 ***Mus musculus* genome view in Map Viewer.** The location of the gene (*Oatp5/Slco1a6*) on chromosome 6 is indicated by a red mark. Below chromosome 6 there is "2" in red, indicating that the search term Oatp-5 retrieved 2 records. In contrast, if the search term is Slco1a6, 102 records are retrieved. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

Slco1a6 under "Map element" retrieves the information shown in Figure 5.37. In order to zoom the view in or out, the line representing the gene can be clicked; a new window appears that provides zoom-in and zoom-out options (Figure 5.37). The view can be zoomed in to view more detail of the *Slco1a6* gene, or zoomed out to view more genes on chromosome 6. Some of these genes are on the plus strand (indicated by a downward arrow in the Orientation ("O") column) whereas others are on the minus strand

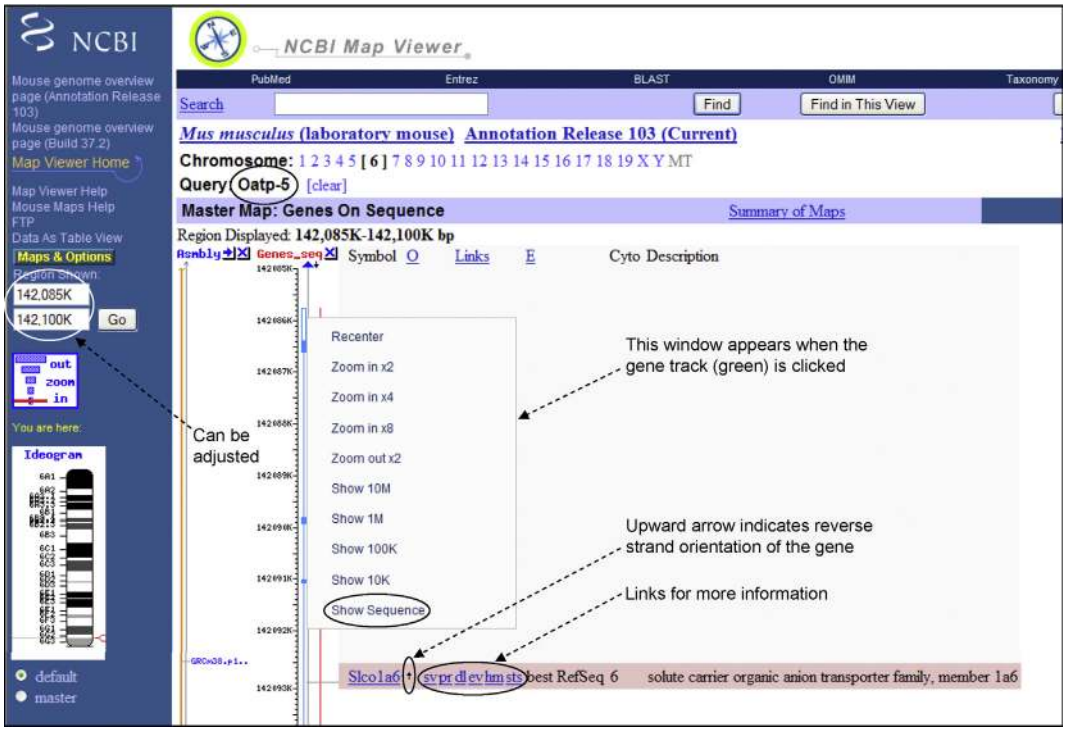


FIGURE 5.37 **Master Map of Oatp-5 in Map Viewer.** Clicking chromosome 6 or Slco1a6 under "Map element" on the page shown in Figure 5.36 retrieves the information shown in this figure. In order to zoom the view in or out, the line representing the gene can be clicked; a new window appears that provides zoom-in and zoom-out options. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

(indicated by upward arrow). *Slco1a6* is on the minus strand. The Map Viewer data is also based on mouse genome-assembly build 38 (Annotation Release 103). In Figures 5.37 and 5.39 there is a link to the previous build (Build 37.2) that can be seen on the left panel. There are a number of links next to the *Slco1a6* gene: **sv** (sequence viewer), **pr** (protein), **dl** (display and download), **ev** (evidence viewer), **hm** (HomoloGene), and **sts** (sequence tagged sites). Clicking each of these links takes the user to a different screen showing specific attributes that can be further explored. For example, clicking "Slco1a6" takes the user to the gene page discussed above. Likewise, clicking "ev" takes the user to the "evidence viewer" page. The evidence viewer is discussed below. The user should play with each of these links to further explore the information available. Therefore, the gene, the mRNA, and the protein sequence information and their various attributes can be retrieved in multiple ways from these links.

5.9.4 VEGA Genome Browser

The **VEGA**[66] (Vertebrate Genome Annotation) genome browser was built on the Ensembl database. The difference between Ensembl and VEGA is that Ensembl displays computationally curated sequences for a large number of vertebrate and invertebrate species, whereas the VEGA database houses high-quality manual annotation of finished vertebrate genomic sequences[s]. The **HAVANA** (Human and Vertebrate Analysis and Annotation) group of the Wellcome Trust Sanger Institute in the UK provides the manual annotation of human, mouse, zebrafish, and other vertebrate genomes that appears in the VEGA browser. Because VEGA is built on Ensembl, the display of information in VEGA is very similar to that in Ensembl. Therefore, only the VEGA home page (http://vega.sanger.ac.uk/index.html) is shown here. At the right-hand side of the home page is a link to the gateway from where a search can be launched (Figure 5.38).

5.10 USING MAP VIEWER TO SEARCH THE GENOME

In the above examples, it was demonstrated how to search and track a specific gene on a chromosome map and retrieve information in specific databases, using

[s]Source: http://www.sanger.ac.uk/resources/databases/vega/.

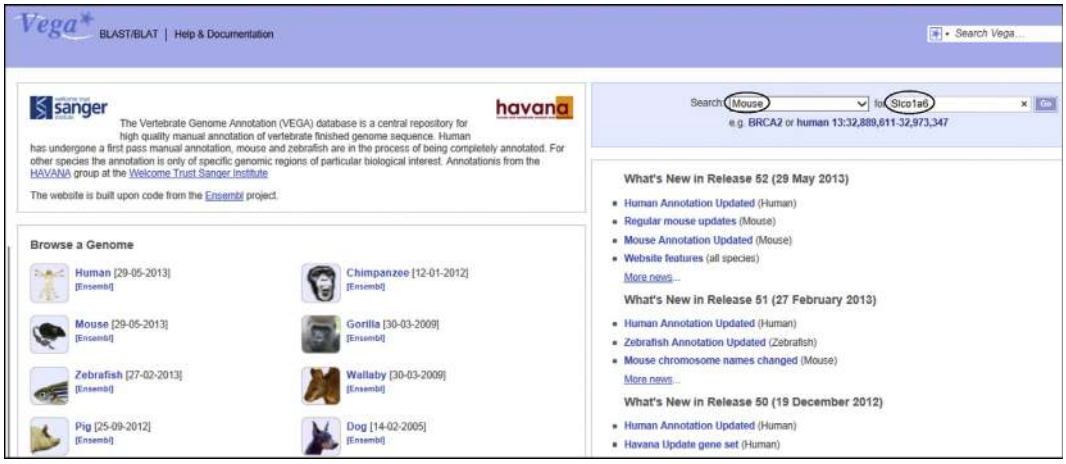


FIGURE 5.38 **VEGA genome browser home page.** *(Source: http://vega.sanger.ac.uk/index.html; as of June 2013)*

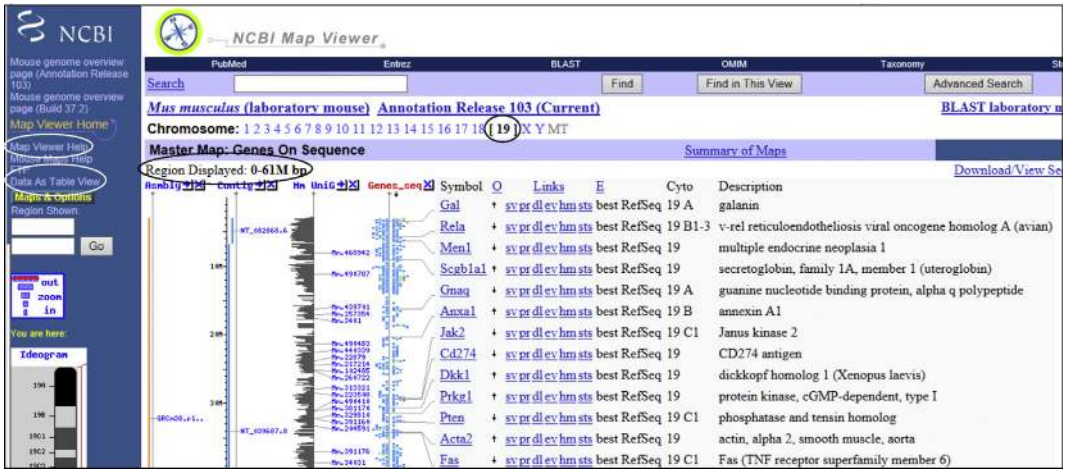


FIGURE 5.39 **Gene distribution in mouse chromosome 19 from Map Viewer.** The list was obtained by selecting "Data As Table View" from the left column. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

the mouse *Oatp5/Slco1a6* gene. However, if one wants to track all the genes identified in a chromosome, one can also do that by using Map Viewer. Entering just *Mus musculus* as the search term on the Map Viewer home page retrieves the mouse genome view in the form of all mouse chromosomes. A particular chromosome can be clicked to open another view with all the genes mapped to that chromosome.

Figure 5.39 shows a partial view of the gene distribution in chromosome 19. Chromosome 19 was chosen because of its small size. The region displayed is 0–61 Mbp. One can select the "Data as Table View" link (circled) from the column on the left to obtain the list of genes in the form of a table. In the same column, there is a link to "Map Viewer Help," which can be clicked to gather some more fundamental information about Map Viewer. For example, the help link explains that there are four levels of details displayed per genome in Map Viewer. Briefly, the **Home Page** for an organism summarizes the resources available for that organism. The **Genome View** provides graphical displays of the complete genome represented in the form of chromosomes. **Map View** displays maps for a selected chromosome and allows one to view regions of interest at different levels of resolution. **Sequence View** displays the sequence data for a specific chromosomal region. In addition, the reader is urged to consult Chapters 20 and 24 of *The NCBI Handbook* (2002, Edited by Jo McEntyre and Jim Ostell; http://www.ncbi.nlm.nih.gov/books/NBK21101/) in order to develop expertise on how to navigate through information in Map Viewer.

Some other uses of Map Viewer links are discussed below. Figure 5.40 shows a partial screenshot of two

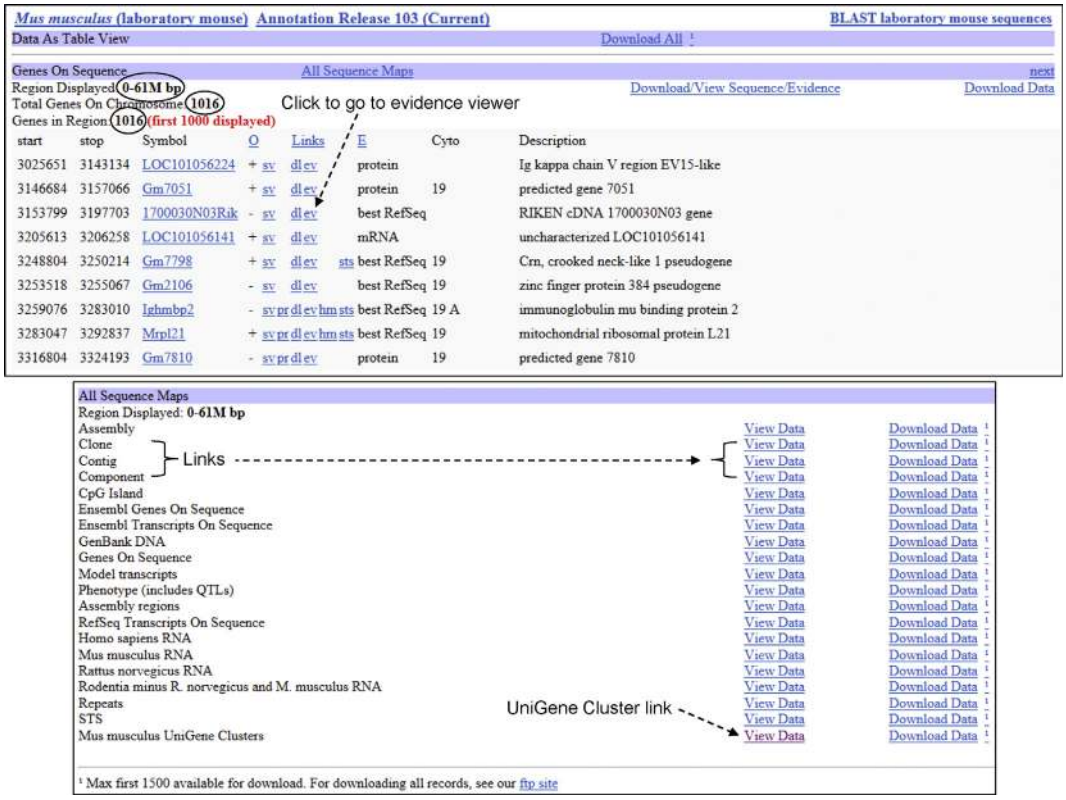


FIGURE 5.40 **Data as Table View.** Clicking the "Data As Table View" link shown in Figure 5.39 retrieves the list of genes in chromosome 19 in the form of a table. The upper and the lower panels are partial screenshots of two fields integrated into one view. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

categories of information integrated into one view; the upper panel shows a partial list of genes in chromosome 19, which contains a total of 1016 genes as of the latest annotation release. In the lower panel is the detail of various attributes of the "Sequence Map" with the option for viewing the relevant data. Clicking the "ev" (Evidence Viewer) link associated with a gene (Figure 5.40, upper panel) opens up the Evidence Viewer screen that shows the evidence for a particular gene model (Figure 5.41A). The gene model is generated based on alignment of mRNA sequences to the human genomic assembly. Thus, the Evidence Viewer displays graphically the cDNAs that align to the genome in a particular region. Mismatches or insertions/deletions are marked. These alignments provide clues to the intron/exon organization of a gene, as annotated on the contigs. Figure 5.41A is a partial screenshot showing only the upper part of the Evidence Viewer display; scrolling down the screen reveals the alignments. A quick discussion on the utility and use of the Evidence Viewer is available at http://www.ncbi.nlm.nih.gov/Web/Newsltr/Fall01/evidence.html.

A few other links labeled in Figure 5.40 are expanded in Figure 5.41 (see legends for Figure 5.41A through 5.41C). In Figure 5.41D, showing the tiling path used to build each genomic contig (the tiling path is the minimum set of closes that encompasses the whole sequence of the contig), there is link to each clone that shows the orientation (+ or −) of the sequence of the clone, the total number of "Bases," and the "Status" (Figure 5.41D). In the "Status" column, "finished HTG" means finished high-throughput genomic sequence[t].

UniGene is not a database of genes; rather, it provides an overview of transcriptomes associated with transcribed loci. Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location. In most organisms, the number of transcribed sequences is usually much larger than the number of genes. This may be

[t]The initial high-throughput genomic (HTG) sequencing data could be single-pass sequencing data with gaps. These initial data are "unfinished" HTG data. Usable data are defined as all sequences existing in contigs of > 2 kb. The unfinished HTG sequence data are eventually converted to the "finished" state (complete contiguity with an error rate of 10^{-4} or less) (see[67]).

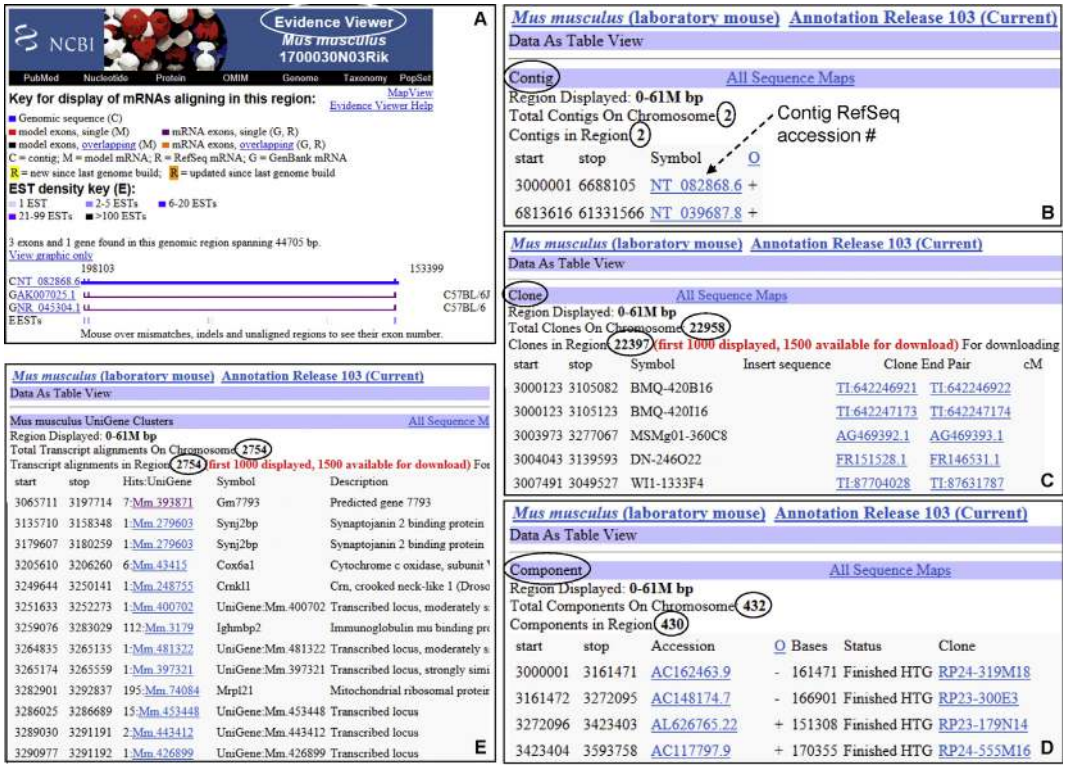


FIGURE 5.41 **Screenshots of individual links (expanded) from Figure 5.40, in June 2013.** (A) Clicking the "ev" link shown in Figure 5.40 retrieves the "Evidence Viewer" screen that shows the evidence for a particular gene model. The NCBI generates gene models based primarily on alignment of mRNA sequences that provide the intron/exon organization of a gene, as annotated on the contigs. (B) Clicking the "Contig" link shown in Figure 5.40 reveals the constructed genomic contig information. There are two constructed genomic contigs covering the sequence of chromosome 19 that spans 0–61 Mbp. Each RefSeq contig accession number can be clicked to obtain further information about the contig, including the sequence. By default, the NT_xxxxxx contigs are shown to reflect the current reference assembly. (C) Clicking the "Clone" link shown in Figure 5.40 reveals that a total of 22,958 clones contain various parts of chromosome 19 sequence, and for the 0–61-Mbp region of chromosome 19, this number is 22,397. The sequence can be obtained by clicking each associated link. (D) The "Component" link in Figure 5.40 provides the tiling path used to build each genomic contig. The tiling path is the minimum set of clones that encompasses the whole sequence of the genomic contig with minimum overlaps (discussed in Chapter 7). The tiling path of chromosome 19 comprises 432 component clones, whereas the tiling path of the 0–61-Mbp region comprises 430 component clones. The details of each clone can be obtained by clicking the associated accession numbers. (E). Clicking the "UniGene Cluster" link shown in Figure 5.40 reveals the transcript information relevant to the region in question. The figure shows a small partial list of transcripts from the UniGene Cluster. Each entry link can be clicked to obtain further information. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

due to multiple reports on the same full-length mRNA (as cDNA), often reported in the database under different names; alternatively spliced variants; multiple partial sequences reported; EST; etc. The existence of many such reported sequences associated with one transcribed locus makes the putative gene assignment a challenging task. This is done computationally as a cluster of transcripts associated with a transcribed locus (hence UniGene Clusters).

In the examples discussed above, only a tiny fraction of the available information has been explored. The user should click the different links, explore, and learn how to harness the wealth of information that is available in and can be accessed through the various genome browsers and databases.

5.11 A NOTE ON THE STATE OF THE SEQUENCE-ASSEMBLY DATA IN DIFFERENT DATABASES

At a given point in time, some inconsistencies may be identified with regard to the genomic data in different databases, or different links within the main database. This is usually owing to the fact that different databases may be updated at different times. The database maintenance team may have limited resources and multiple projects to handle; consequently, a priority is set for handling different projects. Therefore, it is important for the user to take note of the genome-assembly version (build) as well as annotation version when using a genomic database or any link within the database.

References

1. Greenbaum D, et al. *Genome Res* 2001;**11**:1463–8.
2. National Institutes of Health. *GenBank celebrates 25 years of service with two-day conference; leading scientists will discuss the DNA database at April 7–8 meeting*. Available online at: <http://www.nih.gov/news/health/apr2008/nlm-03.htm>; 2008.
3. Benson DA, et al. *Nucl Acids Res* 2013;**41**:D36–42 (Database issue).
4. Kulikova T, et al. *Nucl Acids Res* 2007;**35**:D16–20 (Database issue).
5. EMBL. *EMBL history*. Available online at: <http://www.embl.de/aboutus/general_information/history/>; 2013.
6. Cochrane G, et al. *Nucl Acids Res* 2013;**41**:D30–5 (Database issue).
7. EMBL-EBI. *About ENA*. Available online at: <http://www.ebi.ac.uk/ena/about/about>;2013.
8. Ogasawara O, et al. *Nucl Acids Res* 2013;**41**:D25–9 (Database issue).
9. Kodama Y, et al. *Nucl Acids Res* 2012;**40**:D38–42 (Database issue).
10. Kodama Y, et al. *Nucl Acids Res* 2012;**40**:D54–6 (Database issue).
11. Kaminuma E, et al. *Nucl Acids Res* 2011;**39**:D22–7 (Database issue).
12. Nakamura Y, et al. *Nucl Acids Res* 2013;**41**:D21–4 (Database issue).
13. Brunak S, et al. *Science* 2002;**298**:1333 summarized at: <http://www.insdc.org/policy.html>
14. Acland A, et al. (NCBI Resource Coordinators) *Nucl Acids Res* 2013;**41**:D8–20 (Database issue).
15. Kanz C, et al. *Nucl Acids Res* 2005;**39**:D29–33 (Database issue).
16. Leinonen R, et al. *Nucl Acids Res* 2011;**39**:D19–21 (Database issue).
17. NCBI. *Sequence identifiers: a historical note*. Available online at: <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html>; 2004.
18. Choudhuri S, et al. *Biochem Biophys Res Commun* 2001;**280**:92–8.
19. Cattori V, et al. *FEBS Lett* 2000;**474**:242–5.
20. Choudhuri S, et al. *Biochem Biophys Res Commun* 2000;**274**:79–86.
21. Kitts A, Sherry S. Chapter 5: the single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. In: McEntyre J, Ostell J, editors. *The NCBI handbook*. Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK21088/>; 2002.
22. NCBI. *Sample GenBank Record*. Available online at: <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>; 2006.
23. Pruitt K, et al. Chapter 18. The reference sequence (RefSeq) database. In: McEntyre J, Ostell J, editors. *The NCBI handbook*. Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK21091/>; 2002.
24. Boutet E, et al. *Methods Mol Biol* 2007;**406**:89–112.
25. UniProt Consortium. *How Redundant are the UniProt Databases?* Available online at: <http://www.uniprot.org/faq/33>; 2012.
26. UniProt Consortium. *Nucl Acids Res* 2013;**41**:D43–7 (Database issue).
27. Apweiler R, et al. *Curr Opin Chem Biol* 2004;**8**:76–80.
28. Wu C, et al. *Nucl Acids Res* 2003;**31**:345–7.
29. Berman HM, et al. *Nature Struct Biol* 2003;**10**:980.
30. Andreeva A, et al. *Nucl Acids Res* 2008;**36**:D419–25 (Database issue).
31. Sillitoe I, et al. *Nucl Acids Res* 2013;**41**:D490–8 (Database issue).
32. Sigrist CJA, et al. *Nucl Acids Res* 2013;**41**:D344–7 (Database issue).
33. Attwood TK, et al. *Database (Oxford)* 2012. bas019. doi: 10.1093/database/bas019
34. Punta, et al. *Nucl Acids Res* 2012;**40**:D290–301 (Database issue).
35. Hunter S, et al. *Nucl Acids Res* 2011;**40**:D306–12 (Database issue).
36. Chatr-Aryamontri A, et al. *Nucl Acids Res* 2013;**41**:D816–23 (Database issue).
37. Ceol A, et al. *Nucl Acids Res* 2010;**38**:D532–9 (Database issue).
38. Licata L, et al. *Nucl Acids Res* 2012;**40**:D857–61 (Database issue).
39. Pagel P, et al. *Bioinformatics* 2005;**21**:832–4.
40. Mewes HW, et al. *Nucl Acids Res* 2011;**39**:D220–4 (Database issue).
41. Kerrien S, et al. *Nucl Acids Res* 2012;**40**:D841–6 (Database issue).
42. Fiers MW, et al. *BMC Bioinformatics* 2004;**5**:133.
43. Rustici G, et al. *Nucl Acids Res* 2013;**41**:D987–90 (Database issue).
44. Barrett T, et al. *Nucl Acids Res* 2011;**39**:D1005–10 (Database issue).
45. Barrett T, et al. *Nucl Acids Res* 2013;**41**:D991–5 (Database issue).
46. Tong W, et al. *EHP Toxicogenomics* 2003;**111**:1819–26.
47. Davis AP, et al. *Nucl Acids Res* 2013;**41**:D1104–14 (Database issue).
48. Waters M, et al. *Nucl Acids Res* 2008;**36**:D892–900 (Database issue).
49. McQuilton P, et al. *Nucl Acids Res* 2012;**40**:D706–14 (Database issue).
50. Marygold SJ, et al. *Nucl Acids Res* 2013;**41**:D751–7 (Database issue).
51. Brazma A, et al. *Nat Genet* 2001;**29**:365–71.
52. Brazma A. *Sci World J* 2009;**9**:420–3.
53. Sayers EW, et al. *Nucl Acids Res* 2011;**39**:D38–51 (Database issue).
54. Acland A, et al. *Nucl Acids Res* 2013;**41**:D8–20 (Database issue).
55. Fujibuchi, et al. *Pac Symp Biocomput* 1998;683–94.
56. Choudhuri S, et al. *Biochem Biophys Res Commun* 2001;**280**:92–8.
57. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**:520–62.
58. Bult CJ, et al. *Nucl Acids Res* 2013;**41**:D885–91 (Database issue).
59. Flicek P, et al. *Nucl Acids Res* 2013;**41**:D48–55 (Database issue).
60. Pruitt KD, et al. *Genome Res* 2009;**19**:1316–23.
61. Kent WJ, et al. *Genome Res* 2002;**12**:996–1006.
62. Kuhn RM, et al. *Brief Bioinform* 2013;**14**:144–61.
63. Meyer LR, et al. *Nucl Acids Res* 2013;**41**:D64–9 (Database issue).
64. Karolchik D, et al. *Nucl Acids Res* 2004;**32**:D493–6 (Database issue).
65. Kent WJ. *Genome Res* 2002;**12**:656–64.
66. Loveland J. *Brief Bioinform* 2005;**6**:189–93.
67. Ouellette BFF, Boguski MS. *Genome Res* 1997;**7**:952–5.

CHAPTER

6

Sequence Alignment and Similarity Searching in Genomic Databases: BLAST and FASTA*

OUTLINE

6.1 EVOLUTIONARY BASIS OF SEQUENCE ALIGNMENT

As discussed in Chapter 2, evolution is defined as "descent with modification" from a common ancestor. At the molecular level, the modification means changes in DNA and protein sequence, and corresponding changes in protein function. As mutations accumulate in sequences derived from an ancestral sequence, the derived sequences diverge from one another over time, but sections of the sequences may still retain enough similarity to allow identification of a common ancestry.

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

Evolutionary change in a sequence does not always have to be large; slight changes in certain crucial sections of a sequence can have profound functional consequences.

Expectedly, sequence comparison through sequence alignment is central to most bioinformatic analysis. It is the first step towards understanding the evolutionary relationship and the pattern of divergence between two sequences. The relationship between two sequences also helps predict the potential function of an unknown sequence, thereby indicating protein family relationship.

6.2 THREE TERMS—SEQUENCE IDENTITY, SEQUENCE SIMILARITY, AND SEQUENCE HOMOLOGY—AND THEIR PROPER USAGE

Sequence identity means the same residues being present at corresponding positions in two sequences being compared. For proteins, it means the same amino acids; for nucleic acids, it means the same bases.

Sequence similarity means similar residues being present at corresponding positions in the two sequences being compared. For nucleic acids, sequence similarity and sequence identity are the same. However, for proteins, sequence similarity involves amino acids with similar physicochemical and functional properties. For example, substitution of lysine and arginine by one another will be regarded as similar substitution because both are positively charged hydrophilic amino acids. Likewise, substitution of aspartic acid and glutamic acid by one another will be regarded as similar substitution because both are negatively charged hydrophilic amino acids. Substitution of asparagine by aspartic acid and substitution of glutamine by glutamic acid, or vice versa, are also regarded as similar substitutions. Substitution of isoleucine, leucine, and valine by one another will be regarded as similar substitutions because they have similar aliphatic hydrophobic side chains. Substitution of serine and threonine by one another is also regarded as similar substitution. Similar substitutions are also referred to as **conservative substitutions**[a]. A conservative amino acid substitution is not expected to disrupt the structural/functional attributes of the protein.

Sequence homology is an evolutionary term that has been misused the most in the literature to denote sequence similarity or identity. Sequences are called homologous if they have a common evolutionary origin—that is, if they are derived from a common ancestral sequence. So, sequences are either homologous or not homologous and there is no quantitation of homology. However, even now, expressions like "high homology," "significant homology," and even specifying a "% homology" are very widely used. Such usage has no reference to the evolutionary underpinning of the term **homology**. The root of the term homology goes back to the early evolutionary literature, where organs having similar structure and anatomical origin but performing different functions (hence morphologically different) were called homologous organs. Examples of homologous organs are bats' wings, whales' flippers, and human hands; these are all mammalian forelimbs that are morphologically different because they are adapted to perform different functions. Conversely, organs having different structure and anatomical origin but performing the same function (hence morphologically similar) were called analogous organs. Such a character state (analogous organs) shared by a set of species but not present in their common ancestor is also called **homoplasy**. Examples of analogous organs/homoplasy are bats' wings and butterflies' wings, and dolphins' flippers and sharks' fins. Homoplasy is the result of **convergent evolution** in which unrelated species develop similar morphological structures because of adaptation to the same or a similar environment.

In the case of nucleic acid or protein sequence, a high degree of identity/similarity usually suggests homology as well. However, conclusions about homology are largely conjecture because we cannot go back in time and test the sequence in the ancestor and the descendants. Therefore, it is the quantitative identity/similarity between the two sequences that is used to conclude whether the two sequences are homologous or not. For example, metallothionein-1 proteins in rat and mouse (61 amino acids in both) are 95% identical and 98% similar[b]. Rat and mouse diverged about 33 million years ago.[1] Therefore, based on the substitution of three amino acids in 33 million years, the substitution rate per site per year can be calculated, and it can be concluded with a great deal of certainty that rat and mouse metallothionein-1 were derived from a common ancestor, and have not changed much, probably because of functional constraints; hence, they are **homologous**. Homologous genes in different species performing the same function are called **orthologs**. So, the metallothionein-1 genes in rat and mouse are also orthologs. The problem in drawing conclusions on

[a]*Similar substitution* and *conservative substitution* refer to amino acid substitution in protein. *Synonymous substitution* and *nonsynonymous substitution* refer to nucleotide substitution in DNA. *Synonymous substitution* leads to no changes in amino acids in the encoded protein, while *nonsynonymous substitution* leads to changes in amino acids in the encoded protein.

[b]95% identity (58 identical amino acids; hence $(58/61) \times 100 = 95\%$ identity); 98% similarity (58 identical amino acids + 2 similar substitutions; hence $(60/61) \times 100 = 98\%$ similarity).

homology arises when the similarity between two sequences is low. Conclusions on homology, in this case, are drawn on a case-by-case basis. Two proteins can be considered homologous despite low similarity if one or more of the following conditions are met: (1) the similarity extends over a long stretch of sequence and is statistically significant; (2) despite low sequence similarity, the same pattern of identical and similar amino acid residues is seen in multiple sequences; or (3) the pattern of sequence similarity reflects the similarity between experimentally determined structures of the respective proteins, or at least corresponds to the known key elements of one such structure.[2]

6.3 SEQUENCE IDENTITY AND SEQUENCE SIMILARITY

Sequence identity and sequence similarity can be calculated based on the proportion of identical and similar amino acids, respectively:

$$\%\,\text{Identity(PID)} = (\#\,\text{of identical amino acids}/\#\,\text{of total amino acids}) \times 100; \tag{6.1}$$

$$\%\,\text{Similarity} = \{(\#\,\text{of identical amino acids} + \#\,\text{of similar substitutions})/\#\,\text{of total amino acids}\} \times 100 \tag{6.2}$$

In the above formulae, the denominator (# of total amino acids) can vary. For example, the denominator could be (1) the length of the shortest sequence, (2) the length of the longest sequence, (3) the mean length of the sequences, (4) the length of the aligned region (aligned positions excluding overhangs), etc. Therefore, PID is a rough measure and can be influenced by how it is calculated. However, because of the simplicity of calculation, PID is widely used.[3]

The pairwise alignment in Figure 6.1 (National Center for Biotechnology Information (NCBI) BLAST pairwise alignment; http://blast.ncbi.nlm.nih.gov/ → check the "Align two or more sequences" link) and that in Figure 6.2 (EMBOSS Needle of the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI); http://www.ebi.ac.uk/Tools/psa/) show that there are 560 identical amino acids and 53 similar substitutions (making 560 + 53 = 613 similar amino acids) between the rlst-1a and mlst-1 proteins[c]. This makes the identity 81% and the similarity 88.7%. Note that the NCBI designates % similarity as % positive.

6.4 GLOBAL VERSUS LOCAL ALIGNMENT

A **global sequence-alignment** method aligns and compares two sequences along their entire length, and comes up with the best alignment that displays the maximum number of nucleotides or amino acids aligned. The algorithm that drives global alignment is the **Needleman–Wunsch algorithm**. A global alignment algorithm starts at the beginning of two sequences and adds gaps to each until the end of one is reached. *Global alignment works the best when the sequences are similar in character and length*. Because global alignment displays the best alignment between two sequences using the entire sequence, it may miss a small region of biological importance. This is a trade-off in global alignment.

Two of the available web servers for pairwise global alignment are **EMBL-EBI EMBOSS** (http://www.ebi.ac.uk/Tools/psa/), and **NCBI specialized BLAST** (look for the Global Sequence Alignment Tool link on the NCBI BLAST home page under Specialized BLAST; the URL is too long to include here). For EMBL-EBI EMBOSS, the page that appears by clicking the link provides separate options for protein and nucleotide global alignment. **EMBOSS Stretcher** uses a modification of the Needleman–Wunsch algorithm that allows larger sequences to be globally aligned; it also provides separate options for proteins and nucleic acids.

In contrast to global alignment, **local sequence alignment** is intended to find the most similar regions in two sequences being aligned. The algorithm that drives local alignment is the **Smith–Waterman algorithm**. A local alignment algorithm finds the region of highest similarity between two sequences and builds the alignment outward from this region. If there are multiple regions of very high similarity, the same principle applies. *Obviously, local alignment is useful for sequences that are not similar in character and length, yet are suspected to contain small regions of similarity, such as biologically important motifs.*

The global and local alignments involving two protein sequences that are significantly similar produce identical results. For example, running a global alignment using the Needleman–Wunsch algorithm or a local alignment using the Smith–Waterman algorithm (discussed below) for the rlst-1a and mlst-1 proteins produces identical results. Pairwise global alignment using both RNA (complementary DNA, or cDNA) and protein sequences can identify alternatively spliced variants. Figure 6.3 (EMBOSS Needle of the EMBL-EBI) shows that rlst-1c protein, which is an alternatively spliced form, lacks a segment of 33 amino acids that is present

[c]The original submission accession number of rlst-1a is AF208545 and that of mlst-1 is AB031959.

Query: rlst-1a; 687 amino acids (Accession #: AAF87098.1)
Sbjct: mlst-1; 689 amino acids (Accession #: BAB03272.1)

Score	Expect	Method	Identities	Positives	Gaps	Frame
1166 bits(3016)	0.0	Compositional matrix adjust	560/691(81%)	613/691(88%)	6/691(0%)	

```
Query  1    MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKSSITQIERRFD  60
            MD TQ   KAA  QP RS++TR CDGF++FLAALSFSYICKALGGV+MKSSITQIERRFD
Sbjct  1    MDQTQHPSKAA--QPLRSEKTRHCDGFRIFLAALSFSYICKALGGVIMKSSITQIERRFD  58

Query  61   IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIGCFIMGIGSILTALPHFFMGY  120
            IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIG GCFIMGIGSILTALPHFFMGY
Sbjct  59   IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGTGCFIMGIGSILTALPHFFMGY  118

Query  121  YKYAKENDIGSLGNSTLTCFINQMTSPTGPSPEIVEKGCEKGLKSHMWIYVLMGNMLRGI  180
            Y+YA ENDI SL NSTLTC +NQ TS TG SPEI+EKGCEKG  S+ WIYVLMGNMLRGI
Sbjct  119  YRYATENDISSLHNSTLTCLVNQTTSLTGTSPEIMEKGCEKGSNSYTWIYVLMGNMLRGI  178

Query  181  GETPIVPLGISYLDDFAKEGHTSMHLGTLHTIAMIGPILGFIMSSVFAKIYVDVGYVDLN  240
            GETPIVPLG+SY+DDFAKEG++SM+LGTLHTIAMIGPILGFIMSSVFAK+YVDVGYVDL
Sbjct  179  GETPIVPLGVSYIDDFAKEGNSSMYLGTLHTIAMIGPILGFIMSSVFAKLYVDVGYVDLR  238

Query  241  SVRITPNDARWVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE  300
            SVRITP DARWVGAWWL FIVNGLLCI  SIPFFFLPKIPKRSQ+ERKNS SLH  KTDE
Sbjct  239  SVRITPQDARWVGAWWLGFIVNGLLCIICSIPFFFLPKIPKRSQKERKNSASLHVLKTDE  298

Query  301  EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFIGSFTYLFKFI  360
            +K  +TN T QE+Q P+N+TGFL SLRSILTNE YVIFLILTLLQ+S FIGSFTYLFKFI
Sbjct  299  DKNPVTNPTTQEKQAPANLTGFLWSLRSILTNEQYVIFLILTLLQISSFIGSFTYLFKFI  358

Query  361  EQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLTSVGIAKFVFFTSSVAYAFQF  420
            EQQFG+TASQANFLLG+ITIPTMA+ MFLGGY++K+ KLT +GI KFVFFT+++AY F
Sbjct  359  EQQFGQTASQANFLLGVITIPTMASGMFLGGYLIKRLKLTLLGITKFVFFTTTMAYVFYL  418

Query  421  LYFPLLCENKPFAGLTLTYDGMNPVDSHIDVPLSYCNSDCSCDKNQWEPICGENGVTYIS  480
             YF L+CENK FAGLTLTYDGMNPVDSHIDVPLSYCNSDC CDKNQWEP+CGENGVTYIS
Sbjct  419  SYFLLICENKAFAGLTLTYDGMNPVDSHIDVPLSYCNSDCICDKNQWEPVCGENGVTYIS  478

Query  481  PCLAGCKSFRGDKKPNNTEFYDCSCISNS----GNNSAHLGECPRYKCKTNYYFYIILQV  536
            PCLAGCKSFRGDKK  N EFYDCSC+S S    GN+SA LGECPR KCKT YYFYI  QV
Sbjct  479  PCLAGCKSFRGDKKLMNIEFYDCSCVSGSGFQKGNHSARLGECPRDKCKTKYYFYITFQV  538

Query  537  TVSFFTAMGSPSLILILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKW  596
             +SFFTA+GS SL+LIL++SVQPELKSL MGFHSL++R LGGILAP+YYGA IDRTC+KW
Sbjct  539  IISFFTALGSTSLMLILIRSVQPELKSLGMGFHSLVVRTLGGILAPVYYGALIDRTCMKW  598

Query  597  SVTSCGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYKRNDNKTLENG  656
            SVTSCG RGACRLYNSRLFG  Y+GL++ALKTP L LYV LIY  KRK KRNDNK LENG
Sbjct  599  SVTSCGARGACRLYNSRLFGMIYVGLSIALKTPILLLYVALIYVMKRKMKRNDNKILENG  658

Query  657  RQFTDEGNPDSVNKNGYYCVPYDEQSNETPL  687
            R+FTDEGNP+ VN NGY CVP DE+++ETPL
Sbjct  659  RKFTDEGNPEPVNNNGYSCVPSDEKNSETPL  689
```

FIGURE 6.1 **Pairwise alignment of rlst-1a and mlst-1 proteins using NCBI BLAST.** NCBI BLAST pairwise alignment shows that these two proteins share 81% identity but 88.7% similarity. The similar amino acids are highlighted in gray; many of these are hydrophobic amino acids, charged polar amino acids, and neutral polar amino acids. In the NCBI BLAST pairwise alignment format, the identical amino acids and similar substitutions between the query and the subject sequences are in the middle; and similar substitutions are indicated by a + sign.

```
# Aligned_sequences: 2
# 1: rlst-1a (Accession #: AAF87098.1)
# 2: mlst-1  (Accession #: BAB03272.1)
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 691
# Identity:     560/691 (81.0%)
# Similarity:   613/691 (88.7%)
# Gaps:           6/691 ( 0.9%)
# Score: 2986.0

rlst-1a           1 MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKS     50
                    ||.||...||  |||.||::||.||||::||||||||||||||||:|||
mlst-1            1 MDQTQHPSKA--AQPLRSEKTRHCDGFRIFLAALSFSYICKALGGVIMKS     48

rlst-1a          51 SITQIERRFDIPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIG    100
                    ||||||||||||||||||||||||||||||||||||||||||||||||.|
mlst-1           49 SITQIERRFDIPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGTG     98

rlst-1a         101 CFIMGIGSILTALPHFFMGYYKYAKENDIGSLGNSTLTCFINQMTSPTGP    150
                    ||||||||||||||||||||:||.||||.||.||||||.:||.||.||.
mlst-1           99 CFIMGIGSILTALPHFFMGYYRYATENDISSLHNSTLTCLVNQTTSLTGT    148

rlst-1a         151 SPEIVEKGCEKGLKSHMWIYVLMGNMLRGIGETPIVPLGISYLDDFAKEG    200
                    ||||:|||||||..|:.||||||||||||||||||:||:|||||||
mlst-1          149 SPEIMEKGCEKGSNSYTWIYVLMGNMLRGIGETPIVPLGVSYIDDFAKEG    198

rlst-1a         201 HTSMHLGTLHTIAMIGPILGFIMSSVFAKIYVDVGYVDLNSVRITPNDAR    250
                    ::||:|||||||||||||||||||||:|||||||||.||||||.|||
mlst-1          199 NSSMYLGTLHTIAMIGPILGFIMSSVFAKLYVDVGYVDLRSVRITPQDAR    248

rlst-1a         251 WVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE    300
                    |||||||.|||||||..|||||||||||||||:|||||.|||..||||
mlst-1          249 WVGAWWLGFIVNGLLCIICSIPFFFLPKIPKRSQKERKNSASLHVLKTDE    298

rlst-1a         301 EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFI    350
                    :|..:||.|.||:|.|:|:||||.|||||||.||||||||||:|.||
mlst-1          299 DKNPVTNPTTQEKQAPANLTGFLWSLRSILTNEQYVIFLILTLLQISSFI    348

rlst-1a         351 GSFTYLFKFIEQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLT    400
                    |||||||||||||:||||||||||:||||||||:.||||||::|:.|||
mlst-1          349 GSFTYLFKFIEQQFGQTASQANFLLGVITIPTMASGMFLGGYLIKRLKLT    398

rlst-1a         401 SVGIAKFVFFTSSVAYAFQFLYFPLLCENKPFAGLTLTYDGMNPVDSHID    450
                    .:||.||||||:::||.|...||.|:||||.||||||||||||||||||
mlst-1          399 LLGITKFVFFTTTMAYVFYLSYFLLICENKAFAGLTLTYDGMNPVDSHID    448

rlst-1a         451 VPLSYCNSDCSCDKNQWEPICGENGVTYISPCLAGCKSFRGDKKPNNTEF    500
                    ||||||||||.||||||||:||||||||||||||||||||||..|.||
mlst-1          449 VPLSYCNSDCICDKNQWEPVCGENGVTYISPCLAGCKSFRGDKKLMNIEF    498

rlst-1a         501 YDCSCISNS----GNNSAHLGECPRYKCKTNYYFYIILQVTVSFFTAMGS    546
                    |||||:|.|    ||:||.||||||.||||.|||||..||.:||||||||
mlst-1          499 YDCSCVSGSGFQKGNHSARLGECPRDKCKTKYYFYITFQVIISFFTALGS    548

rlst-1a         547 PSLILILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKW    596
                    .||:|||::|||||||||.|||||::|.|||||||:||||.|||||:||
mlst-1          549 TSLMLILIRSVQPELKSLGMGFHSLVVRTLGGILAPVYYGALIDRTCMKW    598

rlst-1a         597 SVTSCGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYK    646
                    ||||||.||||||||||..|:|||::||||.|.|||.|||..|||.|
mlst-1          599 SVTSCGARGACRLYNSRLFGMIYVGLSIALKTPILLLYVALIYVMKRKMK    648

rlst-1a         647 RNDNKTLENGRQFTDEGNPDSVNKNGYYCVPYDEQSNETPL    687
                    |||||.|||||:|||||||:.||.|||.|||.||:::||||
mlst-1          649 RNDNKILENGRKFTDEGNPEPVNNNGYSCVPSDEKNSETPL    689
```

FIGURE 6.2 **Pairwise global alignment of rlst-1a and mlst-1 proteins using EMBL-EBI EMBOSS.** EMBOSS Needle (Needleman–Wunsch algorithm) shows that these two proteins share 81% identity but 88.7% similarity. The similar amino acids are highlighted in grey.

```
# Aligned_sequences: 2
# 1: rlst-1a (Accession #: AAF87098.1)
# 2: rlst-1c (Accession #: AAF87099.1)
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 687
# Identity:     653/687 (95.1%)
# Similarity:   654/687 (95.2%)
# Gaps:          33/687 ( 4.8%)
# Score: 3388

rlst-1a           1 MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKS     50
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c           1 MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKS     50

rlst-1a          51 SITQIERRFDIPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIG    100
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c          51 SITQIERRFDIPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIG    100

rlst-1a         101 CFIMGIGSILTALPHFFMGYYKYAKENDIGSLGNSTLTCFINQMTSPTGP    150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         101 CFIMGIGSILTALPHFFMGYYKYAKENDIGSLGNSTLTCFINQMTSPTGP    150

rlst-1a         151 SPEIVEKGCEKGLKSHMWIYVLMGNMLRGIGETPIVPLGISYLDDFAKEG    200
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         151 SPEIVEKGCEKGLKSHMWIYVLMGNMLRGIGETPIVPLGISYLDDFAKEG    200

rlst-1a         201 HTSMHLGTLHTIAMIGPILGFIMSSVFAKIYVDVGYVDLNSVRITPNDAR    250
                    ||||||                              :||||||||||
rlst-1c         201 HTSMHL----------------------------------DSVRITPNDAR    217

rlst-1a         251 WVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE    300
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         218 WVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE    267

rlst-1a         301 EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFI    350
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         268 EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFI    317

rlst-1a         351 GSFTYLFKFIEQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLT    400
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         318 GSFTYLFKFIEQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLT    367

rlst-1a         401 SVGIAKFVFFTSSVAYAFQFLYFPLLCENKPFAGLTLTYDGMNPVDSHID    450
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         368 SVGIAKFVFFTSSVAYAFQFLYFPLLCENKPFAGLTLTYDGMNPVDSHID    417

rlst-1a         451 VPLSYCNSDCSCDKNQWEPICGENGVTYISPCLAGCKSFRGDKKPNNTEF    500
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         418 VPLSYCNSDCSCDKNQWEPICGENGVTYISPCLAGCKSFRGDKKPNNTEF    467

rlst-1a         501 YDCSCISNSGNNSAHLGECPRYKCKTNYYFYIILQVTVSFFTAMGSPSLI    550
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         468 YDCSCISNSGNNSAHLGECPRYKCKTNYYFYIILQVTVSFFTAMGSPSLI    517

rlst-1a         551 LILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKWSVTS    600
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         518 LILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKWSVTS    567

rlst-1a         601 CGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYKRNDN    650
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
rlst-1c         568 CGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYKRNDN    617

rlst-1a         651 KTLENGRQFTDEGNPDSVNKNGYYCVPYDEQSNETPL    687
                    |||||||||||||||||||||||||||||||||||||
rlst-1c         618 KTLENGRQFTDEGNPDSVNKNGYYCVPYDEQSNETPL    654
```

FIGURE 6.3 **Pairwise global alignment of rlst-1a and rlst-1c proteins using EMBL-EBI EMBOSS.** EMBOSS Needle (Needleman–Wunsch algorithm) shows that the rlst-1c protein is an alternatively spliced form missing a 33-amino-acid segment that is present in the rlst-1a protein (highlighted).

```
Sequence format is CLUSTAL (CLUSTAL 2.1 Multiple Sequence Alignments)
Sequence 1: rlst-1a      687 aa (Accession #: AAF87098.1)
Sequence 2: rlst-1c      687 aa (Accession #: AAF87099.1)

rlst-1a         MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKSSITQIERRFD 60
rlst-1c         MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKSSITQIERRFD 60
                ************************************************************

rlst-1a         IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIGCFIMGIGSILTALPHFFMGY 120
rlst-1c         IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIGCFIMGIGSILTALPHFFMGY 120
                ************************************************************

rlst-1a         YKYAKENDIGSLGNSTLTCFINQMTSPTGPSPEIVEKGCEKGLKSHMWIYVLMGNMLRGI 180
rlst-1c         YKYAKENDIGSLGNSTLTCFINQMTSPTGPSPEIVEKGCEKGLKSHMWIYVLMGNMLRGI 180
                ************************************************************

rlst-1a         GETPIVPLGISYLDDFAKEGHTSMHLGTLHTIAMIGPILGFIMSSVFAKIYVDVGYVDLN 240
rlst-1c         GETPIVPLGISYLDDFAKEGHTSMHL--------------------------------D 207
                ************************                                :

rlst-1a         SVRITPNDARWVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE 300
rlst-1c         SVRITPNDARWVGAWWLSFIVNGLLCITSSIPFFFLPKIPKRSQEERKNSVSLHAPKTDE 267
                ************************************************************

rlst-1a         EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFIGSFTYLFKFI 360
rlst-1c         EKKHMTNLTKQEEQDPSNMTGFLRSLRSILTNEIYVIFLILTLLQVSGFIGSFTYLFKFI 327
                ************************************************************

rlst-1a         EQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLTSVGIAKFVFFTSSVAYAFQF 420
rlst-1c         EQQFGRTASQANFLLGIITIPTMATAMFLGGYIVKKFKLTSVGIAKFVFFTSSVAYAFQF 387
                ************************************************************

rlst-1a         LYFPLLCENKPFAGLTLTYDGMNPVDSHIDVPLSYCNSDCSCDKNQWEPICGENGVTYIS 480
rlst-1c         LYFPLLCENKPFAGLTLTYDGMNPVDSHIDVPLSYCNSDCSCDKNQWEPICGENGVTYIS 447
                ************************************************************

rlst-1a         PCLAGCKSFRGDKKPNNTEFYDCSCISNSGNNSAHLGECPRYKCKTNYYFYIILQVTVSF 540
rlst-1c         PCLAGCKSFRGDKKPNNTEFYDCSCISNSGNNSAHLGECPRYKCKTNYYFYIILQVTVSF 507
                ************************************************************

rlst-1a         FTAMGSPSLILILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKWSVTS 600
rlst-1c         FTAMGSPSLILILMKSVQPELKSLAMGFHSLIIRALGGILAPIYYGAFIDRTCIKWSVTS 567
                ************************************************************

rlst-1a         CGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYKRNDNKTLENGRQFT 660
rlst-1c         CGKRGACRLYNSRLFGFSYLGLNLALKTPPLFLYVVLIYFTKRKYKRNDNKTLENGRQFT 627
                ************************************************************

rlst-1a         DEGNPDSVNKNGYYCVPYDEQSNETPL 687
rlst-1c         DEGNPDSVNKNGYYCVPYDEQSNETPL 654
                ***************************
```

FIGURE 6.4 **Pairwise alignment of rlst-1a and rlst-1c proteins using DDBJ ClustalW.** Analysis using the multiple alignment program ClustalW (DDBJ). The result is the same as that depicted in Figure 6.3. The missing 33-amino-acid segment in rlst-1c is highlighted. *(DDBJ; http://clustalw.ddbj.nig.ac.jp/)*

in rlst-1a protein, which is the full-length form.[4] The pairwise alignment can also be performed using a multiple alignment program, such as ClustalW (DNA Data Bank of Japan (DDBJ); http://clustalw.ddbj.nig.ac.jp/); the result of the analysis is the same (Figure 6.4). *Note that the alignments in Figures 6.1 through 6.4 have been performed using tools from NCBI, EMBL-EBI, and DDBJ in order to provide visual display of different output formats for marking identical amino acids and similar amino acids.*

6.5 PAIRWISE AND MULTIPLE ALIGNMENT

As the name suggests, pairwise alignment aligns two nucleic acid or two protein sequences to find the best match. Multiple alignment performs the same function using more than two sequences. The purpose of alignment is to identify regions of similarity that may have structural, functional, and evolutionary

TABLE 6.1 Online Pairwise Alignment Tools Using the Smith–Waterman Algorithm

Online Tool	URL
PIR SSEARCH	http://pir.georgetown.edu/pirwww/search/pairwise.shtml[5]
NCBI specialized BLAST	bl2seq resource; look for the Align link on the NCBI BLAST home page under Specialized BLAST
SIM	http://web.expasy.org/sim/
LALIGN*	http://www.ch.embnet.org/software/LALIGN_form.html

*The LALIGN program is William Pearson's, and it implements the algorithm of X. Huang and W. Miller.[6]

consequences. Figures 6.1 through 6.4 are examples of pairwise alignment.

Some widely used online pairwise alignment tools use local alignment strategy (Smith–Waterman algorithm) and are shown in Table 6.1.

The NCBI BLAST pairwise alignment tool, SIM, and LALIGN not only show the overall alignment of the two sequences, but will also display, as separate output, multiple matching subsegments between the two sequences being aligned. For example, Figure 6.5 shows the alignment of the partial sequence of mlst-1 and moatp-2 proteins[d] using LALIGN (http://www.ch.embnet.org/software/LALIGN_form.html), which is also accessible from the EMBL-EBI page (http://www.ebi.ac.uk/Tools/psa/lalign/). A hypothetical sequence "THATISGREATANDFANTASTIC" was added at the beginning of the mlst-1 protein and the end of the moatp-2 protein. The two resulting sequences were then aligned using LALIGN and NCBI BLAST pairwise alignment. Both LALIGN (Figure 6.5) and NCBI BLAST pairwise alignment (Figure 6.6) produced an overall alignment of the two input sequences, and also reported the matching subsegment in these two sequences, which is the added hypothetical sequence. Therefore, these tools are very useful in finding various motifs and conserved sequences between two proteins being compared.

Multiple sequence alignments are useful in identifying conserved sequence segments across the sequences being aligned. Such conserved regions across multiple sequences usually indicate an evolutionary relationship. For an unknown protein, for example, such conserved sequence segments identified through multiple alignment can be used in conjunction with other information to predict functionally important and evolutionarily conserved motifs within the proteins. Multiple alignment is also needed for the construction of phylogenetic trees. Figure 6.7 shows multiple alignment of five transporter proteins (partial sequence used) from mouse and rat using DDBJ ClustalW. The T-Coffee, CBRC (Computational Biology Research Center at the National Institute of Advanced Industrial Science and Technology, Japan) MAFFT, and EMBL-EBI MUSCLE all use ClustalW, so the output format is similar. NCBI COBALT has a very different output format. Multiple alignment is frequently done using Clustal programs, such as **ClustalW** and more recently **Clustal Omega**. Clustal Omega is a scaled-up version that enables thousands of sequences to be aligned. In order to perform multiple alignment, the ClustalW algorithm goes through a number of steps, as follows: it calculates all possible pairwise alignments of the input sequences; computes the score of each alignment, where the score reflects the distance between the two sequences; creates a dendrogram (guide tree) based on the matrix of the distance; and uses the dendrogram as the basis to perform multiple alignment, where closely related pairs of sequences are aligned first.

Multiple alignment programs can also be used to run pairwise alignment. Some online multiple alignment tools are shown in Table 6.2. Sequence input needs to be in FASTA or other formats.

6.6 ALIGNMENT ALGORITHMS, GAPS, AND GAP PENALTIES

An algorithm is a step-by-step procedure that utilizes a finite number of instructions for automated reasoning and the calculation of a function. The algorithm that drives global alignment is the Needleman–Wunsch algorithm, and the algorithm that drives local alignment is the Smith–Waterman algorithm. Both these algorithms are examples of **dynamic programming**. Dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems. In the case of sequence alignment, dynamic programming involves setting up a two-dimensional matrix in which one sequence is listed vertically and the other sequence is listed horizontally; then calculating the scores, one row at a time. For example, a match can be given a 1, a mismatch a 0, and a gap a −1. A 100% perfect alignment will produce a diagonal straight line (with a negative slope) spanning from the top left to bottom right. If the alignment is not perfect, gaps are introduced in the matrix. For the sequence represented horizontally, gaps are introduced vertically, and for the sequence represented vertically, gaps are introduced

[d]The original submission accession number of mlst-1 is AB031959 and that of moatp-2 is AB031814. Partial sequence for each entry is used to save space.

```
LALIGN output
   mlst-1 (BAB03272.1; partial sequence)
   moatp-2 (BAB12445.1; partial sequence).
A hypothetical sequence THATISGREATANDFANTASTIC was added to the beginning
of mlst-1 protein and the end of moatp-2 protein


 49.1% identity in 212 aa overlap (31-241:2-210); score:  681 E(10000): 1.2e-61

               40        50        60        70        80        90
mlst-1 SKAAQPLRSEKTRHCDGFRIFLAALSFSYICKALGGVIMKSSITQIERRFDIPSSISGLI
       .:. . . .. .:  . .. :: ::. .:. :.:.:. :.: .::::::.: ::.:. :::
moatp- GKSEKEVATHGVRCFSKIKAFLLALTCAYVSKSLSGTYMNSMLTQIERQFGIPTSVVGLI
              10        20        30        40        50        60

              100       110       120       130       140
mlst-1 DGGFEIGNLLVIVFVSYFGSKLHRPKLIGTGCFIMGIGSILTALPHFFMGYYRYATEN-D
       .:.::::::::.:.::::::.::::: .::.:: .::.: .: ..::::.:: :.: :
moatp- NGSFEIGNLLLIIFVSYFGTKLHRPIMIGVGCAVMGLGCFLISIPHFLMGRYEYETTILP
              70        80        90       100       110       120

     150       160       170       180       190       200
mlst-1 ISSLHNSTLTCLVNQTTSLTGTSPEIMEKGCEKGSNSYTWIYVLMGNMLRGIGETPIVPL
        :.: .....:  :.: .:   .:      : :  .:  :::::.::..::.::::::.::
moatp- TSNLSSNSFVCTENRTQTL---KPTQDPTECVKEMKSLMWIYVLVGNIIRGMGETPIMPL
             130       140          150       160       170

     210       220       230       240
mlst-1 GVSYIDDFAKEGNSSMYLGTLHTIAMIGPILG
       :.:::.::::  :: .:.: :.:   :::..:
moatp- GISYIEDFAKSENSPLYIGILETGMTIGPLIG
      180       190       200       210

-------------------------------------------------------------------------------

 100.0% identity in 23 aa overlap (1-23:219-241); score:  144 E(10000): 1.5e-07

               10        20
mlst-1 THATISGREATANDFANTASTIC
       ::::::::::::::::::::::::
moatp- THATISGREATANDFANTASTIC
      220       230       240

-------------------------------------------------------------------------------

 40.0% identity in 15 aa overlap (117-131:95-109); score:   50 E(10000): 4.3e+02

        120       130
mlst-1 LIGTGCFIMGIGSIL
       ..: :::...:  .:
moatp- VMGLGCFLISIPHFL
          100
```

FIGURE 6.5 **LALIGN pairwise comparison.** LALIGN output of pairwise comparison of mlst-1 (BAB03272.1; partial sequence) and moatp-2 (BAB12445.1; partial sequence) each containing the hypothetical sequence "THATISGREATANDFANTASTIC." LALIGN produces an overall alignment of two protein sequences and also finds matching subsegments shared by these two input sequences. Note that in LALIGN the identities are reported by two dots and similar substitutions are reported by one dot.

horizontally, and the alignment is determined by a traceback step. The basic sequence alignment method is the dot matrix or dot plot method. In this method, two sequences being compared are written in the vertical and horizontal axes of the matrix. Then each residue is scanned and each match is given a dot; mismatches are left blank. When enough dots are lined up, they are connected (Figure 6.8).

In both global and local alignment, the final output is given an **alignment score**. **Gaps** have to be introduced to improve the alignment. The reason gaps are introduced is because one of the sequences may have gained or lost sequence characteristics (insertion−deletion) during evolution that did not happen with the other sequence. However, the number of gaps is kept to a minimum to keep the

```
Query: mlst-1  (BAB03272.1; partial sequence)
Sbjct: moatp-2 (BAB12445.1; partial sequence)
```

Range 1: 3 to 210 Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
209 bits(533)	2e-71	Compositional matrix adjust	104/211(49%)	143/211(67%)	4/211(1%)	

```
Query  32   KAAQPLRSEKTRHCDGFRIFLAALSFSYICKALGGVIMKSSITQIERRFDIPSSISGLID  91
            K+ + + +   R     + FL AL+ +Y+ K+L G  M S +TQIER+F IP+S+ GLI+
Sbjct  3    KSEKEVATHGVRCFSKIKAFLLALTCAYVSKSLSGTYMNSMLTQIERQFGIPTSVVGLIN  62

Query  92   GGFEIGNLLVIVFVSYFGSKLHRPKLIGTGCFIMGIGSILTALPHFFMGYYRYATEN-DI  150
            G FEIGNLL+I+FVSYFG+KLHRP +IG GC +MG+G  L ++PHF MG Y Y T
Sbjct  63   GSFEIGNLLLIIFVSYFGTKLHRPIMIGVGCAVMGLGCFLISIPHFLMGRYEYETTILPT  122

Query  151  SSLHNSTLTCLVNQTTSLTGTSPEIMEKGCEKGSNSYTWIYVLMGNMLRGIGETPIVPLG  210
            S+L +++  C  N+T +L    P      C K   S  WIYVL+GN++RG+GETPI+PLG
Sbjct  123  SNLSSNSFVCTENRTQTL---KPTQDPTECVKEMKSLMWIYVLVGNIIRGMGETPIMPLG  179

Query  211  VSYIDDFAKEGNSSMYLGTLHTIAMIGPILG  241
            +SYI+DFAK  NS +Y+G L T   IGP++G
Sbjct  180  ISYIEDFAKSENSPLYIGILETGMTIGPLIG  210
```

Range 2: 219 to 241 Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
50.4 bits(119)	4e-12	Compositional matrix adjust	23/23(100%)	23/23(100%)	0/23(0%)	

```
Query  1    THATISGREATANDFANTASTIC  23
            THATISGREATANDFANTASTIC
Sbjct  219  THATISGREATANDFANTASTIC  241
```

FIGURE 6.6 **NCBI BLAST pairwise alignment.** The two partial sequences depicted in Figure 6.5 were also aligned using NCBI BLAST pairwise alignment. Like LALIGN, NCBI BLAST pairwise alignment also produces an overall alignment of two protein sequences, and also finds matching subsegments shared by these two sequences. The hypothetical sequence "THATISGREATANDFANTASTIC" has been identified as a subsegment of 100% identity between the two proteins.

alignment meaningful; otherwise an artificially high alignment score can be obtained even when the two sequences are not related. The **gap penalty** value is subtracted from the gross alignment score to obtain the final alignment score (alignment score and scoring matrix are discussed in the next section). *The insertion of no more than 1 gap per 20 amino acid residues is ideal but that is not possible in most cases*. For each gap opened, a **gap-opening penalty** value is assigned, and for each gap extended, a **gap-extension penalty** value is assigned. A gap-opening penalty is always much higher than a gap-extension penalty. Often, a default value of -10 for a gap-opening penalty and -1 for a gap-extension penalty are used. However, these values can be different and can also be adjusted by the user. This type of differential penalty for gap opening and gap extension is called **affine gap penalty**. There are other types of gap penalties, such as constant gap penalty, linear gap penalty, and proportional gap penalty, but for all practical purposes affine gap penalty is the most relevant for sequence alignment. Affine gap penalty is calculated as follows:

$$G_t = G_o + G_e \times L_n, \tag{6.3}$$

where G_t = total gap penalty, G_o = gap-opening penalty, G_e = gap-extension penalty, and L_n = length of the extension gaps. For any given block of gaps, L_n = # of total gaps $-$ 1, because the first gap is the opening, the rest in the block are extensions.

When running an alignment, it is better to use the default value with the default matrix. This is because there is no rule for setting the best gap-opening and -extension penalty values for a given pair of sequences being compared; thus, changing the gap-opening and -extension penalty values may influence the nature of the alignment. For example, setting gap-opening and -extension penalty values that are a lot higher than the default values creates alignments that contain fewer internal gaps and more end gaps; also local alignments containing gaps may be split into several shorter alignments.

```
Sequence format is CLUSTAL (CLUSTAL 2.1 Multiple Sequence Alignments
Sequence 1: moatp-2      287 aa (Accession #: BAB12445.1; partial sequence)
Sequence 2: moatp-5      287 aa (Accession #: AAG60350.1; partial sequence)
Sequence 3: moatp-1      287 aa (Accession #: BAB12444.1; partial sequence)
Sequence 4: rlst-1a      287 aa (Accession #: AAF87098.1; partial sequence)
Sequence 5: mlst-1       287 aa (Accession #: BAB03272.1; partial sequence)

CLUSTAL 2.1 multiple sequence alignment

moatp-2         MGKSE--------KEVATHGVRCFSKIKAFLLALTCAYVSKSLSGTYMNSMLTQIERQFG 52
moatp-5         MGEPG--------KRVGIHRVRCFAKIKVFLLALIWAYISKILSGVYMSTMLTQLERQFN 52
moatp-1         MEETE--------KKVATQEGRFFSKMKVFLMSLTCAYLAKSLSGVYMNSMLTQIERQFG 52
rlst-1a         MDHTQQSRKAAEAQPSRSKQTRFCDGFKLFLAALSFSYICKALGGVVMKSSITQIERRFD 60
mlst-1          MDQTQHPSKA--AQPLRSEKTRHCDGFRIFLAALSFSYICKALGGVIMKSSITQIERRFD 58
                * ..         :    .  *    :: ** :*  :*:.* *.*. *.: :**:**:*.

moatp-2         IPTSVVGLINGSFEIGNLLLIIFVSYFGTKLHRPIMIGVGCAVMGLGCFLISIPHFLMGR 112
moatp-5         ISTSIVGLINGSFEMGNLLVIVFVSYFGTKLHRPIMIGVGCAVMGLGCFIISLPHFLMGR 112
moatp-1         IPTSVVGFITGSFEIGNLLLIVFVSYFGRKLHRPIIIGVGCVVMGLGCFLMASPHFLMGR 112
rlst-1a         IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGIGCFIMGIGSILTALPHFFMGY 120
mlst-1          IPSSISGLIDGGFEIGNLLVIVFVSYFGSKLHRPKLIGTGCFIMGIGSILTALPHFFMGY 118
                *.:*: *:* *.**:****:*:****** ***** :** ** :**:*.:: : ***:**

moatp-2         YEYETTILPTSNLSSNSFVCTENRTQTLK---PTQDPTECVKEMKSLMWIYVLVGNIIRG 169
moatp-5         YEYETTISPTSNLSSNSFLCVENRSQTLK---PTQDPAECVKEIKSLMWIYVLVGNIIRG 169
moatp-1         YKYETTISPTSNLSSNSFLCIENRTQTLK---PTQDPTECVKEIKSLMWIYVLIGNTMRG 169
rlst-1a         YKYAKEND-IGSLGNSTLTCFINQMTSPTGPSPEIVEKGCEKGLKSHMWIYVLMGNMLRG 179
mlst-1          YRYATEND-ISSLHNSTLTCLVNQTTSLTGTSPEIMEKGCEKGSNSYTWIYVLMGNMLRG 177
                *.* .     ..* ..:: *  *:  : .   *      * *  :*  *****:** :**

moatp-2         MGETPIMPLGISYIEDFAKSENSPLYIGILETGMTIGPLIGLLLGSSCANIYVDTGSVNT 229
moatp-5         IGETPIMPLGISYIEDFAKSENSPLYIGILEVGKMIGPILGYLMGPFCANIYVDTGSVNT 229
moatp-1         IGETPIMPLGISYIEDFAKSENSPLYIGILEMGKIVGPIIGLLLGSFFARVYVDIGSVNT 229
rlst-1a         IGETPIVPLGISYLDDFAKEGHTSMHLGTLHTIAMIGPILGFIMSSVFAKIYVDVGYVDL 239
mlst-1          IGETPIVPLGVSYIDDFAKEGNSSMYLGTLHTIAMIGPILGFIMSSVFAKLYVDVGYVDL 237
                :*****:***:**::****. ::.:::* *.    :**::* ::..  *.:*** * *:

moatp-2         DDLTITPTDTRWVGAWWIGFLVCAGVNILTSIPFFFFPKTLLKEGLQ 276
moatp-5         DDLTITPTDTRWVGAWWIGFLVCAGVNVLTSIPFFFFPKTLPKEGLQ 276
moatp-1         DDLTITPTDTRWVGAWWIGFLVCAGVNILTSIPFFFFPKTLPKKELQ 276
rlst-1a         NSVRITPNDARWVGAWWLSFIVNGLLCITSSIPFFFL---------- 276
mlst-1          RSVRITPQDARWVGAWWLGFIVNGLLCIICSIPFFFLPK-------- 276
                 .: *** *:*******:.*:* . : :  ******:
```

FIGURE 6.7 **Multiple alignment using ClustalW from DDBJ.** Five transporters from rat and mouse have been aligned. Identical amino acids are indicated by a star (*), whereas similar substitutions are indicated by a colon (:). To save space, only the first 287 amino acids from each transporter have been used for the alignment.

TABLE 6.2 Online Multiple Alignment Tools

Online Tool	URL
COBALT (NCBI)	http://www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?link_loc = BlastHomeLink[7]
ClustalW (DDBJ)	http://clustalw.ddbj.nig.ac.jp/index.php?lang = en[8,9]
MAFFT (CBRC)	http://mafft.cbrc.jp/alignment/server/[10]
MUSCLE (EMBL-EBI)	http://www.ebi.ac.uk/Tools/msa/muscle/[11]
T-Coffee	http://www.tcoffee.org/Projects/tcoffee/. Then click any of the server links on this page, such as http://www.tcoffee.org/ and from there the type of alignment program needed for analysis[12]

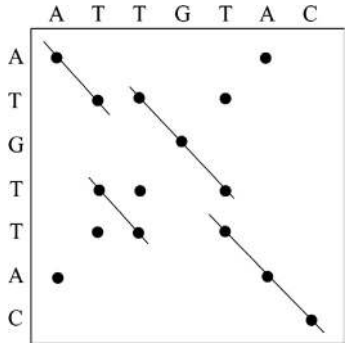


FIGURE 6.8 **Comparison of two sequences using dot matrix or dot plot.**

6.7 SCORING MATRIX, ALIGNMENT SCORE, AND STATISTICAL SIGNIFICANCE OF SEQUENCE ALIGNMENT

A raw alignment score can be calculated based on the following simple formula:

$$S = \Sigma_i + \Sigma_m - G_t, \tag{6.4}$$

where S = raw score, Σ_i = total score for identities, Σ_m = total score for mismatches, and G_t = total gap penalty.

For both nucleic acids and proteins, the alignment score is calculated using a **scoring matrix**. A scoring matrix is a set of values representing the likelihood of one residue being substituted by another during sequence divergence through evolution. This is why the scoring matrix is also known as the **substitution matrix**.

A scoring matrix for comparing DNA sequences can be simple because there are only four nucleotides and the mutation frequencies are assumed to be equal (the Jukes and Cantor assumption). A high positive score (e.g. 5) is assigned for a match and a low negative score (e.g. −4) for a mismatch, thus creating a simple model. However, the frequency of transition mutations (purine replaced by purine or pyrimidine replaced by pyrimidine) is higher than transversion mutations (purine replaced by pyrimidine or vice versa). To deal with this differential mutation frequency, sophisticated statistical models have been developed by Kimura and others. For generating a DNA sequence-alignment score, the simple scoring matrix is still used, such as the NUC4.2 and NUC4.4 DNA scoring matrices. These matrices can be obtained from the NCBI (ftp://ftp.ncbi.nih.gov/blast/matrices/).

Scoring matrices for amino-acid substitutions are more complex, reflecting the similarity of physicochemical properties, as well as the likelihood of one amino acid being substituted by another at a particular position in homologous proteins. The scoring matrices for proteins are 20×20 matrices. Two well-known types of scoring matrices for proteins are PAM and BLOSUM.

6.7.1 PAM Matrices

PAM (point accepted mutation—that is, accepted point mutation—also called percent accepted mutation) matrices were first developed by Margaret Dayhoff and colleagues in 1978 and hence are also known as Dayhoff PAM matrices. A PAM represents a substitution of one amino acid by another that has been fixed by natural selection because either it does not alter the protein function or it is beneficial to the organism. *In a PAM1 matrix, which is the original PAM matrix generated, a PAM unit is an evolutionary time over which 1% of the amino acids in a sequence are expected to undergo accepted mutations, resulting in 1% sequence divergence.* Construction of a PAM1 matrix begins with alignment of the full-length sequences, reconstruction of the phylogenetic tree, and determination of the ancestral sequences for the internal nodes of the tree (see Chapter 9 for a description of the phylogenetic tree). Each computed ancestral sequence is then used to calculate the number and frequency of substitutions in the sequences along each branch arising from the node. The values in the matrix represent the probability that the amino acid in a column will be replaced by the amino acid in row in a given evolutionary time (1 PAM unit in a PAM1 matrix). From the computed probability, the percent probability can be determined. A PAM1 matrix is often displayed after multiplying each entry by 10,000.

The relationship between % amino acid substitution and the number of PAM units is not linear; thus, the above definition applies only when the divergence between two sequences is low. As the divergence increases beyond ~20%, this relationship falls apart. For example, a 100-PAM-unit divergence does not mean 100% substitution. A 100-PAM-unit divergence can be achieved by substituting ~55% of the amino acid residues, and a 200-PAM-unit divergence can be achieved by substituting ~75% of the amino acid residues. The PAM1 matrix was built by aligning closely related protein sequences (71 protein families) that had at least 85% sequence identity.

Subsequently, in order to deal with protein sequences that are more diverged and distantly related, other PAM matrices, such as PAM100 and PAM250, were generated. These later PAM matrices were generated by multiplying the PAM1 matrix by itself hundreds of times. For example, the PAM250 matrix can be obtained by multiplying the PAM1 matrix by itself 250 times over. Figure 6.9 shows the PAM250 substitution matrix. The values in the matrix are **log odds scores** (see Box 6.1).

6.7.1.1 PET91 Matrix

At the time PAM matrices were developed, the number of available protein sequences and the amount of protein family information as well as the knowledge of protein three-dimensional structure were limited. Obviously, PAM matrices could be prone to certain inherent flaws, such as (1) the assumption that each amino acid in a sequence is equally mutable, (2) multiplying a PAM1 matrix n number of times to obtain a PAMn matrix can amplify any error in the original matrix, and (3) the amino-acid-residue profiles of the proteins used to generate a PAM matrix do not

PAM250 matrix

Ala	A	2																	
Arg	R	-2	6																
Asn	N	0	0	2															
Asp	D	0	-1	2	4														
Cys	C	-2	-4	-4	-5	12													
Gln	Q	0	1	1	2	-5	4												
Glu	E	0	-1	1	3	-5	2	4											
Gly	G	1	-3	0	1	-3	-1	0	5										
His	H	-1	2	2	1	-3	3	1	-2	6									
Ile	I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5								
Leu	L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6							
Lys	K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5						
Met	M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6					
Phe	F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9				
Pro	P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6			
Ser	S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2		
Thr	T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	
Trp	W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17
Tyr	Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10
Val	V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

FIGURE 6.9 **A PAM250 substitution matrix made by writing the amino acids in alphabetical order.**

necessarily represent the residue profiles of all protein families.

Jones et al.[13] updated the PAM matrix by taking into account 2621 families of sequences (>16,000 homologous protein sequences) from the Swiss-Prot database. The sequences were clustered at 85% identity level as was done in the original PAM matrix, and the raw mutation frequency matrix was processed in a similar way as in the PAM matrix. This updated PAM matrix is called the PET91 matrix (PET91 = pair exchange table for year 1991). Thus, PET91 takes into account the substitutions that were poorly represented in the original Dayhoff matrix. The overall character of PAM and PET91 matrices is similar.

Each PAM matrix is designed to be used for comparing sequences that are evolutionarily diverged by a specific number of PAM units—that is, by a specific length of evolutionary time. The suffix (number) with PAM indicates evolutionary distance; the greater the number, the greater is the distance. For example, the PAM120 matrix is ideal for comparing sequences that have diverged by 120 PAM units during evolution. Assuming $\sim 10^7$ years (10 million years) as a PAM unit of evolutionary time, 120 PAM units of evolutionary time will correspond to 120×10^7, or 1200 million years. The higher the PAM suffix (number), the better it is in aligning more divergent sequences. PAM matrices have been developed based on the Markovian evolutionary model. The Markovian evolutionary model is the application of the Markov model to predict the probability of the state of a variable over evolutionary time, such as the probability of occurrence of an amino acid at a particular position in a protein sequence. For protein evolution, the Markov model can look at a long sequence of amino acids and analyze the likelihood that an amino acid will substituted by another. The Markov model assumes that each substitution is an independent, "memoryless" process.

6.7.2 BLOSUM

BLOSUM will be referred to as BLOSUM matrix here. BLOSUM (blocks substitution matrices) scoring matrices were proposed by Steven Henikoff and Jorja Henikoff in 1992.[14] BLOSUM represents an alternative set of scoring matrices, which are widely used in sequence-alignment algorithms. Like PAM, BLOSUM matrices are also log-odds matrices. BLOSUM matrices were developed based on multiple alignment of 500 groups of related protein sequences, which yielded > 2000 blocks of conserved amino-acid patterns. Blocks are ungapped multiple sequence alignments corresponding to the most conserved regions of the proteins involved. Henikoff and Henikoff used their BLOCKS database of trusted alignments. In each multiple alignment, the sequences showing similar % identity were clustered into groups and averaged. Using these groups, the substitution frequencies for all pairs of amino acids were calculated and the matrix was developed. Therefore, the blocks of ungapped multiple sequence alignments, which are the cornerstone of BLOSUM matrices, reveal the evolutionary relationship between proteins. The BLOCKS database was developed to host these multiple sequence alignments that reveal the blocks. By 1996, there were $\sim$3000 blocks reported, based on 770 protein families.[15] Different BLOSUM matrices differ in the % sequence identity used in clustering. Therefore, BLOSUM62

BOX 6.1

PROBABILITY, ODDS, LOG-ODDS, SCORING MATRIX

Probability is a measure of how often an event may occur, whereas **odds** is a measure based on the probability that an event may ever occur. Odds is the ratio of probabilities.

1. Probability of event X = # of events X/# of all possible events
   (e.g. when a die is rolled, the probability that the die will land with the six-side up is 1/6. In this case, the probability of the alternative event—that is, the probability *against* the die landing with the six-side up—is 5/6)
2. Odds of event X = probability of event X/probability of the alternative event (i.e. probability *against* event X)
   (e.g. in the above example, the odds of the die landing with the six-side up is the ratio of the two probabilities—that is, (1/6) ÷ (5/6) = 1/5).

In the case of amino-acid substitution (mutation), the odds of substitution means the ratio of the probability that one specific amino acid is preferentially substituted by another specific amino acid during evolution to the probability that such substitution is random. By assigning a score (odds score) to all possible pairs of amino-acid substitution, a scoring matrix can be obtained. Substitution matrices are scoring matrices that use the logarithm of the odd score, called the **log-odds score**. Use of the log-odds score instead of the odds score (which is the ratio of probabilities) allows for addition of the scores instead of multiplication of the probabilities. All algorithms for sequence comparison use some kind of scoring scheme.

If the substitution of two residues i and j is considered, the mathematical logic for the calculation of log-odds will be as follows:

1. The probability that i and j are aligned based on their evolutionary relationship of substitution is $P_e = f_i \times f_{ji}$ (f_i = frequency of residue i and f_{ji} = frequency of residue j substituting for i).
2. The probability that i and j are aligned by random chance is $P_r = f_i \times f_j$ (f_i = frequency of residue i and f_j = frequency of residue j).
3. Hence, the odds = $P_e/P_r = (f_i \times f_{ji})/(f_i \times f_j) = f_{ji}/f_j$.
4. Log odds = log (f_{ji}/f_j).
5. If (f_{ji}/f_j) = 1, then log (f_{ji}/f_j) = 0. This means that the odds of i and j being aligned based on their evolutionary relationship of substitution is the same as that by random chance.
6. If (f_{ji}/f_j) > 1, then log (f_{ji}/f_j) = positive. This means that the odds of i and j being aligned based on their evolutionary relationship of substitution is greater than by random chance.
7. If (f_{ji}/f_j) < 1, then log (f_{ji}/f_j) = negative. This means that the odds of i and j being aligned based on their evolutionary relationship of substitution is lower than even by random chance.

Therefore, a negative log-odds score means that the cost of such substitution to the protein structure and function is high, and normally such substitutions are not encouraged by natural selection. For example, the PAM250 matrix shows that the likelihood of valine being substituted by isoleucine, another hydrophobic amino acid, is higher (4) than by any one of the four hydrophilic and charged amino acids—arginine, lysine, aspartic acid, and glutamic acid (−2 for each one).

means that the sequences used to create this matrix have approximately 62% identity. Substitution frequencies weigh more heavily by protein sequences having less than 62% identity. Therefore, BLOSUM62 is useful for aligning and scoring proteins that show less than 62% identity. Shown below is an example of an ungapped multiple alignment. The conserved amino acids are shaded for identification.

```
GSFEIGNLLLII
GSFEMGNLLVIV
GSFEIGNLLLIV
GGFEIGNLLVIV
GGFEIGNLLVIV
```

Henikoff and Henikoff tested the performance of hierarchical multiple alignment of three serine proteases using BLOSUM45, BLOSUM62, BLOSUM80, PAM120, PAM160, and PAM250 matrices. All BLOSUM matrices performed better than PAM matrices; the number of residues misaligned was three to five times lower when BLOSUM matrices were used compared to PAM matrices. BLOSUM62 performed slightly better than BLOSUM45 and BLOSUM80. The reader is urged to read an excellent short primer by Sean Eddy on how the BLOSUM62 matrix was developed.[16]

BLOSUM62 matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

FIGURE 6.10 **BLOSUM62 substitution matrix made by writing the amino acids in alphabetical order.**

For a PAM matrix, the higher the suffix number, the better it is in dealing with evolutionarily distant protein alignment, and the lower the suffix number, the better it is in dealing with evolutionarily closer protein alignment. In contrast, for BLOSUM matrices, the suffix numbering system is the opposite of PAM matrices; hence, the higher the suffix number, the better it is in dealing with evolutionarily closer protein alignment. In their publication, Henikoff and Henikoff drew equivalence between different PAM and BLOSUM matrices based on relative entropy[e]. For BLOSUM matrices, relative entropy increases nearly linearly with increasing clustering percentage. Based on relative entropy, Henikoff and Henikoff concluded the following:

PAM250 ≈ BLOSUM45 (relative entropy 0.4 bit)
PAM120 ≈ BLOSUM80 (relative entropy 1 bit)
PAM160 ≈ BLOSUM62 (relative entropy 0.7 bit).

BLOSUM62 is the most widely used amino-acid scoring matrix (including by BLAST algorithms) for scoring amino-acid alignment for database searches (discussed below). Figure 6.10 shows a BLOSUM62 matrix. The NCBI FTP site from where various nucleic-acid and protein scoring matrices can be downloaded is ftp://ftp.ncbi.nih.gov/blast/matrices/.

To summarize, PAM and BLOSUM matrices can be compared as follows:

1. PAM matrices are constructed based on an evolutionary model—that is, from the estimation of mutation rates through constructing phylogenetic trees and inferring the ancestral sequence—but BLOSUM matrices are constructed based on direct observation of ungapped multiple alignment-driven sequence relationships. *Thus, PAM matrices are often used for reconstructing phylogenetic trees, whereas BLOSUM matrices are suitable for local sequence alignments*.
2. PAM matrix construction involves global alignment of the full-length sequences consisting of both conserved and diverged regions, but BLOSUM matrix construction involves local sequence alignment of conserved sequence blocks. Additionally, when Henikoff and Henikoff compared the two equivalent matrices PAM160 and BLOSUM62, they found that BLOSUM62 is less tolerant to hydrophilic-amino-acid substitution, but more tolerant to hydrophobic-amino-acid substitution than PAM160. Also, for rare amino acids, such as cysteine and tryptophan, BLOSUM62 is typically more tolerant to mismatches than PAM160.

[e]Relative entropy (also known as Kullback–Leibler divergence) is a measure of the difference between two states or two probability distributions P1 and P2. For example, P1 could be the frequency of occurrence of an amino acid at a given position in a multiple alignment relative to the background frequency, P2, of a random sample. Thus, in the context of sequence alignment, relative entropy can be calculated to determine sequence conservation relative to the background, and it is measured as the average information per residue pair in bit units. When relative entropy is 0, the target (or observed) distribution of pair frequencies is the same as the background (or expected) distribution. Relative entropy increases as two distributions become more distinguishable. An online tool for the calculation of relative entropy within sequence alignment blocks is H-BLOX, which can be accessed at http://gecco.org.chemie.uni-frankfurt.de/h-blox/hblox.html.

Most bioinformatics analysis tools provide users with a default matrix, but the default matrix may not be the most suitable matrix for the user's need. Therefore, it is important to be mindful about the utility of a specific matrix for a specific purpose. There are essentially three levels of similarity-searching alignments: that of closely related sequences, that of divergent sequences, and that of sequences intermediate between the closely related and divergent sequences. Both PAM and BLOSUM matrices can be used for this purpose. The following example shows the PAM−BLOSUM matrix equivalence, and their preferred use:

PAM100 ≈ BLOSUM90 (for less divergent proteins)
PAM120 ≈ BLOSUM80
PAM160 ≈ BLOSUM62 } (for most other proteins)
PAM200 ≈ BLOSUM52
PAM250 ≈ BLOSUM45 (for more divergent proteins)

In general, BLOSUM matrices are widely used for detecting local alignments. BLOSUM62 is the most frequently used matrix for detecting the majority of weak protein similarities, and BLOSUM45 is very suitable for detecting long and weak alignments.

While aligning unknown sequences, if one wants to use the most appropriate matrix based on how similar the sequences are, one has to first try multiple matrices and then use the one that gives the highest ungapped alignment score.

6.7.3 Scoring Sequence Alignment and Statistical Significance of Sequence Alignment

The calculation of alignment scores involves addition of the match/mismatch values from the matrix for every nucleotide base or amino acid residue involved in the alignment to obtain a gross alignment score. Then the total gap penalty is calculated. The total gap penalty value is subtracted from the gross alignment score value to obtain the final alignment score. The terminal gaps may or may not be penalized, depending on the program used. For example, in local alignment (Smith−Waterman algorithm), a terminal gap penalty does not make sense, whereas in global alignment (Needleman−Wunsch algorithm), a terminal gap penalty may be applied depending on the program.

Different alignments should not be directly compared based on their raw score (S). For example, a not-so-good long alignment may get a higher S than a very good short alignment. Thus, different alignments should only be compared after normalization. This is achieved by determining the statistical significance of the score.

The statistical significance of the raw score, S, of an alignment is assessed to determine whether the observed alignment is specific or could be the result of random chance. This is done by creating many random sequences of the same length from one of the two aligned sequences by shuffling the sequence and running the alignment again. Typically this reshuffling and realignment process is repeated 200 times or more. Each alignment using these random sequences produces an alignment score (s). These scores ($s_1 \ldots s_n$) are plotted to generate a distribution pattern, a threshold of significance is set, and the original score (S) is compared against this distribution. If the S is located at one end of the distribution (extreme value distribution) that means that the alignment is not likely to be produced by random chance.

6.7.3.1 P-*Value*

The ***P*-value** of an alignment represents the **probability** of obtaining a score $\geq S$ by chance. For example, if the P-value is 10^{-5}, it means that the probability of obtaining an alignment with a score $\geq S$ is 1 out of 10^5. Thus, different alignments can be compared based on their P-values. The P-value ranges from 0 to 1; the closer it is to 0, the better is the alignment.

6.7.3.2 Z-*Score*

In the statistical sense, Z is the distance between S and the mean of scores obtained using randomized sequences. The Z-score is calculated by repeating the reshuffling and realignment process, as described above, and noting the raw score (s) of each alignment using the randomized sequences ($s_1 \ldots s_n$). The mean ($\bar{x}$) and the standard deviation (σ) of $s_1 \ldots s_n$ are calculated and from these the Z-score of the target alignment can be determined.

The calculation of the Z-score assumes that the alignment of the shuffled random sequences shows a normal distribution. Hence, the farther the alignment raw score S is away from the $\bar{x}$ of $s_1 \ldots s_n$, the more likely it is to be significant. In a statistical sense, the Z-score reflects the extent to which S is an outlier from the population. A $Z = 5$ means the S is 5σ above the $\bar{x}$ of $s_1 \ldots s_n$. By convention, a $Z > 7$ indicates a significant alignment and it is likely that the two sequences being aligned are homologs; it also indicates that the alignment of the two sequences likely reflects the alignment of structurally and functionally related amino acid residues of the proteins. Another interpretation of the Z-score is as follows[17]:

$Z > 20$: two sequences are definitely homologous (Family)
Z between 10 and 20: two sequences most likely homologous (Family/Superfamily)
Z between 6 and 8: two sequences are less likely to be homologous
$Z < 6$: not significant.

PRSS (current version PRSS3; http://www.ch.embnet.org/software/PRSS_form.html)[18] is freely available web-based software that can be used to evaluate the significance of a protein or DNA sequence-similarity score. PRSS compares two sequences and calculates the optimal similarity scores, and then repeatedly shuffles the second sequence, and calculates optimal similarity scores using the Smith–Waterman algorithm. An extreme value distribution (EVD) is then fit to the shuffled-sequence scores. In the PRSS output, the left-most column represents the normalized similarity scores; and the E ( ) column on the right represents the number of sequences expected to achieve the score in the first column.

6.7.3.3 E-*Value*

This is particularly relevant in relation to sequence-similarity searching using BLAST and FASTA, which are discussed later in this chapter. The ***E*-value** is the **expectation** value that indicates the number of alignments with a score $\geq S$ that one can expect to find by chance in a database of size N. Hence, the E-value is dependent on the database size and the query length. The closer the E-value to 0, the better is the alignment. For $E < 1e-2$ $(=1\times10^{-2}=0.01)$, $P \approx E$. *The* E*-value is the most widely used measure for estimating the quality of sequence alignment—that is, the extent of sequence similarity*.

The typical threshold for the E-value when judging homology, particularly using BLAST, is $E \leq 1e-5$ $(=1\times10^{-5})$, and the lower the value, the better it is. For BLAST (both nucleotide and protein), the default E-value is set at 10 in the **Expect threshold** box under **Algorithm parameters** (lower left corner of the BLAST home page). This means that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990).[19] It also means that the BLAST output will not report any alignment with an E-value greater than 10. Obviously, when the E-value is increased from the default value of 10, a larger number of chance matches will be reported. In contrast, lowering the default value makes the search more stringent and fewer chance matches are reported. The default E-value should be increased if searching for short sequence matches, because setting a lower E-value will automatically exclude the short matches as spurious and these will not be reported. In such cases, the default value in the "Expect threshold" box can be manually changed. *Alternatively, the nucleotide and protein BLAST programs of the NCBI automatically adjust the* E*-value* if the query, either nucleotide or amino acid, is of length 30 or less.

6.7.3.4 *Bit Score*

The **bit score** (S') is a normalized raw score expressed in *bits*; it is an estimate of the search space one has to search through—that is, the number of sequence pairs one has to score—before one can come across a raw alignment score $\geq S$, by chance.

For example, a bit score of 30 means that, on average, one has to score 2^{30} $(=1$ billion) sequence pairs before one will come across a score $\geq S$, by chance. Usually, good alignments produce a bit score > 50. *It should be emphasized that the bit score is dependent on sequence length, and short sequences may not produce high bit scores despite very high identity*.

To summarize the utility of the statistical estimates of sequence alignment in simple terms, **the better the alignment (e.g. homologous sequences), the lower the *P*- and *E*-values, and the higher the *Z*- and bit scores.**

6.8 DATABASE SEARCHING WITH THE HEURISTIC VERSIONS OF THE SMITH–WATERMAN ALGORITHM—BLAST AND FASTA

Alignment programs that use dynamic programming algorithms, such as the Needleman–Wunsch and Smith–Waterman algorithms, require long processing times, particularly when searching a huge database. In order to circumvent this computational limitation, heuristic methods have been developed. A **heuristic** method (algorithm) estimates the best solution without considering every possible outcome; thus, a heuristic method does not guarantee to find the best solution, but finds good solutions, and thereby has high speed and is time efficient. Two examples of heuristic methods are the Basic Local Alignment Search Tool (BLAST) and FAST-All (FASTA). FASTA is pronounced "fast A". It stands for "FAST-All" because it is an extension of "FAST-P" for proteins and "FAST-N" for nucleotides; therefore, FASTA works with all alphabets associated with proteins and nucleic acids.

6.8.1 BLAST and its Utility

Currently, the most widely used heuristic algorithm is BLAST, developed by Altschul and colleagues.[20] The BLAST algorithm allows a DNA or protein **query** sequence to be compared with sequences in the database. The main idea behind BLAST searching is that homologous sequences are likely to contain a short, high-scoring similarity region, called a **word** or **hit** (**W**). Each word (hit) gives a **seed** that triggers

the alignment and BLAST tries to extend on both sides of the seed. The word size—i.e. the length of the seed—may vary. For nucleotides (**blastn**), the default word size is 11 and the smallest word size is 7; for proteins (**blastp**), the default word size is 3 and the smallest word size is 2. For **megablast** (highly similar sequences), the default word size is 28 and the smallest word size is 16 for nucleotides. These parameters can be adjusted by clicking "**Algorithm parameters**" in the lower left corner of the BLAST page. For a nucleic-acid sequence alignment, the seed should match completely in order to trigger the alignment; for proteins, the match may or may not be exact. In order to create an alignment, the BLAST algorithm breaks the query sequence into short subsequences. Typically, BLAST is designed to find local regions of similarity, but can be expected to run about two orders of magnitude faster than the Smith–Waterman algorithm. An important parameter governing the sensitivity of BLAST searches is the length of the initial words (hits).

Database searching is done for various reasons, such as finding relationships between the query sequence and other sequences in the databases, understanding the likely function of a sequence, identifying regulatory elements, understanding genome evolution, or assisting in sequence assembly. In designing probes and primers, the selected nucleic acid sequence is compared with other sequences in the database to determine the specificity and uniqueness of the selected sequence. Therefore, a BLAST search can help determine the identity of nucleic acid and protein sequences, reveal whether these sequences represent new genes and proteins, discover variants of existing genes and proteins, discover potential orthologs and paralogs of a sequence, determine whether a gene or protein is present in other organisms, or determine whether a nucleic acid sequence is expressed.

In a BLAST search, the sequence that is subject to comparison is termed the **query**. This query sequence is subjected to BLAST search against all sequences in the database. The search retrieves all sequences showing similarity with the query sequence. These sequences are called **subject** (or **target**).

6.8.2 Various BLAST Programs for Analysis

At the NCBI, there are several BLAST resources, which can be grouped as **basic BLAST** and **specialized BLAST**.

Basic BLAST offers a few options, such as **blastn** (searches a nucleotide database using a nucleotide query), **blastp** (searches a protein database using a protein query), **blastx** (searches a protein database using a translated nucleotide query), **tblastn** (searches a translated nucleotide database using a protein query), and **tblastx** (searches a translated nucleotide database using a translated nucleotide query).

Specialized BLAST provides many specialized/advanced options, such as Primer-BLAST, trace archives, conserved domains, conserved domain architecture, gene expression profile (GEO), immunoglobulin search (IgBLAST), single nucleotide polymorphism (SNP) flank search, vector contamination screening (vecscreen), Align, PubChem BioAssay search, searching SRA transcript and genomic libraries, Multiple Alignment Tool, Global Sequence Alignment Tool, or searching the RefSeqGene database.

For a detailed description of each of these different BLAST programs and their use, refer to the NCBI reference resource (http://blast.ncbi.nlm.nih.gov/).

6.8.2.1 Megablast, Blastn, and Discontinuous Megablast

Currently, the nucleotide BLAST program offers three options for searching sequences for hits in the database with different degrees of similarity. These are megablast, blastn, and discontinuous megablast.

Megablast is optimized for highly similar sequences. It efficiently finds long alignments between highly similar (>95%) sequences, and thus is the best tool to find the identical match to the query sequence. The default word size is 28 and the lowest word size is 16.

Blastn is optimized for somewhat similar sequences. The reason blastn is more sensitive than megablast is because it uses a shorter default word size (11). Because of this, blastn is better than megablast at finding alignments to related nucleotide sequences from other organisms. Reducing the word size from 11 (default) to 7 (lowest) increases the **sensitivity** of search—that is, increases the number of positive hits.

Discontinuous megablast is optimized for more dissimilar sequences. Instead of using the exact word match as seed for an alignment extension, discontinuous megablast uses a noncontiguous word within a longer window of template. As a result, discontinuous megablast using the same size of the initial hit is even more sensitive and efficient than standard blastn using the same word size.

6.8.2.2 Searching for Short, Nearly Exact Matches

For searching short nucleotide-sequence matches, algorithm parameters can be manually adjusted as follows: select blastn→select the non-redundant (nr) nucleotide database (unless a specific database is needed)→select "Somewhat similar sequences (blastn)"→click on

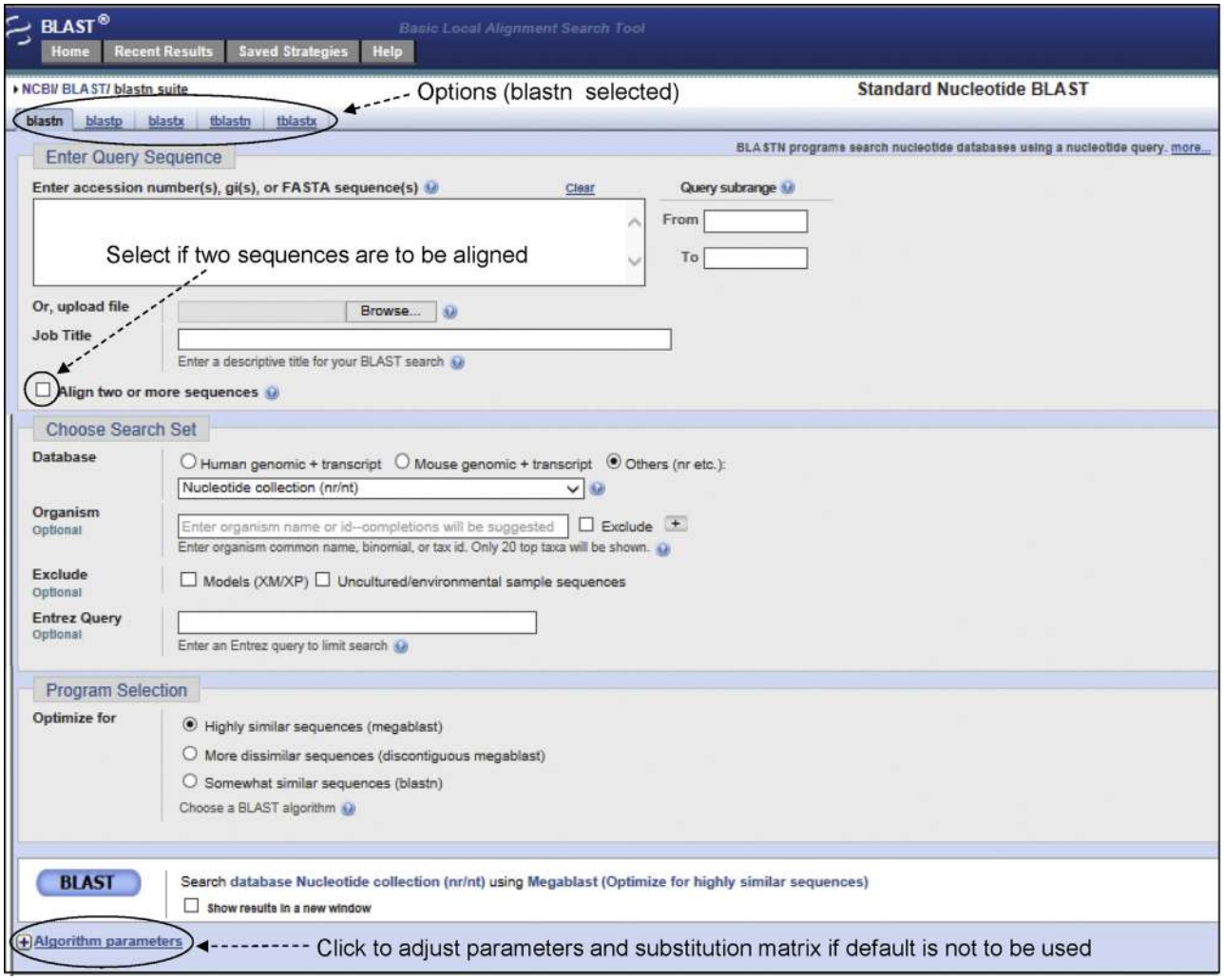


FIGURE 6.11 **NCBI BLAST home page of nucleotide blast.** By clicking the tabs at the top (circled), other BLAST tools can be obtained. For regular BLAST, the sequence can be entered in plain text format. For pairwise alignment, the small box (indicated by an arrow) can be checked and a second box appears where the other sequence can be entered. The "Algorithm parameters" can be clicked and the default setting can be changed.

"Algorithm parameters"→check the short queries box→filter[f] setting to remain off→select the word size 7→change expect threshold to 1000 (or as necessary). For searching short protein-sequence matches, algorithm parameters can be manually adjusted as follows: select blastp→select the non-redundant (nr) protein database (unless a specific database is needed)→check the short queries box→filter setting to remain off→select the word size 2→change expect threshold to 10000 (or as necessary)→select PAM30 as the scoring matrix. The query needs to be at least twice the word size. Theoretically therefore, a query of four amino acid residues should be searchable, but at least five residues are recommended.[21] Figure 6.11 shows a partial screenshot of the BLAST home page. *Alternatively, the nucleotide and protein BLAST programs of NCBI automatically adjust the E-value if the query, either nucleotide or amino acid, is of length 30 or less.*

[f]Because sequence-similarity searching aims to detect sequences that indicate structural and/or functional similarity, a sequence filter is used to remove low-complexity regions during similarity searching. Examples of low-complexity regions are repeat sequences (e.g. polyA tails, nucleotide sequences like AAAATTAAAAAT, proline-rich regions, amino-acid sequences like GGGGKDKKKKDD), compositionally biased sequences etc. that are naturally abundant in most sequences. If low-complexity regions are not removed, then the sequence alignment may produce artificially high scores that would not be a true reflection of homology. Blastn filters low-complexity nucleotide sequences with the DUST algorithm, and blastp filters low-complexity amino-acid sequences with the SEG or XNU algorithms. Low-complexity nucleotide sequence is substituted by "N" (e.g. NNNNNNN), whereas low-complexity amino-acid sequence is substituted by "X" (e.g. XXXXXXX), and removed from the search.

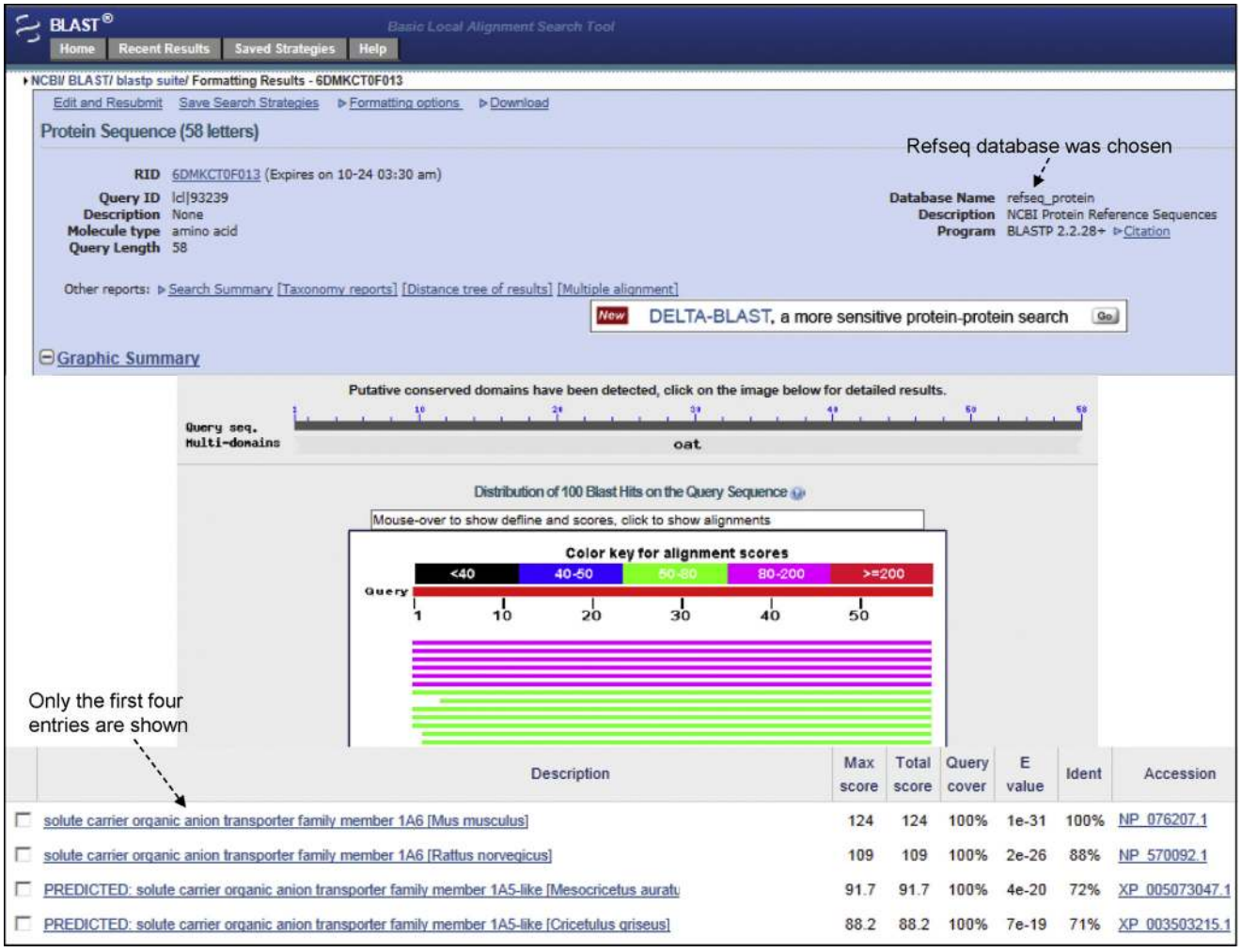


FIGURE 6.12 **Result of the BLAST analysis of Slco1a6.** The screenshot was captured in three different pieces (the upper, middle and lower segments), which are put together in the figure. A 58-amino-acid segment was used for BLAST (blastp). The RefSeq protein database was chosen to minimize the number of redundant hits. Alternatively, the Swiss-Prot could be chosen to obtain non-repetitive specific hits. The result shows on the top that putative conserved domains have been detected. These are the Kazal domain and the MFS domain. Refer to Chapter 8 for a more detailed discussion on this topic. From the analysis, only the first four entries are shown. From the BLAST hit diagram, a specific line can be clicked to get to the alignment. The color key for alignment score is self explanatory.

6.8.2.3 *Suggested BLAST* E-*Value Cut-Off*

For nucleic-acid-based search, the suggested threshold (minimum significant hit) for the *E*-value is ≤ 1e-6 ($=10^{-6}$), and a sequence identity of ≥ 70%. For protein-based search, the suggested threshold for the *E*-value ≤ 1e-4 ($=10^{-4}$), with a sequence identity of ≥ 35%[g]. However, typically for protein-based homology search, the threshold used is $E \leq$ 1e-5 ($=10^{-5}$), and the lower it is, the better. For example, an *E*-value of 1e-25 ($=10^{-25}$) will indicate a clear homology.

It should be borne in mind that the E-value is influenced by the query length. A moderately good alignment involving two very long sequences will produce a higher E-value than an extremely good alignment involving two smaller sequences.

6.8.3 Typical Basic BLAST Output

Figure 6.12 shows the result of a BLAST search. A 58-amino-acid segment was searched in the NCBI database using BLAST. In order to tailor the search to

[g]It has been reported that protein pairs with similar structure and function are likely to have > 35% sequence identity[22]. The author analyzed more than a million sequence alignments between protein pairs of known structures and noted that sequence alignments could unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity was > 40% for long alignments. The signal, however, became blurred when the sequence identity was between 20 and 35%; this 20–35% range was termed the **twilight zone** of sequence identity.

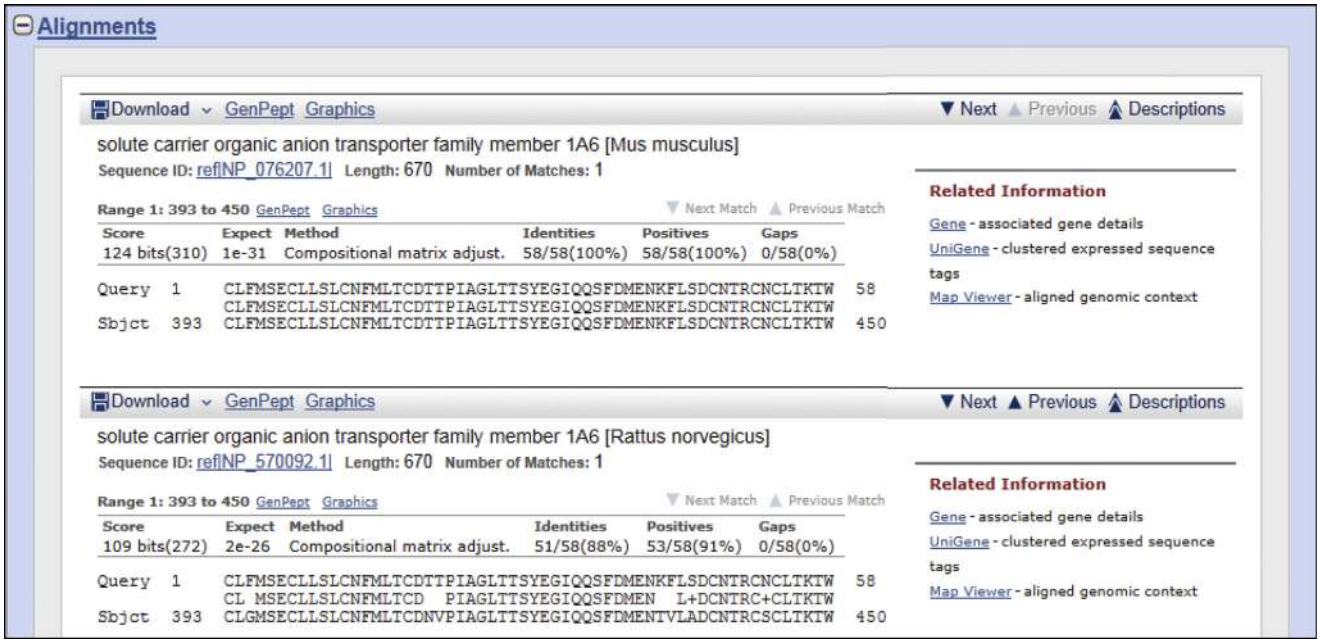


FIGURE 6.13 **The details of two alignments from Figure 6.12.** In the alignment, the upper sequence is the **query** sequence (the sequence submitted for search) and the lower sequence is the **subject** sequence (from the database); the identities and the similarities are in the middle. The number of amino acids showing identity/similarity is indicated; **identities** indicate identical amino acids between the query and subject sequences whereas **positives** indicate identical amino acids plus similar amino acids at the corresponding positions. Similar substitutions are indicated by a + sign. Each individual alignment also provides direct link to the original sequence in the database. If the subject sequence is from an organism whose whole genome is known and sequenced, the alignment also provides links to the Gene and Map Viewer databases, indicated on the right-hand side.

reduce the amount of less relevant output, the organism (*Mus musculus*) and the database (RefSeq protein database) were chosen on the BLAST home page. The search returns many entries; the highest similarity was (predictably) with mouse Slco1b2 protein (Refseq ID NP_065241). In the output, the subject sequences are listed from the highest similarity at the top to progressively lower similarities going down the list, as depicted by the bit score (score) and the *E*-value. The bit scores are listed from the highest value at the top to progressively lower values going down the list, whereas the *E*-values are listed from lowest value at the top to increasingly higher values going down the list. The detailed alignments are shown in Figure 6.13.

6.8.3.1 Searching for Distantly Related Proteins—PSI-BLAST

Many homologous proteins have similar three-dimensional structure, but in pairwise alignment they may not show significant sequence similarity. Therefore, regular protein BLAST (blastp) is not useful in identifying these proteins. Position-Specific Iterative BLAST (PSI-BLAST) is designed to detect weak relationships between the query sequence and other sequences in the database that are not necessarily detectable by standard BLAST searches. When a new genome is sequenced, PSI-BLAST can be used to identify the homology of the predicted protein products. The procedure of PSI-BLAST involves the following steps:

First step in PSI-BLAST involves standard protein–protein BLAST using the default substitution matrix, such as BLOSUM62. The input protein sequence is compared to proteins in the database to generate similarity hits. The high-scoring hits (default threshold *E*-value = < 0.005) are used to generate a multiple alignment. The original query sequence serves as the template to drive the multiple alignment. PSI-BLAST analyzes the alignments position by position and assigns a score to every position. If the amino acid residue is highly conserved at a particular position, that residue is assigned a high positive score, and others are assigned high negative scores. At weakly conserved positions, all residues receive scores near zero. Using these scores, a **profile** or **position-specific scoring matrix** (**PSSM**) is built. In the next iteration of BLAST search, this PSSM replaces the substitution matrix used in the previous iteration of BLAST search; thus more proteins are identified using this PSSM. The newly identified proteins are then incorporated in the multiple alignment to create a new PSSM, which replaces the previous one. This process is repeated (iterative) until no new

proteins are found. In each repetition, a new PSSM is generated, which replaces the old one and is used for the new round of search. The PSI-BLAST output looks like regular BLAST output.

Because of the nature of the algorithm, the main source of error in PSI-BLAST is the corruption of the profile (PSSM). In other words, for reasons unrelated to true homology/functional characteristics (e.g. amino-acid compositional bias), a position-specific amino acid may be wrongly identified as a conserved residue and assigned a high score. That position in the profile will then adversely influence the next iteration to identify more related proteins. Repeated iteration will amplify the error corrupting the subsequent profiles. There are several ways to address this problem, such as filtering out compositionally biased regions using a filtering algorithm, lowering the *E*-value from the default 0.005, or visually inspecting each output and applying judgment to discard the hits that appear spurious.

6.8.3.2 Searching for Pattern Hit—PHI-BLAST

Many proteins contain signature sequences (motifs) that are characteristics of a protein family. These signature sequences are part of important structural or functional domains. Pattern-hit-initiated (PHI)-BLAST is designed to search the database for proteins that are significantly related to the query sequence and also contain a pattern. In other words, PHI-BLAST searches for significantly similar sequences to both a query sequence and a signature. This dual requirement is supposed to reduce the number of database hits that contain the pattern but are likely to have no true homology to the query.

6.8.4 BLAT

Blast-like alignment tool (BLAT) has been discussed in the context of the University of California Santa Cruz (UCSC) Genome browser in Chapter 5. Also refer to Figure 5.32 for BLAT output. Therefore, the discussion here will be brief. BLAT is an alignment tool like BLAST, but it is structured differently. BLAT is commonly used to map the location of a query sequence in the genome, or to determine the exon structure of an mRNA. DNA BLAT works well within humans and primates, while protein BLAT works well for terrestrial vertebrates and even earlier organisms for conserved proteins.

6.8.5 FASTA

FASTA was developed for rapid biological-sequence comparison.[23] It was derived as a more sensitive and versatile program from its predecessor program FASTP, which was developed by the same authors 3 years earlier for rapid protein-sequence comparison. Like BLAST, FASTA also allows the user to compare a DNA or protein query sequence against a large database. FASTA searches for matching sequence patterns called ***k*-tuples** (***ktup***), which are akin to the "words" (W) in BLAST. The *ktup* length is usually user defined (e.g. defining *ktup* = 6 for a search involving DNA sequence will prompt the algorithm to use 6 nucleotides as the matching sequence pattern for the search). The FASTA search strategy involves searching for words of length *ktup* common to the query and target sequences. Using *ktup*, FASTA builds a local alignment. Finally, FASTA scores this alignment and provides the output as a list of sequences similar to the query in descending order. *The default ktup is 2 for amino acids and 6 for nucleotides; hence, the default window size in FASTA is smaller than that in BLAST.*

Some web-based FASTA servers are provided in Table 6.3.

TABLE 6.3 Web-Based FASTA Servers

FASTA Server	URL
GenomeNet, Japan	http://www.genome.jp/tools/fasta/
EMBL-EBI	http://www.ebi.ac.uk/Tools/sss/fasta/
University of Virginia	http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

6.8.5.1 Comparison of BLAST and FASTA

BLAST and FASTA are both heuristic algorithms that perform database searches to find sequences related to a query sequence. However, there are some differences between the two:

1. BLAST begins a search by looking for matches that include exact matches and conservative substitutions; FASTA begins a search by looking at exact matches.
2. BLAST scans a larger window size than FASTA; hence, FASTA may produce better coverage for homologs.
3. BLAST may produce multiple best-scoring alignments (also called **high-scoring segment pairs** or **HSPs**) from the same sequence; FASTA returns only one alignment from one sequence.
4. BLAST automatically masks low-complexity regions; FASTA does not employ such automatic masking. Therefore, if the query sequence has non-unique segments, such as repeats, compositionally biased segments, etc., FASTA search may return alignments with artificially high scores.
5. For a given sequence search, the BLAST output is larger than that of FASTA.
6. For a given sequence search, BLAST is faster than FASTA.

6.9 SEQUENCE COMPARISON, SYNTENY, AND MOLECULAR EVOLUTION

Comparative genomics is the study of the evolutionary relationships between the genes and genomes of different species. Comparative genomic studies are helpful in elucidating the structure, function, and evolution of genomic elements and sequence features that influence various aspects of genome biology. From the macro to the micro scale, the similarity between two genomic sequences can be studied at the level of the whole genome, at the level of chromosomal segments, and also at the level of specific genomic markers. This is because the genomes of the descendants of a common ancestor are likely to preserve at least some of the same genes in the same order. A chromosomal segment that has been inherited from the common ancestor during evolution without a major rearrangement of the order of genes is called a **syntenic block** (or **synteny block**). Syntenic blocks contain specific non-repetitive genomic markers that are in the same order and orientation in the genomes being compared. These genomic markers could be protein-coding genes, RNA-coding genes, noncoding sequences, pseudogenes, etc., and are called **syntenic anchors** (or **synteny anchors**).[24] In other words, syntenic blocks are composed of syntenic anchors present in consecutive order. Genes within a syntenic block are likely to be orthologous. While comparing two genomes, the overall sequence similarity can be enhanced if the genomes are segmented into syntenic blocks. For example, approximately 40% of the human genome can be aligned with the mouse genome, but over 90% of mouse and human genomes can be segmented into blocks of conserved synteny. Comparison of mouse chromosome 16 with the human genome shows regions of conserved synteny with human chromosomes 3, 8, 12, 16, 21, and 22. A total of 11,822 syntenic anchors map to chromosome 16; the mean length and identity of these anchors are 198 bp and 88.1%, respectively. Over 50% of these anchors are in runs of at least 128 in a row in the same order and orientation between mouse chromosome 16 and the human chromosomes sharing blocks of conserved synteny.[24] Charting the blocks of conserved synteny creates a **synteny map**, which shows the large-scale evolutionary relationships between genomes that are related through a common ancestor, but have diverged during evolution. Shared genomic synteny and shared protein functions can be used to enhance the identification of orthologous gene pairs.[25]

References

1. Nei M, et al. *Proc. Natl. Acad Sci USA* 2001;**98**:2497–502.
2. Koonin EV, Galperin MY. *Sequence-evolution-function: computational approaches in comparative genomics*. Boston, MA: Kluwer; 2003. p. 25–49
3. Raghava GPS, Barton GJ. *BMC Bioinformatics* 2006;**7**:415.
4. Choudhuri S, et al. *Biochem Biophys Res Commun* 2000;**274**:79–86.
5. Wu CH, et al. *Nucl. Acids Res* 2002;**30**:35–7.
6. Huang X, Miller W. *Adv Appl Math* 1991;**12**:337–57.
7. Papadopoulos JS, Agarwala R. *Bioinformatics* 2007;**23**:1073–9.
8. Larkin MA, et al. *Bioinformatics* 2007;**23**:2947–8.
9. Thompson JD, et al. *Nucl Acids Res* 1994;**22**:4673–80.
10. Katoh K, et al. *Nucl Acids Res* 2002;**30**:3059–66.
11. Edgar RC. *Nucl Acids Res* 2004;**32**:1792–7.
12. Notredame C, et al. *J Mol Biol* 2000;**302**:205–17.
13. Jones DT, et al. *Comput Appl Biosci* 1992;**8**:275–82.
14. Henikoff S, Henikoff JG. *Proc Natl Acad Sci USA* 1992;**89**:10915–9.
15. Pietrokovski S, et al. *Nucl Acids Res* 1996;**24**:197–200.
16. Eddy SR. *Nat Biotechnol* 2004;**22**:1035–6.
17. Kim D, et al. *Protein Eng* 2003;**16**:641–50.
18. Pearson WR. *Meth Enzymol* 1996;**266**:227–58.
19. Karlin S, Altschul SF. *Proc Natl Acad Sci USA* 1990;**87**:2264–8.
20. Altschul SF, et al. *J Mol Biol* 1990;**215**:403–10.
21. NCBI. *NCBI BLAST Help*. Available online at: <http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD = Web&PAGE_TYPE = BlastDocs&DOC_TYPE = ProgSelectionGuide>; 2013.
22. Rost B. *Protein Eng* 1999;**12**:85–94.
23. Pearson WR, Lipman DJ. *Proc Natl Acad Sci USA* 1988;**85**:2444–8.
24. Mural, et al. *Science* 2002;**296**:1661–71.
25. Zheng, et al. *Bioinformatics* 2005;**21**:703–10.

CHAPTER

7

Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences*

OUTLINE

7.1 GENOME SEQUENCING

The traditional sequencing method involves the following steps: the DNA fragment to be sequenced is cloned into a vector that provides known primer-binding sites flanking the cloned sequence. The first set of sequencing primers is designed based on these known primer-binding sites. The sequencing runs on both strands produce two sequencing reads. New primers are designed from the 3′-end of the newly obtained sequences (Figure 7.1A). In this process, the sequence reads generated in one direction have sequence overlaps. Using the sequence overlaps, these contiguous sequence reads are assembled into a larger sequence, called a **contig**[a] (from contiguous) (Figure 7.1B; upper and lower panels). The sequencing method described above involves sequential designing of primers followed by new sequencing; hence, this sequencing method is called **primer walking**. Primer walking works well for sequencing a complementary DNA (cDNA) or a large DNA fragment of finite size. However, primer walking is costly and slow, and it involves cloning of the fragment. Although it can be scaled up, primer walking is still not a high-throughput strategy for sequencing a genome.

Primer walking is an example of **directed sequencing** because the primer is designed from a known region of DNA to guide the sequencing in a specific direction. In contrast to directed sequencing, **shotgun sequencing** of DNA is a more rapid sequencing strategy. As the name suggests, shotgun sequencing involves random fragmentation of the DNA into small pieces followed by sequencing of these small fragments. Shotgun sequencing can adopt either a **hierarchical shotgun sequencing** (top-down) approach, or a **whole-genome shotgun** (**WGS**) **sequencing** (bottom-up) approach. In the hierarchical shotgun sequencing approach, the chromosomes are sorted, broken into

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

[a]A sequence read should not be confused with a sequence contig. In theory, at least two overlapping sequence reads are needed to construct one sequence contig. In reality, a sequence contig is constructed from many sequence reads.

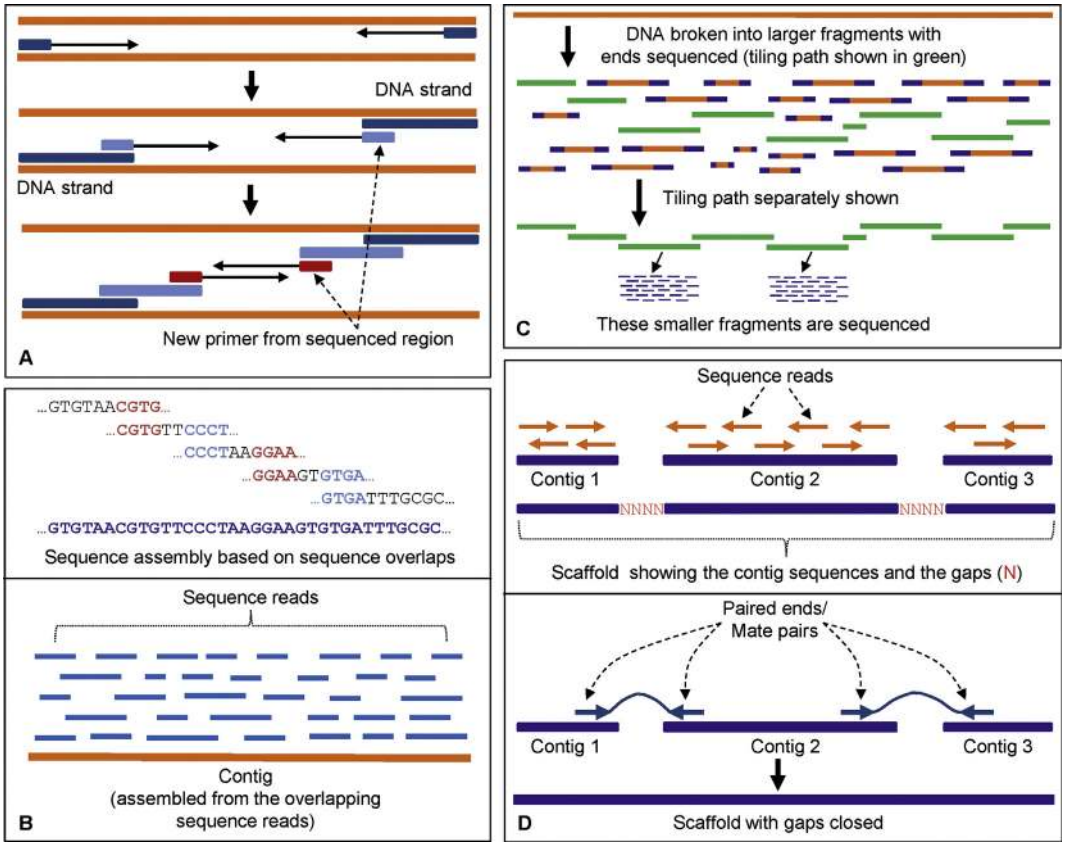


FIGURE 7.1 **Sequencing strategy.** (A) Directed DNA sequencing by primer walking. This involves sequential designing of primers from a known region. The first set of sequencing primers are designed based on the primer-binding sites flanking the cloned DNA. New primers are designed from the 3′-end of the newly obtained sequences. (B) The sequence reads have sequence overlaps that help put the contiguous sequences together in proper order (upper panel). Many such sequence reads are assembled to obtain a sequence contig (lower panel). (C) In the hierarchical shotgun sequencing approach, the chromosomes are sorted and broken into large fragments. Both ends of each clone are sequenced and the tiling path is determined based on sequence overlaps. The tiling path (shown as green fragments) is the smallest set of overlapping clones that covers the entire chromosome or contig. Once the clones in the tiling path are identified, the larger fragments in these clones are broken down into smaller fragments, which are then sequenced using a shotgun sequencing strategy. The sequence is put together by a sequence assembler. (D) A scaffold, or supercontig, is a portion of the chromosome (or genome) sequence that is composed of contigs put together in correct order. Scaffolds have gaps (upper panel); once the gaps are identified, the goal becomes sequencing those regions and closing the gaps. The lower panel shows that the scaffold of these three contigs is held together by mate pairs. The thin lines connect the paired ends.

large fragments and cloned into vectors that can hold large DNA fragments, such as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs)[b]. Both ends of each clone are sequenced, producing an approximately 500−800-bp read each, together called **paired ends** or **mate pairs**, and the tiling path is determined based on sequence overlaps. This is part of the physical mapping process[c]. The **tiling path** is the smallest set of overlapping clones (i.e. clones with overlapping DNA fragments) that covers the entire chromosome or contig (Figure 7.1C). Therefore, the clones that produce the tiling path

[b]BACs can hold DNA fragments up to 300 kbp, whereas YACs can hold fragments up to 3000 kbp.

[c]A physical map of a chromosome is a set of cloned DNA fragments whose position relative to each other in the chromosome is known. In physical mapping, a large number of clones from the recombinant library of each chromosome are end sequenced to obtain a fingerprint for each clone. A fingerprint is a unique sequence signature that identifies a specific clone. The information about such signatures can be obtained by random sequencing or by examining sequence information already existing in the database. For example, the sequence of a known unique gene in the chromosome will provide the fingerprint for a clone that contains this sequence. This type of short DNA sequence (usually less than 500 bp) that occurs only once in the chromosome (or genome) is known as a sequence tagged site (STS). Appropriate overlaps between clones are determined based on such clone-specific fingerprints. Fingerprinting the clone contigs generates many genomic landmarks along the length of the chromosome. These landmarks help in the process of accurate sequence assembly, particularly if the genome is rich in repetitive sequences.

constitute a set of **clone contigs** (contiguous clones). Once the clones in the tiling path are identified, the larger fragments in these clones are broken down into smaller fragments, which are then sequenced using a shotgun sequencing strategy. The sequence is put together by a sequence assembler. During assembly, the contigs are assembled in correct order to produce longer **supercontigs**, also called **scaffolds**. Scaffolds usually have gaps (Figure 7.1D; upper panel). Once the gaps are identified, special care is taken to sequence the gapped regions; this is part of the finishing process for genome sequencing and assembly (Figure 7.1D; lower panel).

In the bottom-up WGS sequencing approach, the DNA is randomly sheared into small pieces, fragments are size selected and subcloned into a "universal" cloning vector containing "universal" priming sites. Clones are sequenced. Numerous sequence reads are generated from numerous small fragments. The sequence is put together by a sequence assembler with very high computing capacity. In 1988, Eric Lander and Michael Waterman published a paper in which they demonstrated mathematically that at least 8–10-fold sequencing coverage is needed for the successful assembly of most of the genome, assuming an even distribution of sequence reads.[1]

Both hierarchical shotgun sequencing and WGS sequencing have advantages and disadvantages. Hierarchical shotgun sequencing creates a physical map of the genome; hence, it produces genomic landmarks that can be helpful in sequence assembly if the genome is rich in repetitive sequences (like the human genome). However, hierarchical sequencing is slow because it proceeds through many steps. The WGS sequencing approach is rapid and direct, but the assembly of sequences may run into problems if the genome is rich in repetitive sequences. The number of sequencing reads generated in WGS sequencing is very high; therefore, the computing power needed for WGS sequence assembly is very high. Currently, the computing power is less of an issue, but it was an issue in early days of genome sequencing. Current genome-sequencing efforts adopt a combination of both strategies for speed and accuracy. Use of the next-generation (next-gen) sequencing technique has further added to the speed because it does not need cloning of the fragments.

7.2 SEQUENCE ASSEMBLY

Genome assembly from sequence reads is an algorithm-driven automated process. DNA-sequence-assembly programs have utilized sequence overlaps for sequence assembly in correct order. The computational aspect of assembly algorithms is beyond the scope of this book. Nevertheless, a few terms will be discussed in plain language for the sake of familiarity. Sequence assembly can be done using one of three approaches: (1) **greedy**, (2) **overlap-layout-consensus (OLC) and Hamiltonian path**, and (3) **de Bruijn graph and Eulerian path**[d].

Greedy is a rapid-assembly algorithm, which joins together the sequence reads that are the most similar to each other based on as much sequence overlap as possible. In doing so, the greedy algorithm first compares all fragments in a pairwise fashion to identify sequences that have overlaps; next, the sequences that have the best overlaps are merged; this merging process continues (iterative process) until all the sequences with overlaps have been merged. In this process, some reads may not be assembled, which are shown as gaps. Paired-end sequencing is used to close the gaps. Many early assemblers were based on the greedy algorithm and were extremely useful, such as **Phrap, TIGR assembler**, and **CAP**. The **Phred–Phrap–Consed** suite of programs has been widely used. Phred and Phrap were developed by Drs Phil Green and Brent Ewing at the University of Washington, Seattle, in 1998 for the Human Genome Sequencing project. Phred is base-calling software that assigns a quality score to each base called. Phrap is de novo shotgun sequence-assembly software. Consed is the sequence-assembly editor companion to Phrap, and it is a tool for viewing, editing, and finishing sequence assemblies created with Phrap. Many such assembly suites also include sequence-alignment tools.

The overlap-layout-consensus (OLC) algorithm is based on all pairwise comparisons, and it generates a directed graph using reads and overlaps[e]. In the graph, each sequence is created as a node and an edge is created between any two nodes whose sequences overlap. The algorithm then tries to find the Hamiltonian traversal path of the graph, which contains all the nodes (sequences) exactly once, and combines the overlapping sequences in the nodes into the sequence of the genome. Some assemblers that utilize

[d]If the reader is interested to learn more about the computational aspects behind the key methods in simple terms, a good source to consult is *Bioinformatics for Biologists*.[2]

[e]A graph is represented by a set of nodes (vertices) and a set of edges (arcs) between the nodes; hence, it can be conceptualized as balls (nodes) in space with arrows (edges) connecting them. If the edges can be traversed in only one direction, the graph is known as a directed graph. Each directed edge represents a connection from one "source node" to one "sink node"; the sink node of one edge forms the source node for any subsequent nodes. The assembly process is like finding the path through the graph in a way that the path visits every node only once.[3]

the OLC algorithm are **Arachne**, **CABOG** (**Celera Assembler**), **Newbler**, **Minimus**, **Edena**, and **MIRA**. Overlap-based approaches have been mostly used for longer reads (>200 bp). However, overlap-based assemblers for short reads have also been developed.[4]

The de Bruijn–graph-based approach has been successfully employed in assembling short reads (<100 bp). However, de Bruijn graph assemblers have also been successfully used with longer reads.[4] Some assemblers that utilize the de Bruijn–graph algorithm are **Euler-SR**, **Oases**, **Velvet**, **ALLPATH**, **ABySS**, and **SOAPdenovo**. Sequence assembly based on significant sequence overlap, as done using the standard Sanger method, works well when there are a finite number of sequence reads to be assembled. However, next-gen sequencing generates hundreds of millions of sequence reads. The assembly of such a large number of sequence reads cannot be done easily using this traditional method. The problem of scalability is solved by using the de Bruijn graph. The de Bruijn graph does not use the actual sequence reads for assembly, but breaks each sequence read down to smaller sequences called *k*-mers. These *k*-mers are aligned using $(k-1)$ sequence overlaps. The actual size of k depends on sequence coverage, read length, etc., but usually is not less than half of the actual read length. For example, a 106-base read can be divided into 49 overlapping 58-mers (sequence read length − *k*-mer length + 1 = # of *k*-mers; hence, $106-58+1=49$). Because breaking one sequence read into *k*-mers increases the number of short sequence reads (e.g. just one 106-base read generates 49 *k*-mers, each one 58 bases long), it is likely that the resulting *k*-mers generated from all sequence reads will represent nearly all *k*-mers from the genome for sufficiently small k. This process seemingly compensates for missing sequence reads—that is, the sequence reads that could not be generated through sequencing for a variety of technical reasons.[5] Therefore, computational application of the de Bruijn graph helps alleviate many problems of de novo sequence assembly, but it is still not a fool-proof process.

With the improvement of sequence coverage and computing power, software is being constantly being developed or improved based on newer algorithms. Sequence reads can now be accurately assembled based on overlaps as small as 15 bp.[6]

A genome sequence assembly can be performed in two ways: **mapping and assembly**, or **de novo assembly**. If the genome has been sequenced before and a **reference genome** sequence already exists, then the newly obtained resequence reads are first mapped to the reference genome through alignment and then assembled in proper order; this mode of assembly is called "mapping and assembly." Bowtie is an ultrafast, memory-efficient short-read aligner that helps in mapping and assembly. It rapidly aligns large sets of short sequencing reads to a reference sequence, at a rate of over 25 million 35-bp reads per hour. For reads longer than about 50 bp, **Bowtie 2** is generally faster, more sensitive, and uses less memory than the original Bowtie (http://bowtie-bio.sourceforge.net/index.shtml).

In contrast, if there is no reference genome sequence then the assembly is called "de novo assembly." For de novo assembly, paired reads work better than single reads because paired reads help generate scaffolds. Therefore, genome assembly is a hierarchical process; it is performed in steps beginning from the assembly of the sequence reads into contigs, assembly of the contigs into scaffolds (supercontigs), and assembly of the scaffolds into chromosomes. Many genome assemblies remain restricted to scaffold level for a long time because the gaps can not be easily sequenced. Some scaffolds can be placed within a chromosome, while the chromosomal assignment of other scaffolds may remain difficult.

The de novo genome assembly can be assessed based on a number of parameters, such as the number of contigs and scaffolds available and their size, and the fraction of reads that can be assembled. One widely used metric to evaluate the quality of assembly is the contig and scaffold **N50** value (see Box 7.1). An N50 contig is the size of the shortest contig such that the sum of contigs of that size or longer constitutes at least 50% of the total size of the assembled contigs. For example, an N50 contig of 100 kb means that when contigs of 100 kb or longer are added up, the resulting size represents at least 50% of the total size of all assembled contigs. Likewise, an N50 scaffold size is the length of the shortest scaffold such that the sum of the scaffolds of that size or longer constitutes at least 50% of the total size of all assembled scaffolds.

Although genome sequencing has become high throughput and very cheap, and the computational power in genome-sequence assembly has tremendously increased, the current methods have many problems, partly owing to the nature of the genome sequence itself and partly owing to problems inherent in the sequencing method. Consequently, de novo sequence assembly is still a major challenge and can be fraught with errors and missing sequence.[7] This makes finishing a genome sequence and assembly a continuous and long-drawn-out process.

7.3 GENOME ANNOTATION

Genome annotation is the process by which biological information is assigned to the genome sequence. It involves the prediction of exons, introns, regulatory elements, various signal sequences, alternatively

BOX 7.1

The N50 contig value can be determined by first sorting all contigs in decreasing order of size, then adding the contigs until the total added size reaches at least half of the total size of all assembled contigs. The size of the smallest contig used in this addition process represents the N50. The scaffold N50 is calculated in the same fashion using the scaffold size. For example, if the contigs assembled are 0.43, 0.75, 1, 0.6, 0.8, 0.55, 0.32, and 0.25 Mbp, the total assembled size of all contigs is 4.7 Mbp. Now, organizing the contigs in decreasing order of size, we get: 1, 0.8, 0.75, 0.6, 0.55, 0.43, 0.32, and 0.25 Mbp. Adding just 1, 0.8, and 0.75 yields 2.55 Mbp, which is 54% of the total assembled size of all contigs. The smallest contig used in this addition process is 0.75 Mbp. Therefore, the N50 contig is 0.75 Mbp. The larger the N50 value, the better is the assembly. Using the same concept, higher values of N are also used, such as N60 and N80. If the N50 scaffold length is too short, additional rounds of shotgun sequencing are recommended.

spliced variants, noncoding RNAs, etc., that ultimately reflects the function and sheds light on molecular (sequence) evolution. Therefore, annotation has a structural aspect and a functional aspect. Annotation can be done computationally or manually; the latter requires human expertise. In reality, both computational and manual annotations are used to optimize the annotation process. Expectedly, the existence of similar annotated genomes greatly facilitates the annotation of newly sequenced genome. *The median gene lengths are roughly proportional to genome size; hence, bigger genomes have bigger genes*. Thus, accurate annotation of a larger genome requires a more contiguous genome assembly in order to avoid splitting genes across scaffolds.[8]

In brief, at the beginning of genome annotation, repeats are identified and masked computationally (e.g. using **RepeatMasker**; created by Smit, A.F.A., Hubley, R., and Green, P.; http://www.repeatmasker.org) because repeats, if not removed, can produce false evidence of gene annotations through spurious BLAST alignments. Repeats include low-complexity sequences (homopolymeric runs of nucleotides) and transposable elements, including long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Computational masking of repeat sequence frequently involves replacing the sequence with "N".

After repeat masking, the genome assembly is aligned to known expressed sequence tag (EST), RNA, and protein sequences; these sequences may include previously identified transcripts and proteins from the same organism whose genome is being annotated, or they may be from other organisms. When sequences from other organisms are used, evolutionarily conserved proteins provide useful information. The alignment process uses BLAST and BLAT (discussed in Chapters 2, 5, and 6) in order to rapidly identify approximate regions of homology. BLAT can also map these sequences to the genome. The alignment data are filtered to eliminate marginal alignments as revealed by low % identity or % similarity. The filtered alignment data are then inspected for the presence of redundant sequences, which would be removed. Further alignment is performed to obtain greater precision of exon boundaries using splice-site detecting alignment algorithms, such as **Splign** (http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi) and **Spidey** (http://www.ncbi.nlm.nih.gov/spidey/spideydoc.html). Both Splign and Spidey compute mRNA/cDNA-to-genome alignments, including spliced sequence alignments. Splign was developed by Kapustin et al.[9] and Spidey was developed by Wheelan et al.[10] Figure 7.2 shows how Splign can be used online. The example used is mouse *Slco1a6* mRNA (cDNA) (RefSeq NM_023718.3), which was mapped to and aligned with the mouse genome to find the genomic location of the exons and splice-junction sites. Figure 7.3 shows partial information of Splign output.

The final stage of annotation is best done manually but is being increasingly done computationally. Although manual annotation is high quality, it is time consuming, expensive, and labor intensive. In the age of massive genomic data generation, available genomic information, and increased computational power, genome annotation projects are increasingly utilizing automated annotation. The ultimate goal of annotation is to obtain a synthesis of alignment-based evidence with gene predictions to obtain a final set of gene annotations. Annotation of a genome undergoes repeated quality-control checks and it is a long ongoing process. The target for annotation is to generate a "high-quality draft" assembly that is at least 90% complete.[8] RNA sequencing (RNA-seq) data can be used to greatly improve the accuracy of gene annotations because such data provide strong evidence for exons, splice sites, and alternatively spliced exons. The interested reader is urged to read an excellent overview of eukaryotic genome annotation by Yandell and Ence.[8]

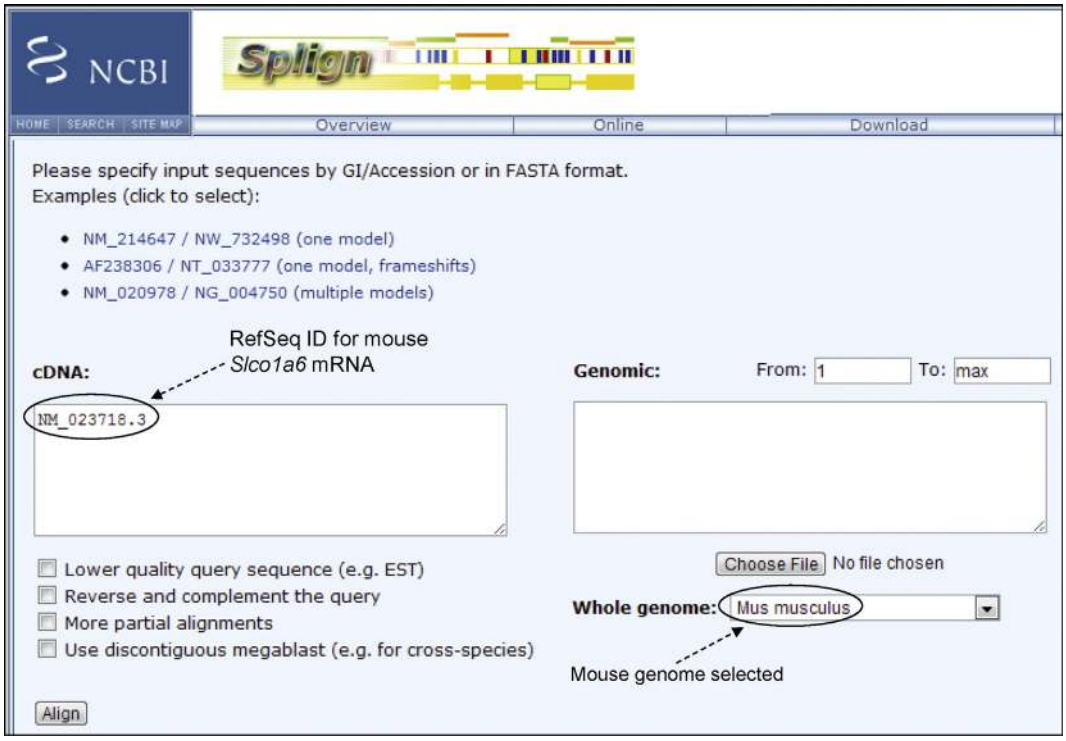


FIGURE 7.2 **The use of Splign online.** In the box for cDNA, either the sequence or the accession number/GI number can be entered. The sequence has to be entered in FASTA format. The example used is mouse *Slco1a6* mRNA (cDNA) (RefSeq NM_023718.3). The goal is to map the sequence to and align it with the mouse genome to find the genomic location of the exons and splice-junction sites. The default settings were maintained.

7.3.1 Gene Prediction

Gene prediction, which is part of genome annotation, involves the identification of putative coding exons in an unannotated DNA sequence. In other words, gene prediction attempts to predict putative coding sequences. The process is probabilistic and the putative exons are scored for the probability of being a true exon.

Gene prediction in prokaryotes (Bacteria and Archaea) involves fewer confounding factors than in eukaryotes because in prokaryotes the genome size is small and gene density is high, with ~88% of the genome containing coding sequences.[11] Bacteria do not have introns (Archaea have introns in rRNA and tRNA genes[12]), and the genomes have fewer repeat sequences. This is in contrast to eukaryotic genomes that are very large and full of repeat sequences; the majority of the eukaryotic genome is non-protein-coding, and the protein-coding genes contain large introns. Bacterial genes also have **Shine–Dalgarno sequence** (consensus AGGAGGT), which is the ribosomal binding site that lies upstream of the translational initiation codon (ATG) but downstream of the transcription start site. The end of the transcriptional unit (operon) has a terminator sequence that can form a stem–loop structure followed by a string of "T"s. The frequency of certain codons is much higher because of known codon preferences. These telltale signals, coupled with high gene density and fewer repeat sequences in the genomes, tend to make gene prediction in prokaryotes easier than in higher eukaryotes.

Gene prediction in an unannotated genome can be performed by **intrinsic** or ***ab initio*** prediction, **extrinsic** or **evidence-based** prediction, and **homology-based** prediction.

In the absence of any reference sequence (genome, EST, protein) from a related organism, gene prediction relies on **intrinsic** or ***ab initio*** prediction—that is, prediction based on the identification and analysis of telltale signals of protein-coding genes. In other words, the prediction is based on the information contained in the genomic sequence itself. Some of these signals are: start and stop codons, known codon preferences, intron splice signals, poly(A) signal sequence, TATA boxes, cap sites, transcription-factor-binding sites, Kozak sequence, and termination signals. In addition, the nucleotide composition differences known to exist between coding and noncoding regions as well as many essential features of gene structure are also taken into account, such as gene density, typical number of exons/gene, typical exon length, and open reading frame (ORF)-specific hexamer composition versus

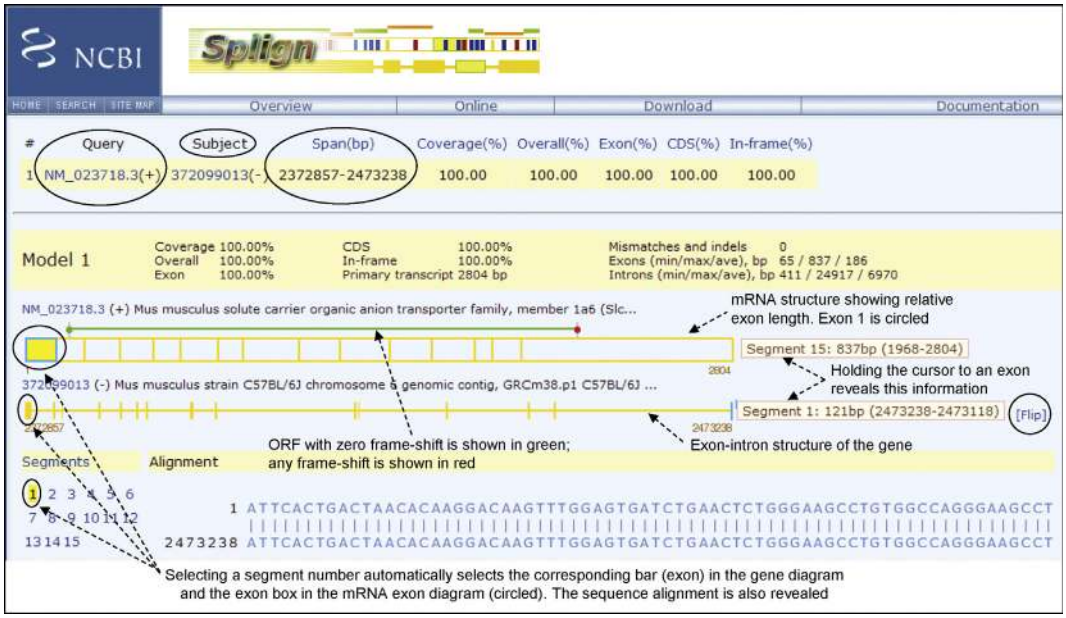


FIGURE 7.3 **Partial Splign output.** Splign has aligned the input sequence to the mouse genome, and has created 15 segments, displayed under "Segments" link on the left-hand side. In this example, each segment corresponds to one exon. Above the "Segments" link is the exon−intron organization of the gene, in which each exon is represented by a vertical line. Above the gene diagram is the mRNA diagram, in which each exon is represented by a box and the length of each box is proportional to the length of the exon. So, exon 15 (the last exon) is the longest. Above the mRNA, the open reading frame (ORF) is represented by a line. The green line here shows that there is no frameshift in the input sequence. Any frameshift would be represented by a partial red line. The green dot at the beginning and the red dot at the end of the ORF denote the start and the stop codon, respectively. Although not shown here, mismatches are denoted by vertical red lines and insertions/deletions (indels) are denoted by vertical blue lines inside the rectangular boxes representing exons. If the cursor is held close to an exon in the gene (vertical line), its genomic location appears as long as the cursor is held in place (segment 1 in this example); similarly, if the cursor is held close to an exon in the mRNA (rectangular box), its location in the mRNA appears (segment 15 in this example). Note that for the mRNA, the orientation is 5′→3′ from left to right; hence, segment 15 (exon 15) is at the right, whereas for the gene, the orientation is 5′→3′ from right to left; hence segment 1 (exon 1) is at the right. This is because the gene is located in the reverse orientation in the genome, which is indicated by the word "Flip" (right-hand side, circled). In the figure, the location of exon 15 (segment 15) of the mRNA and segment 1 (exon 1) in the genome are shown; one of them is copied and pasted separately in the figure. This is because only one at a time can be obtained, not both. As soon as a segment is selected, the corresponding vertical line in the gene diagram becomes blue and the corresponding rectangular box in the mRNA diagram becomes highlighted in yellow with its border becoming blue (in the figure, exon 1). Also, the alignment with the genomic sequence is displayed.

ORF-independent hexamer composition (in introns and intergenic regions).

The nucleotide composition of coding versus noncoding regions is analyzed using probabilistic statistics, such as various versions of Markov models. For example, the wobble base (third position in a codon) tends to be higher in G + C content in a coding region. Thus, if the local G + C content in a genomic region is significantly higher than the background, it suggests the likelihood of an ORF in that region. The sequence can be translated in all six frames (three sense, three antisense). Because there are 3 stop codons plus 61 amino-acid codons, a random unbiased distribution of bases should produce approximately 1 stop codon for every 20 codons in an ORF search. If the region is rich in A + T, a stop codon is expected even before 20 codons because the stop codons (TAA, TAG, TGA) are A + T rich (7 A + T out of 9 bases). These features and generalizations are expected for noncoding regions, but not for coding regions. Therefore, if an ORF search of a genomic region produces a translated ORF that shows a significantly high number of codons, such as >50 or so, before a stop codon appears, it suggests the likelihood of a legitimate ORF. With some exceptions, the number of codons in most ORFs is far greater than 60; in fact, proteins containing <200 amino acids are still considered to be small proteins and are known to play important roles in development.[13] Therefore, the *ab initio* approach combines statistical analyses along with other gene signals for gene prediction.

AUGUSTUS (http://bioinf.uni-greifswald.de/augustus/submission) is an *ab initio* gene-prediction program that uses the hidden Markov model (HMM; see Box 7.2). The program has used a diverse training set of approximately 60 genomes belonging to four different groups of organisms: animals; Alveolata (single-celled eukaryotes); plants and algae; and fungi, and is therefore able to predict genes in a wide range of species. The original version of AUGUSTUS utilized a purely *ab initio* method and was

BOX 7.2

THE HIDDEN MARKOV MODEL

Gene-prediction algorithms have become more sophisticated with the incorporation of statistical methods, particularly the Markov model and its variants. A **Markov model** is a stochastic model—that is, a model to predict the outcome of a stochastic (random) process. The simple Markov model is a **Markov chain** that represents an ordered sequence of discrete events, moving from one "state" (event) to another with a certain probability, called the **transition probability**. In a Markov chain, at any given point in time, each current state has a previous state s_i, which has evolved into the current state s_j with a transition probability p_{ij}, and the current state s_j will evolve into a future state s_k with a transition probability p_{jk}. In this sequence of events, p_{jk} depends on s_j but not s_i. In other words, a Markov model assumes that the probability of the future state depends on the current state but NOT on the past state.

A Markov model predicts the evolution of an observable event that depends on internal factors. The observable event can be called an "output signal" and the internal factor can be called a "state." In a Markov model prediction, both the "output signal" and the "state" are observable. Markov models are used to predict many events in day-to-day life, such as stock market performance, to make weather forecasts, and so on. In contrast to Markov models, in the hidden Markov model (HMM) the "output signal" is observable but the "state" is not. Examples of HMM from biology are DNA and protein sequences. A DNA sequence is an observable output signal (from sequence determination) but the state of the sequence—that is, whether the sequence belongs to exon or intron or regulatory element or intergenic region—is not directly observable. Similarly, the sequence of amino acids in a protein is an observable output signal (from sequence determination), but the state of the sequence—that is, whether the sequence is part of a specific domain (e.g. a transmembrane domain)—is not directly observable. These hidden states can be modeled and predicted with certain probabilities by HMM. Consequently, HMMs have been used in, among other things, gene prediction, pairwise and multiple sequence alignment, base-calling, modeling DNA sequencing errors, protein secondary structure prediction, noncoding RNA (ncRNA) identification, RNA structural alignment, acceleration of RNA folding and alignment, and fast noncoding RNA annotation.[14]

Markov models can be **fixed order** or **variable order**, as well as **inhomogeneous** or **homogeneous**. In a fixed-order Markov model, the most recent state is predicted based on a fixed number of the previous state(s), and this fixed number of previous state(s) is called the **order** of the Markov model. For example, a **first-order** Markov model predicts that the state of an entity at a particular position in a sequence depends on the state of one entity at the preceding position (e.g. in various *cis*-regulatory elements in DNA and motifs in proteins). A **second-order** Markov model predicts that the state of an entity at a particular position in a sequence depends on the state of two entities at the two preceding positions (e.g. in codons in DNA). Similarly, a **fifth-order** Markov model predicts the state of the sixth entity in a sequence based on the previous five entities (e.g. in hexamers in coding sequence). It has been observed that the probability of occurrence of pairs of codons (hexamers) in a coding sequence is significantly higher than in noncoding sequence. A fifth-order Markov model calculates the probability of the sixth base based on the previous five bases in the sequence. In addition to the order, if the probability of occurrence of the state also depends on the position within the sequence, the model is called an inhomogeneous Markov model. In contrast, in a homogeneous Markov model all positions in the sequence are described by the same set of conditional probabilities.

Fifth-order Markov models are often used in gene prediction. For example, **GeneMark** (http://opal.biology.gatech.edu/GeneMark/) is a family of gene-prediction programs that uses an inhomogeneous fifth-order Markov model. However, a potential problem with a higher-order (e.g. fifth-order) Markov model is having enough data for the training set. For example, a fifth-order Markov model will require 4^5 (=4096) probabilities (probable combinations) to be estimated from the training data. In order to estimate these probabilities, many occurrences of all possible *k*-mers must be present in the data. The lack of availability of such huge amount of data may limit the usefulness of a higher-order Markov model. The **interpolated Markov model** (**IMM**) overcomes this problem by combining probabilities from contexts of varying lengths to make predictions, and by only using those contexts (oligomers) for which sufficient data are available.[15] The IMM method involves sampling dimers ($k = 1$) to nine-mers ($k = 8$) and adding the probabilities of all weighted *k*-mers, placing less weight on rare *k*-mers and more weight on more abundant *k*-mers. Therefore, the probability of the model is the sum of all probabilities of all weighted *k*-mers for which sufficient data are available. **GLIMMER** (Gene Locator and Interpolated Markov ModelER) is a microbial gene prediction and genome annotation tool that uses IMM and is available to run online at the NCBI (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi). The majority of gene-prediction software uses HMM for prediction.

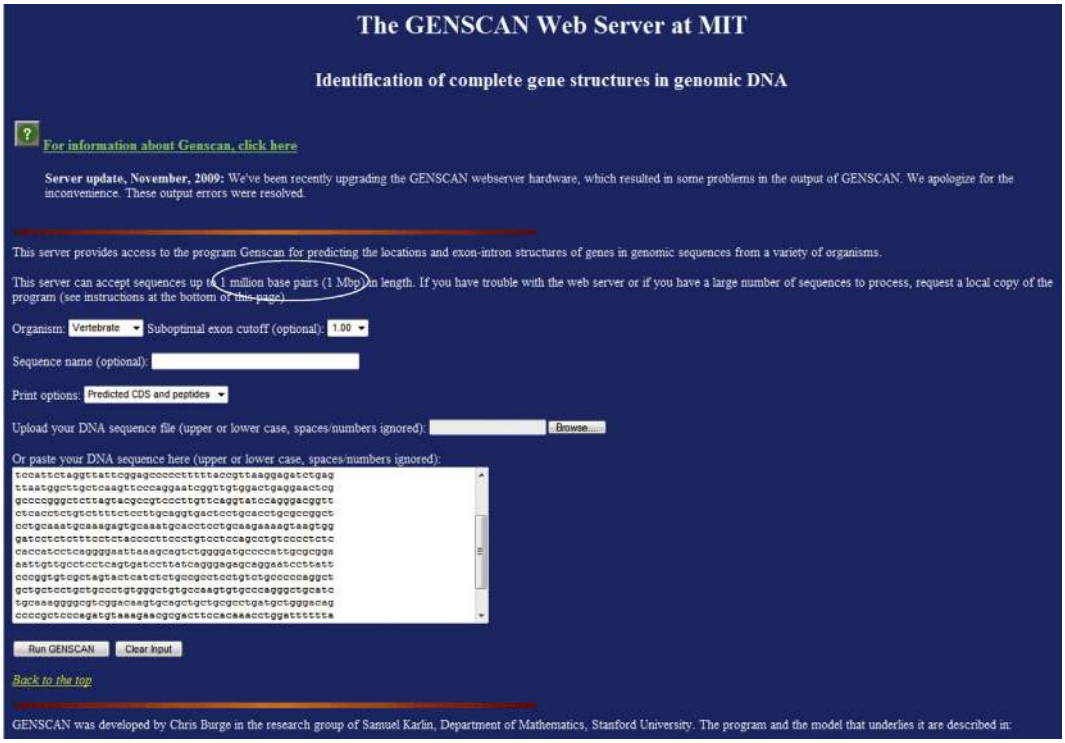


FIGURE 7.4 **GENSCAN home page.** Currently, GENSCAN can analyze an input sequence of up to 1 million bases (circled).

found to be one of the best *ab initio* algorithms for gene prediction.[16] **FGENESH** is a very fast and accurate *ab initio* gene-prediction program. The SoftBerry home page (http://linux1.softberry.com/berry.phtml) provides link to FGENESH and to a diverse set of other bioinformatics applications. **GENSCAN** (http://genes.mit.edu/GENSCAN.html) is another *ab initio* prediction tool developed early on by Dr Chris Burge in the research group of Samuel Karlin at Stanford University[17]; it also utilizes HMM. GENSCAN was trained using 570 vertebrate gene sequences.[18] When tested on standardized sets of human and vertebrate genes, GENSCAN accurately predicted 75 to 80% of exons.[17] Figure 7.4 shows the GENSCAN home page, and Figure 7.5 shows a GENSCAN analysis of a 932-bp input DNA fragment[f].[19] Based on the G + C content, the input sequence is predicted to belong to **isochore 3**[i] (circled).

Ab initio prediction algorithms fail to accurately predict alternative splicing, very long or short exons, nested and overlapping genes, any non-canonical features associated with the gene (e.g. non-ATG start codon, selenocysteine codons, split start or stop codons, etc.). Purely *ab initio* predictions are generally 50% or less accurate at the gene level.

Another approach is **extrinsic** or **evidence-based** prediction, in which some information is available, such as mRNA, EST, or protein product information. As more and more genomes have been sequenced and annotated, and more and more genomic information has become available, the pure *ab initio* prediction algorithms have been modified to incorporate genomic information and develop extrinsic prediction algorithms. For example, the newer version of AUGUSTUS combines the prediction ability of an *ab initio* algorithm with extrinsic information, such as matches to protein databases or alignments of genomic sequences, to improve the prediction accuracy. Because of this improvement, the new version of AUSGUSTUS is also able to predict splice variants, which the original algorithm could not do. **MAKER 2** (http://www.yandell-lab

[f]GenBank: NC_000016.9, Region: 56642478 − 56643409

[i]**Isochores** have been defined as >300-kb-long DNA segments in warm-blooded vertebrates (birds and mammals) with a characteristic, relatively homogeneous base composition. Based on the G + C content, isochores are classified in two "G + C-poor" types (L1 and L2) and three "G + C-rich" types (H1−H3). The average G + C content of isochore 3 (H3) is the highest (~ 54%) and it constitutes ~ 3% of the genome. In general, genes with higher G + C content belong to G + C-rich isochores (types H1−H3). The H2 and H3 isochores together have been termed the "genome core" because of their higher gene concentrations, which makes up about 12% of the genome (9% for H2 and 3% for H3). In the human genome, the H3 isochore apparently contains 25% of the genes, and the genome core (H2 + H3 combined) contains about 54% of the genes.

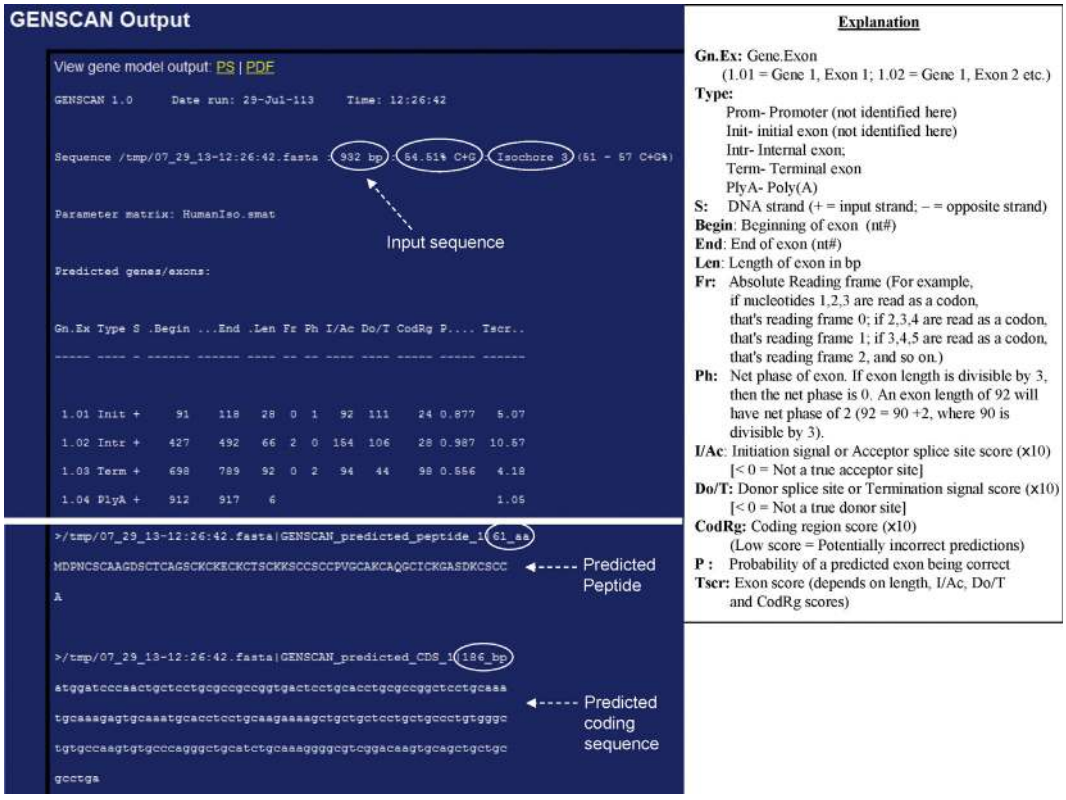


FIGURE 7.5 **GENSCAN analysis of a input DNA sequence fragment.** The upper left panel shows the analysis output and also the length of the input sequence (932 bp) and its G + C content (54.51%) (circled). Based on the G + C content, the input sequence is predicted to belong to isochore 3 (circled). The lower left panel shows a 186-bp predicted ORF and a 61-amino-acid predicted protein. The abbreviations are explained in the right-hand panel.

.org/software/maker.html) is another gene-prediction and genome-annotation program that combines *ab initio* and extrinsic approaches to produce gene annotations having evidence-based quality values. **GenomeScan** (http://genes.mit.edu/genomescan.html) is the successor of GENSCAN and it performs gene prediction in humans and other vertebrates. The algorithm utilizes two principal sources of information: (1) models of exon−intron and splice-signal composition; and (2) sequence similarity information, such as BLASTX hits. The probabilistic model used by GenomeScan is based on that used by GENSCAN.

Homology-based prediction relies on identifying significant matches of the query sequence with sequences in known and annotated genome sequences from related species. Thus, homology-based prediction relies on comparative genomics, and has been made possible because the genomes of many organisms have been sequenced. Homology-based prediction is based on the molecular evolutionary principle that functionally important parts of the genome evolve at a slower rate compared to the rest of the genome; therefore, many gene sequences, particularly in related species, should be highly conserved and therefore be recognizable by the prediction algorithm. Consequently, homology-based prediction has a high level of accuracy, and the greater the number of available genomes of related species, the greater the accuracy and completeness of prediction. The homology-based gene prediction tools align syntenic regions of unannotated genomes, and utilize a probabilistic framework for gene structure prediction. Several programs have been developed for homology-based prediction, such as **SLAM** (http://baboon.math.berkeley.edu/~syntenic/slam.html), **CEM**, and **Twinscan/N-SCAN** (http://mblab.wustl.edu/software.html), and **EuGene'Hom** (http://tata.toulouse.inra.fr/apps/eugene/EuGeneHom/cgi-bin/EuGeneHom.pl); for plant genomes. Comparative-genomics-based gene-finding programs outperform *ab initio* gene-finding programs.[20,21]

Many of these software programs can be downloaded for noncommercial and research purposes to carry out sequence analysis and gene prediction. A list of many gene-prediction software programs is available at the geneprediction.org website (http://www.geneprediction.org/software.html). Many of these can be accessed and run online by simply entering the input sequence either in plain text format or in FASTA format. The reader can try these links using known

genomic sequence (containing a known gene) and learn firsthand how each algorithm performs gene prediction and what the different outputs look like. A flow-chart for practice activity is given below.

Go to the NCBI home page → select "Gene" from the drop-down list of databases → enter Oatp-5 (or Slco1a6) in the "Search" space and hit enter → from the "Results" page, click "Mus musculus Slco1a6" → scroll down the Slco1a6 page → under the "NCBI Reference Sequences (RefSeq)" bar, locate the section "Reference GRCm38.p1 C57BL/6J" → under this section, locate the heading "NC_000072.6" → under this heading, click the "GenBank" link[j].

This will take the user to the RefSeq nucleotide sequence page of chromosome 6 showing VERSION NC_000072.6 and GI: 372099104. The sequence is 100,382 bases long. Copy the sequence. Now open a new web browser page → Google "Readseq" (the file conversion tool) → open Readseq from any of the sites, such as EBI, NIH, or Indiana University link → paste the sequence → from the "Output format" drop-down menu select the format as "Plain/Raw" if plain text format is desired or "Pearson/Fasta" if FASTA format is desired, and check the box for "Remove gap symbols" (or "degap" if using the EBI link). "Submit" the sequence and the desired sequence format will be returned without base numbers and gaps. Now copy this sequence and paste it in any of the gene prediction tools and run gene prediction. The Readseq link at the Indiana University site (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi) provides an option to download the sequence file, but the default is "View in browser."

Although the three approaches have been discussed here separately; in reality they are combined to increase the prediction accuracy. The sequencing and annotation of an ever-increasing number of prokaryotic and eukaryotic genomes have made it possible to successfully combine all three approaches. A common current approach for gene finding involves the following activities: several sets of gene predictions by different gene finders are compiled, and alignments from ESTs and proteins to the genome are constructed. All these data are combined to find the most plausible gene sequence, either manually or by using meta tools that combine several predictions and alignments.[16]

7.4 PREDICTION OF PROMOTERS, TRANSCRIPTION-FACTOR-BINDING SITES, TRANSLATION INITIATION SITES, AND THE ORF

Many free software packages are available online for the prediction of putative promoter sequences, transcription start sites, *cis*-regulatory elements, translation initiation sites, and the ORF.

Transcription of all classes of RNA (rRNA, mRNA, tRNA) in prokaryotes is catalyzed by one RNA polymerase, which is a multi-subunit enzyme. It contains a core polymerase that is composed of five subunits (α^{I}, α^{II}, β, β', ω), and a sigma (σ) factor. The **sigma factor** is the **initiation factor** that helps position the core polymerase to the promoter. The promoter has two consensus sequences, one at the −10 position (TATAAT in *Escherichia coli*), also known as the **Pribnow** box, and the other at the −35 position (TTGACA in *E. coli*) relative to the transcription start site. Bacteria possess different types of sigma factors. In *E. coli* and other bacteria, the sigma factor that initiates transcription of housekeeping genes and many other genes has a molecular weight of 70 kDa (hence σ^{70}). In prokaryotes, a transcriptional unit (i.e. an **operon**) may contain one gene or a number of genes under the control of one promoter. The transcription of one gene produces **monocistronic** RNA, whereas the transcription of many genes produces **polycistronic** RNA. Therefore, the promoter is located upstream of the first gene in a polycistronic transcriptional unit. Wang et al.[22] predicted operons in *Staphylococcus aureus* with $>90\%$ accuracy using a scoring system to annotate the intersection between two genes. In other words, this method identified whether two adjacent genes belong to the same operon. The scoring system was based on a number of parameters, such as intergenic distance, presence/absence of a terminator, comparison with other known prokaryotic genomes, etc.

Transcription in eukaryotes is carried out by three different RNA polymerases—RNA polymerases I, II, and III—which all bind to the promoter regions of the respective genes that will be transcribed. Of these, RNA polymerase II (pol II) produces translatable mRNAs. RNA pol II binds to the promoter, and also interacts with various other proteins for transcription. The DNA-binding proteins bind to specific sequence elements, called *cis*-response elements or *cis*-regulatory elements, that are all located at variable distances upstream of the transcription start site. The eukaryotic promoter can be divided into the **core** (or basal), **proximal**, and **distal** promoter, based on function and distance from the transcription start site.

In general, the transcription start site is determined by the TATA box (consensus TATAAA) and initiator (Inr) element (consensus: Y-Y-+1-N-T/A-Y-Y, where Y = pyrimidine, +1 = transcription start site, N = any nucleotide), or by the Inr element and downstream promoter element (DPE; consensus: $(A/G)_{+28}$ G(A/T) (C/T)$(G/A/C)_{+32}$) in the case of TATA-less promoters.

[j]These commands are current as of July, 2013. They may change if the mouse genome assembly version changes.

Typically, the core promoter is about 35 bp long, and can extend either upstream or downstream of the transcription start site (−35 to +35).[23] The core promoter may contain two or more of the following sequence motifs: TATA box, Inr element, and DPE. In most higher eukaryotic genes, the TATA box is located approximately 25-nt upstream (usually between −30 and −25) from the transcription start site. In many genes, a variation of the classic Inr may be present.[24]

The proximal promoter is about 250 bp long and can extend between the −250 and +250 nt positions, relative to the transcription start site.[25] Two transcription-activating response elements found in the proximal promoter are the CAAT box (binds the transcription factor NF-I) and the GC box (binds the transcription factor Sp1). The CAAT box is located ~75 nt upstream of the transcription start site and has a consensus sequence GG(T/C)CAATCT. The GC box is located ~90 nt upstream of the transcription start site and has a consensus sequence GGGCGG. The CAAT box and the GC box operate as enhancer elements because they can activate transcription in an orientation-independent manner.

Distal promoter sequences are further upstream of the proximal promoter elements.[26] The majority of transcription-regulatory protein-binding sites are located within 500 bp upstream of the transcription start site. Some regulatory-protein-binding sites can also be located downstream of the transcription start site.

Prediction of the translation initiation site (TIS) in a genomic sequence is an important problem to address. TIS prediction at the genome level is still not a trivial task because of the noise in the data. Some algorithms take into account weighted signal-based translation initiation site scores as well as the coding potential of sequences flanking TISs. At the gene level, an important sequence feature relevant for translation initiation and identification of the correct ATG codon by the translation initiation complex is the **Kozak sequence**. The original functional Kozak sequence (in the sense strand of DNA) was described as 5′-GCCRCC**ATG**G-3′ (where R is a purine, which in most vertebrate mRNAs is an "A"; **ATG** is the translation initiation codon). A shorter and more effective version (5′-ACC**ATG**G-3′) of the original Kozak sequence was also described later. The translation initiation region is characterized by certain features. Many genes contain the consensus Kozak sequence while others contain some variant. Still others may not have any Kozak sequence at all. The "G" after the ATG (i.e. ATG**G**) is the most prevalent base in the vast majority of mRNAs. If there is an ATG codon before the actual start codon, the sequence context of that ATG codon—such as lack of Kozak sequence around it, lack of a "G" immediately following the ATG, etc.—can help the ribosome bypass the incorrect ATG and detect the right ATG codon through scanning (known as **leaky scanning**). The incorrect ATG is usually out of frame with respect to the true initiation codon. If translation is initiated from the incorrect ATG codon that precedes the correct ATG codon, the ribosome encounters a premature stop codon, which is in-frame with the incorrect ATG codon. In such cases, translation is initiated again (**reinitiation**) from the correct initiation codon.

The National Center for Biotechnology Information (NCBI) ORF prediction tool **ORF Finder**[k] (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) is a graphical analysis tool that finds all ORFs of a selectable minimum size in the six frames (three sense; three antisense), using the standard or alternative genetic codes. The ORF translation in three frames is achieved by sliding the translational frame one base at a time. Because the genetic code is triplet, moving by three bases will find all possible frames. Figure 7.6A shows the graphics of computational translation of mouse *Slco1a6* mRNA in six frames. When the longest predicted ORF (top frame) is clicked, the sequence and other details of the sequence are displayed (Figure 7.6B). The entire sequence is not displayed in the figure. Clicking the "SixFrames" link shows the six frames (Figure 7.6C). In each of these frames, the blue vertical lines represent the in-frame ATG codons and the red lines represent in-frame STOP codons. As is evident, each of these frames except the top one is full of in-frame stop codons. The total number of entries on the right-hand side (15), each with a small blue square, corresponds to the total number of translational reading frames present in all six frames combined; hence, each entry on the right corresponds to one translational reading frame. Clicking any blue square reveals the corresponding translational reading frame (both turn red), and the sequence of the reading frame is revealed.

There are many online tools available for the prediction of promoters and *cis*-regulatory elements. These programs are not all trained on the same training data set; consequently, the prediction outputs may not be identical. Thus, it is a good idea to check the prediction using multiple programs to find out at least the common elements predicted by different programs. *It should be remembered that the bioinformatic predictions of the* cis*-regulatory elements (regulating transcription) as well as the translation initiation site (i.e. the beginning of the ORF) need to be experimentally verified. A more than 10% error rate in computationally predicted ORFs compared to experimentally derived values has been reported.* The errors are due to the variation in predicting the translation initiation site. Such error is partly due to

[k]Tatiana Tatusov and Roman Tatusov are credited on the ORF Finder home page.

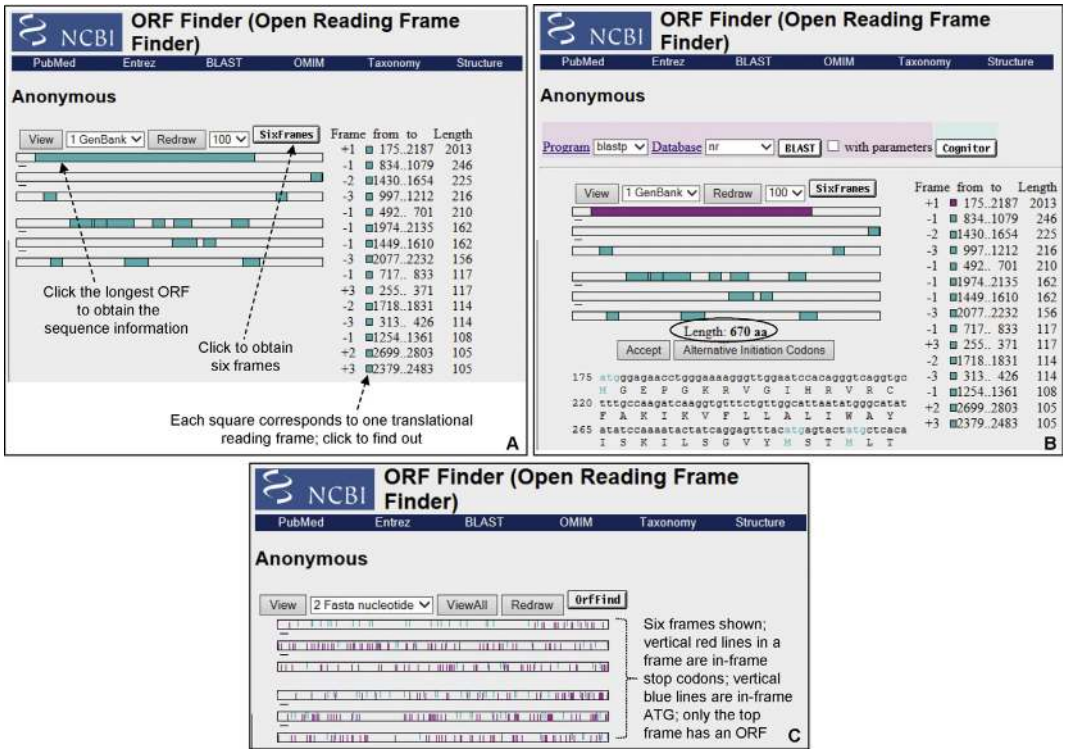


FIGURE 7.6 **NCBI ORF Finder.** (A) Computational translation of mouse *Slco1a6* mRNA in six frames, three sense and three antisense. (B) When the longest predicted ORF (top frame) is clicked, the sequence and other details of the sequence are displayed. Only the upper portion of the entire sequence is displayed. (C) Clicking the "SixFrames" link shows the six frames.

the ORF-prediction algorithm used, and partly due to the taxon examined. For example, genomes having high G + C content are particularly susceptible to ORF-prediction errors because of the existence of the alternative start codon GTG.[27]

Some of the publicly available online tools for the prediction of promoters, *cis*-regulatory elements, transcription start sites, translation initiation sites, and the ORF are listed in Table 7.1. There are many more prediction tools available. The reader can use these tools to obtain a rapid prediction about an input sequence, and compare the predictions of different tools.

7.5 RESTRICTION-SITE MAPPING OF THE INPUT SEQUENCE

Experiments involving DNA often require the experimenter to use various restriction enzymes. Restriction enzymes may be used to simply cut the DNA for gel electrophoresis or for advanced manipulation of DNA, such as making a vector, or a transgenic or knockout construct. Two online resources that can be used to analyze various restriction-enzyme cutting sites and generate a restriction map of an input DNA sequence are Webcutter 2.0[1] (http://rna.lundberg.gu.se/cutter2/) and NEBCutter 2.0[38] (http://tools.neb.com/NEBcutter2/).

7.6 RNA SECONDARY-STRUCTURE PREDICTION

RNA is single stranded but it can form significant secondary structure because of intrastrand base pairing. The three-dimensional shape of an RNA is its secondary structure. Some secondary structures observed in RNA are **short duplexes**, **stem−loops** (**hairpin stem−loops**), **bulges**, **internal loops**, **pseudoknots**, etc. (Figure 7.7A). The secondary structure of an RNA plays an important role in its maturation, regulation, and function. In fact, the formation of RNA secondary structure is the key to some of its functions regulating gene expression. For example, during **translational reprogramming**, or **recoding**, the gene-encoded reading frame is altered during translation, which allows for the generation of multiple ORFs from the same basic ORF encoded by the gene. This is achieved by

[1]©1997 Max Heiman.

TABLE 7.1 Some Online Tools for Prediction of Promoters, *Cis*-Regulatory Elements, Transcription Start and Initiation Sites, and the ORF

Online Analysis Tool	Comments and URL
BPROM	**Bacterial promoter prediction**. A SoftBerry utility that predicts putative transcription start positions of bacterial genes regulated by sigma70 promoters. The prediction accuracy is about 80%; the specificity is also about 80% when tested on equal numbers of promoter and non-promoter sequences. It uses the signal and content information of the sequence (e.g. consensus sequence). BPROM should be run on a region between two neighboring ORFs located on the same strand, or on a sequence upstream from an ORF (most promoters are located within 150 bp upstream of the ORF). BPROM should not be used for whole genomes, to avoid the many false positives (http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb)
Virtual Footprint	**Prokaryotic promoter prediction**. Virtual Footprint is a software suite for analyzing transcription-factor-binding sites in whole bacterial genomes and their underlying regulatory networks. The result is a list of potential binding sites and corresponding genes defining the whole regulon. There are two types of analysis: analysis of a whole prokaryotic genome with one regulator pattern, and analysis of a promoter region with several regulator patterns[28] (http://www.prodoric.de/vfp/vfp_promoter.php)
BDGP (Berkeley *Drosophila* Genome Project)	**Prokaryotic and eukaryotic promoter prediction.** Neural network promoter prediction (NNPP)-based. NNPP is method that consists mainly of two recognition features for predicting eukaryotic promoters; one for recognizing the TATA-box and one for recognizing the initiator element. Both features are combined into one output unit, which gives output scores between 0 and 1. The default score is set at 0.8. The prediction accuracy for prokaryotic promoters is greater than that for eukaryotic promoters[29] (http://www.fruitfly.org/seq_tools/promoter.html)
FindTerm	**Rho-independent-terminator prediction in the bacterial genome**. A SoftBerry utility that predicts terminators in the bacterial genome. The search utilizes certain known features of bacterial terminators, such as T-rich regions, possible combinations of spacer lengths, all hairpins etc., and the result output shows all putative terminators (http://linux1.softberry.com/berry.phtml?topic=findterm&group=programs&subgroup=gfindb)
Promoter 2.0	**Vertebrate pol II transcription start site (TSS) prediction**. The program builds on principles that are common to neural networks and genetic algorithms[30] (http://www.cbs.dtu.dk/services/Promoter/)
Tfsitescan	**Eukaryotic promoter sequence and putative transcription-factor-binding site prediction**. Works best with sequences of ~500 nt. The output is in graphic display and shows expectation scores for the putative binding sites[a] (http://www.ifti.org/cgi-bin/ifti/Tfsitescan.pl)
SoftBerry Search for promoters/ functional motifs	**SoftBerry utility providing a suite of prediction tools for promoter/functional motif prediction**. For example: **1.** Plant promoter prediction (**TSSP**) **2.** Human pol II promoter prediction (**TSSG** and **TSSW**) **3.** Human promoter prediction (**FPROM**) **4.** Promoter prediction using orthologous sequences in eukaryotic genome (**PromH(G)** and **PromH(W)**) **5.** Regulatory motif prediction (**Nsite**) (http://linux1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=promoter)
WWW Signal Scan	**Eukaryotic transcriptional elements prediction** based on scoring homologies of published *cis*-regulatory transcriptional signal sequences (e.g. in TFD, TRANSFAC databases) in the input sequence[b,31] (http://www-bimas.cit.nih.gov/molbio/signal/)
WWW Promoter Scan	**Eukaryotic promoter prediction** based on scoring homologies with eukaryotic pol II promoter sequences. If the program finds a putative promoter sequence, it reports the sequence range of the putative promoter, including the TATA box (if present) and the estimated transcription start site[32] (http://www-bimas.cit.nih.gov/molbio/proscan/)
Human Core-Promoter Finder	**Transcription start site (TSS) prediction in human core-promoters**. The input genomic DNA sequence should be longer than 240 bp and less than 2001 bp. The functional core-promoter is assumed to span between −60 and +40 nt with respect to the TSS (+1). The program is able to localize a TSS to a 100-bp interval ~60% of the time[c]. (http://rulai.cshl.org/tools/genefinder/CPROMOTER/human.htm)

(*Continued*)

TABLE 7.1 (Continued)

Online Analysis Tool	Comments and URL
EP3 (Easy Promoter Prediction Program)	**Eukaryotic core promoter prediction**. Performs very well in identifying regions in human genes that are associated with transcription initiation. EP3 uses universal properties of the promoter to detect those regions in a whole-genome context[33] (downloadable) (http://bioinformatics.psb.ugent.be/webtools/ep3/)
Eponine	**Transcription start site prediction in mammalian genomic sequence**. A probabilistic method with good specificity and excellent positional accuracy. Eponine is estimated to detect > 50% of transcription start sites, with ~70% specificity[34] (downloadable from Sanger Center) (http://www.sanger.ac.uk/resources/software/eponine/)
Footprinter	**Prediction of regulatory elements in DNA sequences based on phylogenetic footprinting**. Phylogenetic footprinting method identifies regions of DNA that are highly conserved across a set of orthologous sequences[35] (downloadable from the University of Washington (Motif Discovery link) (http://bio.cs.washington.edu/software)
ORF Finder	**Open reading frame (ORF) prediction**. A very user-friendly ORF finder on the web. It is a graphical analysis tool that finds all ORFs in the input sequence, using the standard or alternative genetic codes. The putative ORFs are displayed in six frames, three sense and three antisense[d] (http://www.ncbi.nlm.nih.gov/gorf/gorf.html)
NetStart 1.0	**Translation initiation site prediction**. NetStart produces neural network predictions of translation start sites in vertebrate and *Arabidopsis thaliana* nucleotide sequences. The program has been trained on cDNA-like sequences; therefore, it shows better performance for cDNAs and ESTs. It has not been tested on genomic data[36] (http://www.cbs.dtu.dk/services/NetStart/)
ATGPr	**Translation initiation site prediction**. ATGpr can be used to predict whether an initiation codon is present or absent in a piece of cDNA, and predict which ATG is the initiation codon for cases where there are multiple ATG codons. The method uses linear discriminant analysis, and has been tested on a non-redundant data set of 660 sequences[37] (http://atgpr.dbcls.jp/)

[a] *Made available by the Institute for Transcriptional Informatics (IFTI) at the IFTI-MIRAGE website.*
[b] *WWW implementation by Robin Hart and Rao Parasa.*
[c] *The web version is offered by Michael Zhang.*
[d] *Tatiana Tatusov and Roman Tatusov are credited on the ORF Finder home page.*

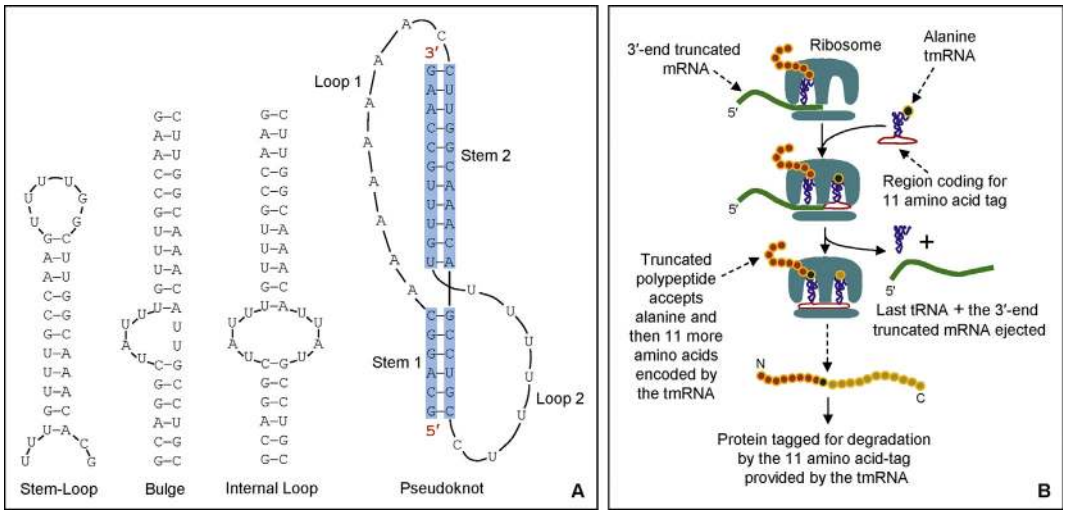


FIGURE 7.7 RNA secondary structure. (A) Some secondary structures of RNA. RNA pseudoknots can be more complex than the one shown here. (B) The transfer-messenger RNA (tmRNA; 10Sa RNA) and *trans*-translation. Alanine-charged tmRNA helps resume translation of a 3′-end-truncated mRNA by first providing alanine and then providing its own coding sequence, which adds the 11-amino-acid sequence to the C-terminal of the previously translated truncated polypeptide. The 11-amino-acid sequence tags the protein for degradation.

switching the reading frame during translation by one base, the so-called −1 or +1 frameshift mechanism. The efficiency of frame shifting is directly correlated with the extent of ribosomal pause. The *cis*-acting structural motifs of the mRNA that apparently facilitate ribosomal pause and consequent frame shifting include a heptanucleotide **slippery sequence** at the shift site, and a **pseudoknot** secondary structure that begins five or six nucleotides downstream from the shift site.

It is well recognized that the secondary structures of tRNA and ribozyme are necessary for their function. The telomerase RNAs in different species of ciliates and vertebrates have very different sequences but they all fold into similar secondary structures, strongly suggesting that the conserved secondary structure is important for the specific function of telomerase RNA.[39]

The **transfer-messenger RNA** (**tmRNA**) in bacteria that mediates ***trans*-translation** also has a unique secondary structure that is needed for its function. The phenomenon of *trans*-translation involves **ribosomal hopping**, involving two distinct RNA templates in succession. In various bacteria, this 10Sa RNA species acts as an alanyl tRNA because it is charged with alanine by alanyl-tRNA synthetase. The 10Sa RNA also has mRNA features because it encodes an 11-amino-acid oligopeptide that tags proteins for degradation. Because 10Sa RNA possesses such dual features of tRNA and mRNA, it is called transfer-messenger RNA (tmRNA). When ribosomes carrying a peptidyl-tRNA pause at the end of a 3′-end-truncated mRNA and accept the alanyl-10Sa RNA molecule as the alanyl-tRNA surrogate, the alanyl-10Sa RNA first provides the alanine and then provides its internal reading frame for the translation of the 11-amino-acid oligopeptide tag. This results in the incorporation of the oligopeptide tag to the already synthesized truncated polypeptide, which is thus flagged for degradation (Figure 7.7B).

An example of the importance of RNA secondary structure in its maturation is the biogenesis of micro RNA (miRNA). Transcription of a miRNA gene produces primary miRNA (pri-miRNA), which has a stem−loop structure with additional internal loops. Processing of pri-miRNA in the nucleus by Drosha produces precursor miRNA (pre-miRNA) which has a shortened stem−loop structure compared to pri-miRNA. Processing of pre-miRNA in the cytoplasm produces miRNA. The secondary structure of these precursors is necessary for the biogenesis of miRNA. An RNA hairpin is an essential secondary structure of RNA that can guide RNA folding, determine interactions in a ribozyme, protect mRNA from degradation, serve as a recognition motif for RNA-binding proteins, and also regulate gene expression.[40] A recent study using a high-throughput sequencing-based structure-mapping approach in *Drosophila melanogaster* and *Caenorhabditis elegans* transcriptomes identified both paired (double-stranded) and unpaired (single-stranded) RNA components. The authors observed that these RNAs are significantly correlated with specific epigenetic modifications. They also uncovered highly base-paired RNAs, many of which likely encode lncRNAs (long noncoding RNAs). Additionally, they identified conserved features of mRNA secondary structure that indicate that RNA folding demarcates regions of protein translation. Finally, they identified and characterized 546 mRNAs whose folding pattern is significantly correlated between these two species even though they are so far apart in evolution, thereby suggesting that the observed mRNA secondary structure has some function.[41]

The formation and stability of RNA secondary structure are dependent on a number of factors. For example, more GC base pairs and longer stem regions result in greater stability of the secondary structure, whereas unpaired bases, such as bulges and internal loops, tend to decrease the stability of the secondary structure. Similarly, the formation of hairpin loops with more than 10 or less than 5 bases requires more energy; hence, it reduces the stability of the secondary structure. In general, a secondary structure is thermodynamically favored (hence more stable) if its formation releases energy (ΔG is negative, i.e. negative free energy). Conversely, a secondary structure becomes thermodynamically unfavorable (hence less stable) if its formation requires energy (ΔG is positive, i.e. positive free energy). This fact is used to predict the secondary structure of a particular sequence. Free energies are additive, so one can determine the total free energy of a secondary structure by adding all the component free energies (as kcal/mole).

Given the importance of RNA secondary structure, a number of prediction algorithms have been developed and are available online to analyze an RNA sequence to predict its putative secondary structure. Some of the publicly available online tools for RNA secondary-structure prediction are listed in Table 7.2.

Secondary-structure-predicting algorithms often generate an output made up of brackets and dots (sometimes brackets and hyphens). The character string denoted by brackets and dots represents the number of residues of the input sequence and their base-pairing status. In the bracket notation, the base pairs are indicated by opening and closing parentheses. Some program outputs have these brackets and dots above the bases. Some program outputs may contain the base-pairing probability as well (Figure 7.8).

RNA secondary-structure prediction based on thermodynamic parameters has been in practice since the 1980s. Such predictions owe their success to the application of various experimentally verified thermodynamic parameters. However, like every other method, thermodynamic predictions have their limitations. In

TABLE 7.2 Some Online Tools for RNA Secondary-Structure Prediction

Online Analysis Tool	Comments and URL
RNAfold	RNAfold predicts secondary structures of single-stranded RNA or DNA sequences based on the classic minimum-free-energy algorithm of Zuker and Stiegler[42] as well as the partition-function algorithm of McCaskill.[43] Current limits are 10,000 nt for minimum-free-energy-only predictions and 7500 nt for partition-function calculations. The server function can be tested using the sample sequence provided[44] (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi)
RNAsoft	RNAsoft is a collection of online services for the computational prediction and design of RNA/DNA structures based on a standard free-energy model.[45] The underlying algorithms have been designed and implemented by members of the Bioinformatics, Empirical and Theoretical Algorithmics (BETA) Lab at the Department of Computer Science of the University of British Columbia (http://www.rnasoft.ca/)
CONTRAfold	CONTRAfold is a novel secondary-structure prediction method based on conditional log-linear models (CLLMs), a flexible class of probabilistic models with high prediction accuracy[46] (http://contra.stanford.edu/contrafold/server.html)
RNAstructure	RNAstructure uses several secondary-structure prediction algorithms, including thermodynamic and partition-function algorithms. It is a complete package for RNA and DNA secondary-structure prediction and analysis. It can take different types of experiment mapping data to constrain or restrain structure prediction[47] (http://rna.urmc.rochester.edu/RNAstructureWeb/)
IPKnot	IPKnot performs integer-programming (IP)-based prediction of RNA pseudoknots. IPknot can also predict the consensus secondary structure when a multiple alignment of RNA sequences is given[48] (http://rna.naist.jp/ipknot/)
CYLOFOLD	RNA secondary-structure (including pseudoknot) prediction tool. Some examples of RNA sequences are provided that can be used to perform a test run. The bracket notation output is in brackets and hyphens instead of brackets and dots* (http://cylofold.abcc.ncifcrf.gov/)
CentroidHomfold and CentroidFold	CentroidHomfold predicts the secondary structure of an input RNA sequence by employing automatically collected homologous sequences of the target[49,50] CentroidFold uses the CONTRAfold model as the default setting to calculate base-pairing probabilities, and predicts RNA secondary structure using a γ-centroid estimator. Currently, the input sequence should be less than or equal to 2000 bases[51] (http://www.ncrna.org/)
pknotsRG	pknotsRG is a tool for predicting RNA secondary structures, including the class of simple recursive pseudoknots. It uses the thermodynamic energy model extended by some pseudoknot-specific values.[52] The program on the BiBiserv is limited to sequences of length up to 800 bases (http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/submission.html) pknotsRG will be discontinued and replaced by pKiss in the near future (http://bibiserv2.cebitec.uni-bielefeld.de/pkiss)

*Made available by Dr Bruce A. Shapiro and his research group at the National Cancer Institute, Frederick, MD.

order to circumvent this problem, various probabilistic and statistical models have been developed that seemingly outperform thermodynamic-parameter-based predictions.[54] Figure 7.8A shows secondary-structure prediction of the input RNA sequence based on minimal-free-energy (MFE) calculation by pknotsRG-MFE. Figure 7.8B shows secondary-structure prediction of the input RNA sequence based on the partition functions and base-pair probabilities model[m] by IPKnot; the output is the McCaskill model. In contrast, Figure 7.8C shows an alternative output by IPKnot, based on a conditional log-linear probabilistic model known as CONTRAfold.[46] The figure also shows the respective bracket notations of each model. The free energy of a secondary structure is calculated by summing energy parameters of respective loop substructures, which can be experimentally determined and computationally estimated.[55]

7.7 MICROARRAY ANALYSIS

Most researchers doing microarray experiments use the analysis software provided by the manufacturer of

[m]Partition functions estimate statistical properties of a system with respect to thermodynamic probabilities, such as melting behavior and base-pair probabilities; properties and probabilities of a myriad of alternative structures in thermodynamic equilibrium.

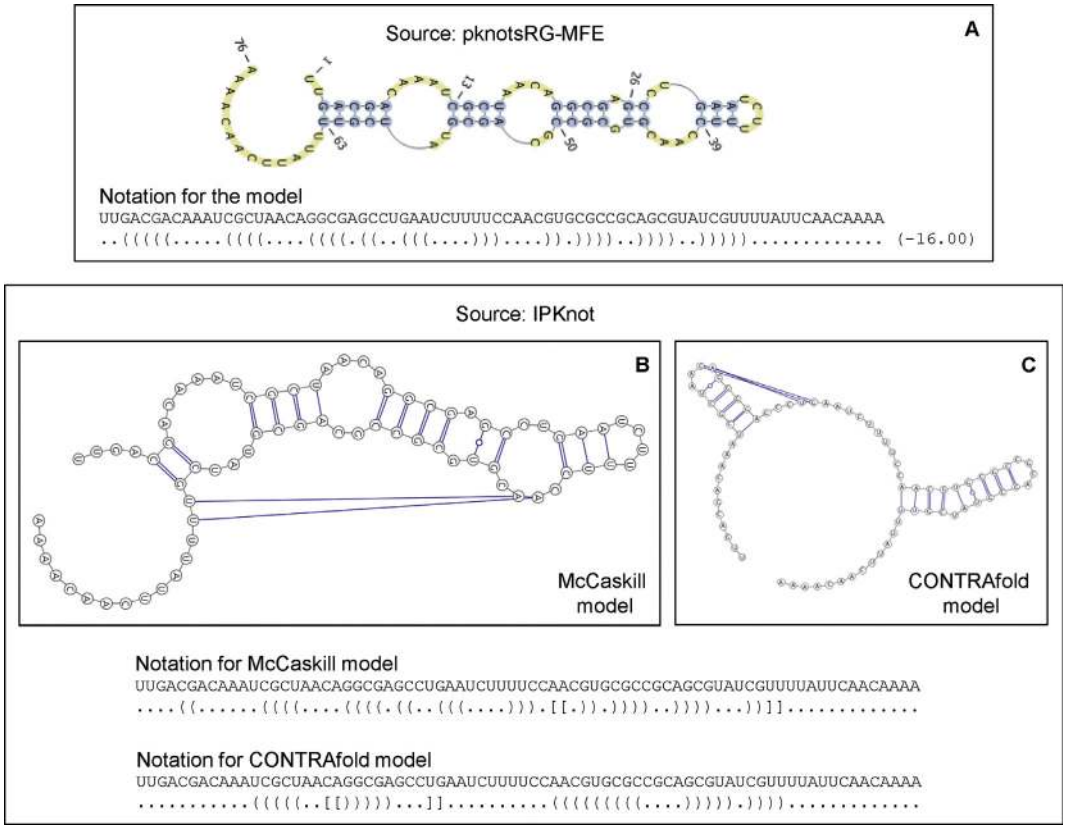


FIGURE 7.8 **RNA secondary-structure prediction by two web-based programs using default parameters.** (A) Prediction using pknotsRG-MFE of the Bielefeld University Bioinformatics Server (BiBiServ).[53] (B and C) Integer-programming (IP)-based prediction using IPKnot of the Nara Institute of Science and Technology, Japan. The default is the McCaskill model shown in (B); an alternative is the CONTRAfold model shown in (C). The respective bracket notations are also shown. In the bracket notation, the base pairs are indicated by opening and closing parentheses. Residues not involved in base pairing are denoted by dots. Every base with a "(" notation below is base-paired with a downstream base with a ")" notation below it. Some program outputs may also contain the base-pairing probability.

the microarray platforms. Therefore, some basic concepts of microarray data analysis are discussed here.

An outline of the microarray technique has been discussed in Chapter 3. The system described is also called **two-color** or **two-channel microarrays** because it involves the use of two different fluorescently labeled probes; one labeled with the fluorescent dye Cy3[n] (fluorescein, with fluorescence emission at ~565 nm; hence green), and the other labeled with the fluorescent dye Cy5 (biotin, with fluorescence emission at ~665 nm; hence red). The goal of DNA microarray is to screen the expression profile of genes, and the technique is useful because of its high-throughput nature.

Scanning of the microarray slide is the first step following post-hybridization processing and drying. The slide is scanned by a laser scanner hooked to a confocal laser microscope. The laser excites each spot in the microarray and the fluorescence emission is captured through a photomultiplier connected to the confocal laser microscope. The scanning is done in both green and red channels (at both wavelengths), each producing an individual image. The individual images are merged to obtain a composite image, in which the spot images can be green, red, or yellow; yellow means there are equal amounts of green and red fluorescence. However, the color of all the spots may not be perfectly green, red, or yellow, and may show a range, such as black/dark blue, blue, green, yellow, orange, and red. The image is usually reported as the ratio of Cy5 and Cy3 fluorescence intensity.

The next step is **image processing**. The features on the array—that is, what is contained in each grid/spot—are already defined. The image captured is a

[n]Cy3 (cyanine 3) dye is red (dark pink) in color and Cy5 (cyanine 5) dye is blue in color. However, the absorption and florescence emission maxima for Cy3 are ~547 and ~565 nm, respectively, whereas those of Cy5 are ~647 and ~665 nm, respectively. Hence, Cy3 is detected as green florescence in the green channel, and Cy5 is detected as red florescence in the red channel. Therefore, the physical colors of these dyes are not to be confused with their fluorescence emission colors.

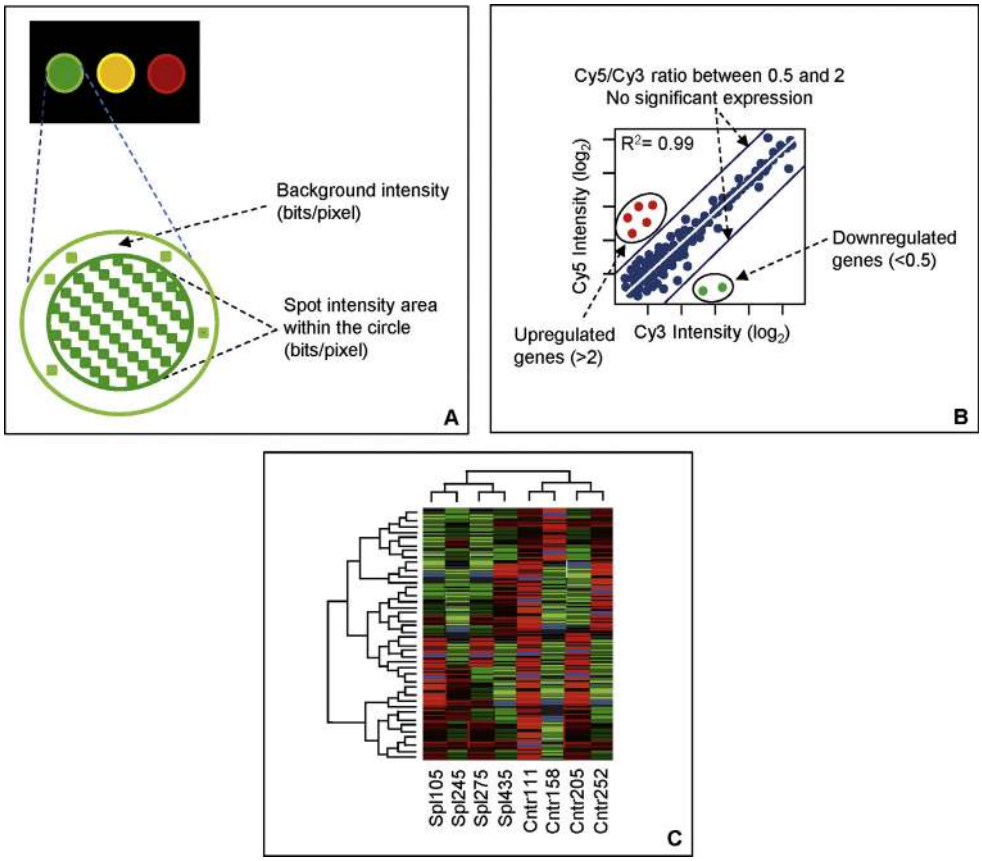


FIGURE 7.9 **Microarray image normalization and clustering.** (A) The captured microarray image is digital in nature. A digital image is composed of pixels, its smallest individual elements; each pixel has a value that represents the brightness of a given color at a point. Microarray scanners typically capture the color images as 16 bits/pixel. Therefore, the higher the bits/pixel, the greater is the color depth. For each spot, the true signal intensity is determined by subtracting the median background value. (B) Following image processing, the data are normalized in order to adjust for differences in labeling and detection efficiencies for Cy5 and Cy3. In the Lowess (locally weighted scatterplot smoothing; regression) method of normalization, it is assumed that mRNAs from closely related samples should cluster, producing a straight line in a scatter plot of Cy5 versus Cy3 intensities (or their $\log_2$ values), with a slope value close to 1. If such linearity is missing, the data are normalized to create the desired slope. If the cutoff for significant changes in expression is set at 2, the values ranging between 0.5 and 2 are not considered to be significant. (C) Hierarchical clustering dendrogram and heat map commonly used to display microarray data. The dendrogram represents relationships amongst genes and the branch lengths represent the degree of similarity in terms of their expression. In this method, using a distant matrix method, the algorithm first joins the two closest genes into a cluster; then the next most similar genes are joined together, and so on. This repetitive agglomeration first creates smaller clusters, which are similarly joined to form larger clusters. This process continues until all of the genes are joined into one giant cluster.

digital image, which is a rectangular array of intensity values in the spot; each intensity value is a pixel. The color depth is expressed as bits/pixel; hence the higher the bits/pixel, the greater is the color depth. During image processing, the spot boundaries are defined so that the true signal and the background values can be assigned. The median background value is then subtracted to obtain the true signal value (Figure 7.9A). The true signal is the fluorescence intensity due to specific hybridization, whereas the background signal is the fluorescence intensity due to non-specific hybridization that has survived post-hybridization washing, as well as non-specific binding of the fluorescently labeled nucleic acid fragments to a "sticky" surface, or even any dirt on the slide.

The next step is **data normalization**. Following image processing and analysis, the data are normalized. The purpose of normalization is to adjust for differences in labeling and detection efficiencies for Cy5 and Cy3, as well as to adjust for any differences in the RNA samples. Without normalization, the Cy5/Cy3 ratio could be artificially skewed. Normalized samples are ready for further analysis. Normalization can be done by (1) the total intensity normalization method, (2) the regression method, or (3) the ratio statistics method. The regression method is called the "**Lowess**" (locally weighted scatterplot smoothing) method, which is a locally weighted linear regression used to estimate systemic biases in the data. In the regression method, which is often used, it is assumed that mRNAs from closely related samples should be

expressed at similar levels. Under this assumption these mRNAs should cluster, producing a straight line in a scatter plot of Cy5 versus Cy3 intensities (or their $\log_2$ values). The scatter plot is thus a **ratio-intensity (R-I) plot**. If the labeling and detection efficiencies were the same for both samples, the slope of the scatter plot should be 1 or close to 1. If such linearity is missing, Lowess normalizes the data to create the desired slope. Normalized data are then used to report the expression ratios of genes between the samples, such as between the control and the experimental sample, or between normal and disease tissue samples. The cutoff for significant changes in expression can be set at 2—that is, values ranging between 0.5 and 2 are not considered to be significant. In this scenario, >2-fold difference means significant upregulation of expression, and <0.5-fold difference means significant downregulation of expression. However, these can be adjusted depending on the experiment, as well as the variability of the data (Figure 7.9B).

Cluster analysis of microarray data is a very widely used way to demonstrate gene-expression differences between the objects being studied, such as normal versus diseased tissue, control versus treatment group. Because genes involved in a common pathway, genes that are coordinately regulated, and genes involved in similar physiological response may be expressed similarly, the expressions of these genes are related. Microarray expression data can be used to find the relationships between genes in terms of their expression and consequently categorize such genes. This method is called cluster analysis. Therefore, in cluster analysis, the genes that are upregulated or downregulated in response to a specific condition (exposure, disease), can be identified and the biological relevance of such gene expression can be further investigated. Additionally, such gene expression can also be used as a biological marker of specific physiological response. Clustering can be supervised or unsupervised. In **supervised clustering**, the expression pattern of the gene(s) is known and this knowledge is used to group genes into clusters. In **unsupervised clustering**, there is no prior knowledge regarding the expression pattern of the gene(s) in a specific condition. Similar expression profiles are then connected to form the groups until all expression data have been included.

The most widely used method of unsupervised clustering is known as **hierarchical clustering**. Hierarchical clustering is commonly used in microarray as well as in phylogenetic analysis because it computes a tree (dendrogram). In DNA microarray analysis, the tree represents relationships amongst genes and the branch lengths represent the degree of similarity in terms of their expression. *Hierarchical clustering is a bottom-up agglomerative approach*. In this method, the algorithm starts by calculating the pairwise distance matrix for all of the genes in the so-called "gene space." Next, the algorithm joins the two genes that are the closest into a cluster. If there are multiple gene pairs that share the same degree of similarity, then the first cluster is formed based on some predetermined rule. Then, the next most similar genes are joined together, and so on. Once the small clusters are formed, the algorithm computes the pairwise distance matrix for all of the clusters in the so-called "cluster space." Next, the algorithm joins the two small clusters that are the closest into a larger cluster. This repetitive agglomeration process continues until all of the genes are joined into one giant cluster (Figure 7.9C). The other means of unsupervised clustering is known as ***k*-means clustering**. Contrary to the hierarchical clustering, *k-means clustering is a top-down divisive approach*. Obviously it does not produce dendrograms; instead, in this method data are partitioned into a prespecified set of *k*-clusters. Another divisive clustering method based on neural networks is **self-organizing maps** (SOM). The *k*-means clustering and SOM methods will not be further discussed here.

The **TM4 suite** of tools (http://www.tm4.org/)[56] consists of four major applications, Microarray Data Manager (MADAM), The Institute for Genomic Research (TIGR) Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV). **TIGR Spotfinder** is a microarray image-processing and quantification tool, whereas **TIGR's MIDAS** is a normalization and filtering tool. Another microarray image-analysis tool, **ScanAlyze**, is provided by the Eisen Lab at http://rana.lbl.gov/EisenSoftware.htm. The same link at Eisen Lab also provides **Cluster** and **TreeView**, which are cluster-analysis and graphical visualization software tools. They can perform hierarchical clustering, self-organizing maps (SOMs), *k*-means clustering, and principal component analysis.[57] Another web server for the normalization and standardization of DNA microarray data is **SNOMAD**[o] (http://pevsnerlab.kennedykrieger.org/snomadinput.html), made available by the Pevsner Lab at Johns Hopkins University School of Medicine.

7.8 DETECTION OF SEQUENCE POLYMORPHISM AND THE SNP DATABASE

Mutations can be point mutations, small deletions and insertions, or large-scale changes in the chromosome. Point mutations can be common or rare types of mutations. By definition, a point mutation that occurs

[o]© 2000 by Carlo Colantuoni, George Henry, and Jonathan Pevsner.

in at least 1% of the population is called a **single nucleotide polymorphism** (**SNP**; pronounced "**snip**").

SNPs constitute a very important class of mutations; they generally occur at a frequency of at least 0.1% (1/1000 bases) in the genome but may occur more frequently in certain regions. In the human genome, >65% of all SNPs involve C→T transition mutations. A set of linked SNPs that tend to inherit together as a unit is referred to as **SNP haplotype**. SNPs can occur in both coding and noncoding regions of genes. SNPs in the coding region may alter the characteristics of the protein while SNPs in the regulatory regions may alter the expression profile of genes.

Some SNPs can predispose people to disease or influence their response to a drug. For example, two SNPs in the *ApoE* gene result in three possible alleles of the gene: *E2*, *E3* (wild type), and *E4*. The corresponding protein product of each gene differs by one amino acid ($ApoE2^{C112,C158}$, $ApoE3^{C112,R158}$, $ApoE4^{R112,R158}$). Individuals inheriting two *E4* alleles have the highest chance of getting Alzheimer's disease, while those inheriting two *E2* alleles are the least likely to get the disease; so the order of risk associated with various *ApoE* alleles is *E4* > *E3* > *E2*. Apparently, one amino acid change in the ApoE protein alters its structure and function enough to influence the risk of disease development associated with each allele.[58]

The **International HapMap Project** is a multi-country (USA, UK, Canada, Japan, China, and Nigeria) effort to identify and catalog genetic similarities and differences in human beings. In doing so, the project expects to identify and catalog SNPs and SNP haplotypes that confer susceptibility/resistance to disease or therapy.

Sequence polymorphisms can be detected through pairwise alignment of two DNA sequences from two individuals. Deep resequencing of specific regions of the genome can also identify sequence polymorphisms.

The NCBI SNP database (**dbSNP**; http://www.ncbi.nlm.nih.gov/projects/SNP/ or http://www.ncbi.nlm.nih.gov/snp/) is the largest public database of short genetic variations (SNVs). The dbSNP is a broad collection of simple genetic polymorphisms, which includes single-base nucleotide substitution (**SNPs**), small-scale multi-base deletions or insertions (deletion–insertion polymorphisms or **DIPs**[P]), and retroposable element insertions and microsatellite repeat variations (also called short tandem repeats or **STRs**). Each dbSNP entry includes the sequence context of the actual polymorphism, such as the surrounding sequence; the occurrence frequency of the polymorphism (by population or individual); and the experimental method(s), protocols, and conditions used to assay the variation.[60]

A new submission to dbSNP is assigned a unique **ss#** (**submitted SNP ID number**). The submission is verified by alignment to the appropriate genomic contig. If several ss# entries map to the same position, the records are merged into a cluster that is given a unique **rs#** (**reference SNP cluster ID**).

A search was made for the mouse *Slco1a6* gene in dbSNP. The search produced 2092 hits as of July 2013 (Figure 7.10).

Selecting "Summary" from the "Display Settings" drop-down menu returns the summary of information on that SNP (figure not shown). Selecting "Graphic Summary" from the drop-down menu returns the display shown in Figure 7.10. Clicking "rs266211819" returns its **cluster report**. The top portion of the cluster report is shown in Figure 7.11A. The "Variation Class" field shows that it is a single nucleotide variation (SNV), the "RefSNP Alleles" field shows that the SNV is either A or C (circled). In other words, one of the alleles would be termed the "A" allele and the other allele would be termed the "C" allele, and the SNP is located on the "forward strand" ("Fwd"; circled). The information is organized into a few sections, such as GeneView, Map, etc. Figure 7.11B shows that rs266211819 is an intronic SNP. Clicking "view" in the "Neighbor SNP" field (circled in Figure 7.11A) shows that there are two SNPs within 100 bases upstream and four SNPs within 100 bases downstream of rs266211819 (Figure 7.12).

Figure 7.13 shows the graphic view of SNP rs266211819.

The SNP cluster page also has a section on the submitted SNP ID number (ss#) (Figure 7.14A). The ss370364874 has the longest flanking sequence and is shown. Clicking on the ss# (Figure 7.14A; circled) returns the details of the submitted SNP (Figure 7.14B). In the left-hand top corner there is "Submitter" information. The "Handle" field provides the submitter information. Clicking "SC_MOUSE_GENOMES" reveals the submitter contact information. In this case, the submitter is from the Wellcome Trust Sanger Institute, Cambridge, UK. In the right-hand top corner is "Resource Links." The submission can be viewed by clicking the "view" field (circled). Figure 7.15A shows the details of the original submission, including the SNP (A/C) as well as the 5′- and 3′-flanking sequences. Note that the original submission shows the SNP as A/C, but in the NCBI cluster report (the FASTA sequence part from the cluster report is displayed in Figure 7.15B) this (A/C) is replaced by M. This substitution of the original SNP is done following the IUPAC (International Union of Pure and Applied Chemistry) nucleotide codes shown in Table 7.3.

[P]DIP (deletion–insertion) or indel (insertion–deletion) polymorphisms consist of the presence or absence of short sequences (typically 1–50 bp).[59]

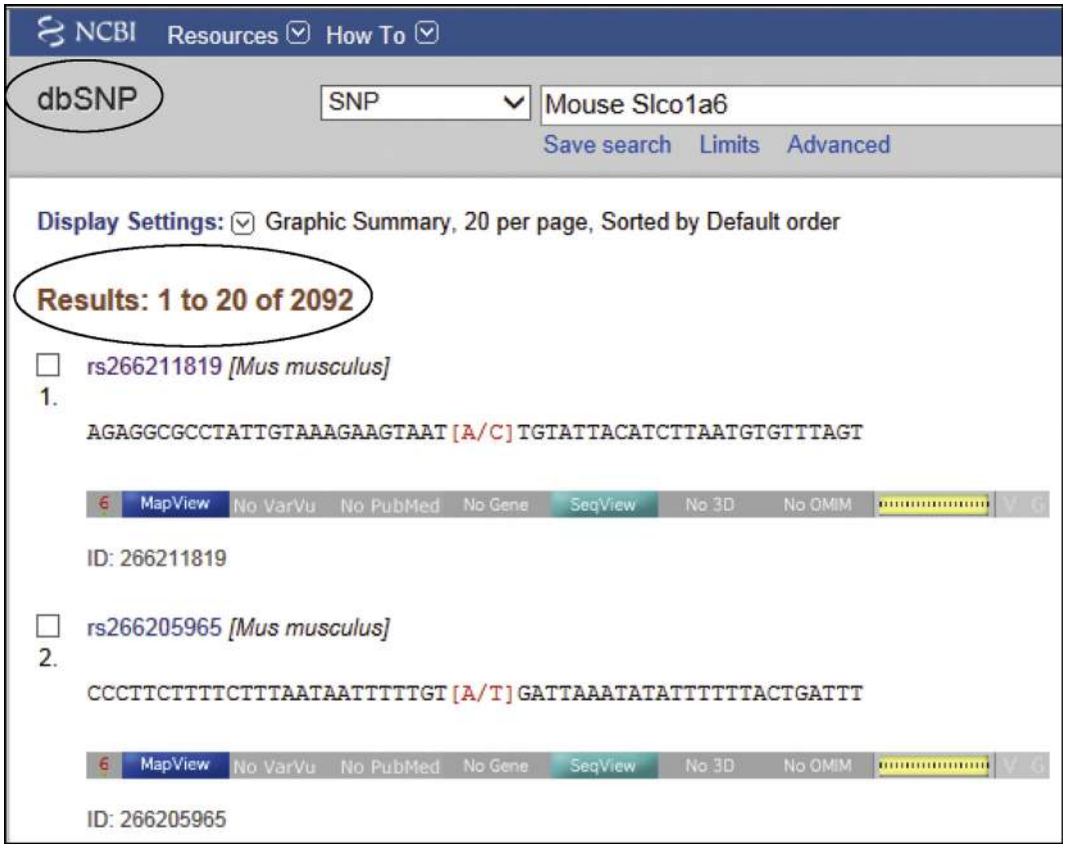


FIGURE 7.10 **A search for the mouse *Slco1a6* gene in the SNP database.**

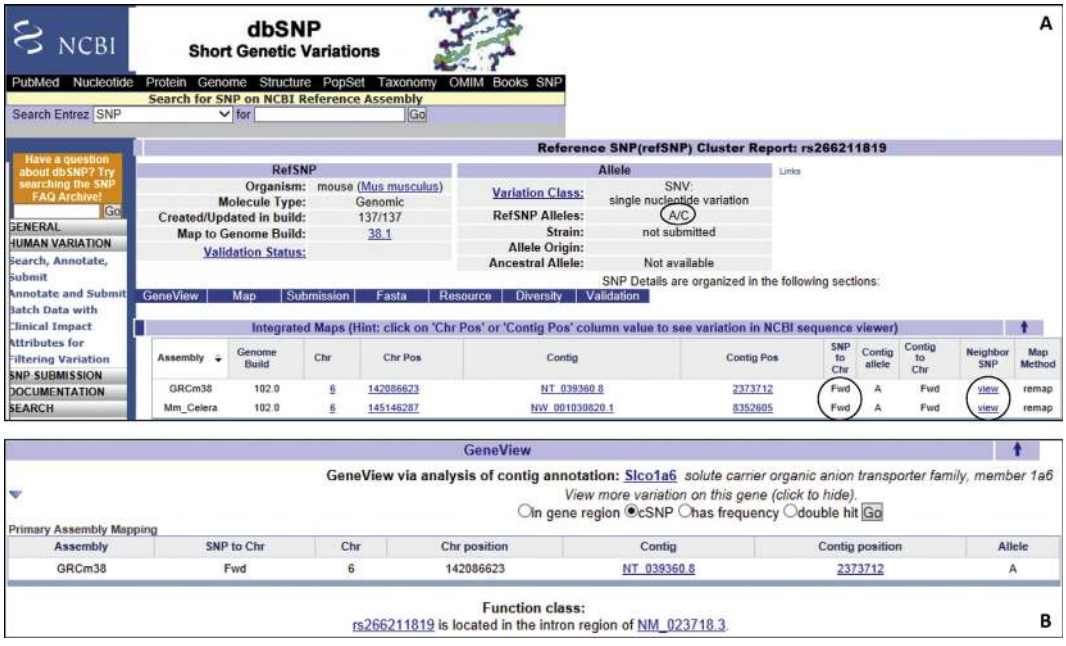


FIGURE 7.11 **Clicking the first entry rs266211819 returns its cluster report.** (A) The top portion of the cluster report is shown, see text for explanation; (B) GeneView shows that the rs266211819 is an intronic SNP.

Neighbor (within 100 bases) SNP for rs266211819:

distance (base)	rs	map weight	validation	assembly	Contig accession	Contig position
-76	rs240960243	1		Mm_Celera	NW_001030820.1	8352529
-62	rs244941523	1	N.D.	Mm_Celera	NW_001030820.1	8352543
0	rs266211819	1	N.D.	Mm_Celera	NW_001030820.1	8352605
34	rs228719522	1	N.D.	Mm_Celera	NW_001030820.1	8352639
54	rs215448038	1	N.D.	Mm_Celera	NW_001030820.1	8352659
72	rs236998816	1	N.D.	Mm_Celera	NW_001030820.1	8352677
74	rs254203577	1	N.D.	Mm_Celera	NW_001030820.1	8352679

Note:
- When distance is negative, it means the neighbor snp is upstream to the rs266211819.
- When distance is 0 and the snp is other than rs266211819, then it means these snps hit on the same contig position on the assembly, which means:
A: These rs will be merged in future builds.
B: Some snp that hit the same positions are not merged because they have different variation class.

FIGURE 7.12 **Neighboring SNPs of rs266211819.** Information retrieved by clicking "view" in the "Neighbor SNP" field circled in Figure 7.11A, showing six flanking SNPs.

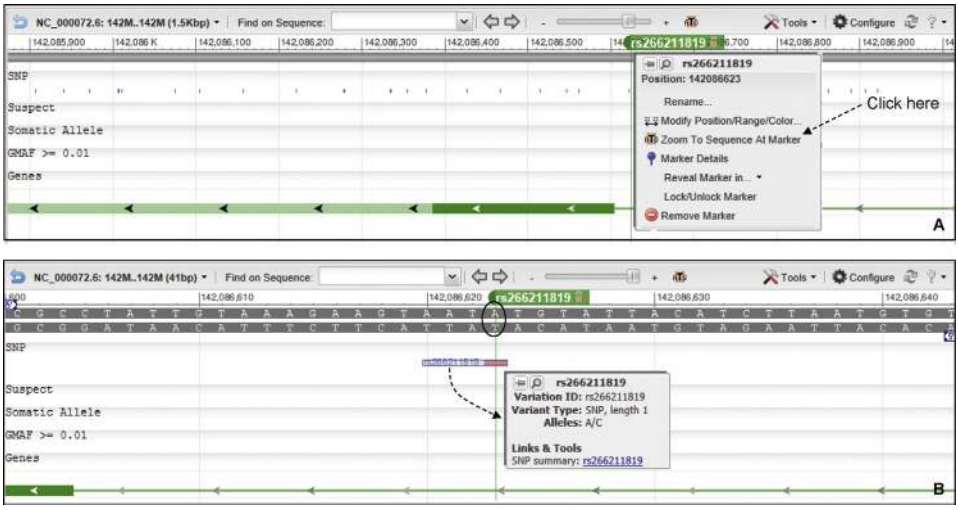


FIGURE 7.13 **The graphic view of rs266211819.** (A) Holding the cursor next to the green bar with the rsID (rs266211819) produces a drop-down menu. (B) Selecting "Zoom to Sequence At Marker" from this drop-down menu returns the sequence and the SNP. Selecting the bar with the rs# returns the drop-down menu shown. The drop-down menu contains information about the SNP (A/C).

A

Submitter records for this RefSNP Cluster

The submission **ss370364874** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs266211819** during BLAST analysis for the cur

NCBI Assay ID	Handle|Submitter ID	Validation Status	ss to rs Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date	Update Date	Build Added	Molecule Type
ss370364874	SC_MOUSE_GENOMES|MGP_WTSI_6_142035144		fwd/	A/C	agatagaggcgcctattgtaaagaagtaat	tgtattacatcttaatgtgtttagtgaaag	04/20/11	04/20/11	137	Genomic

B

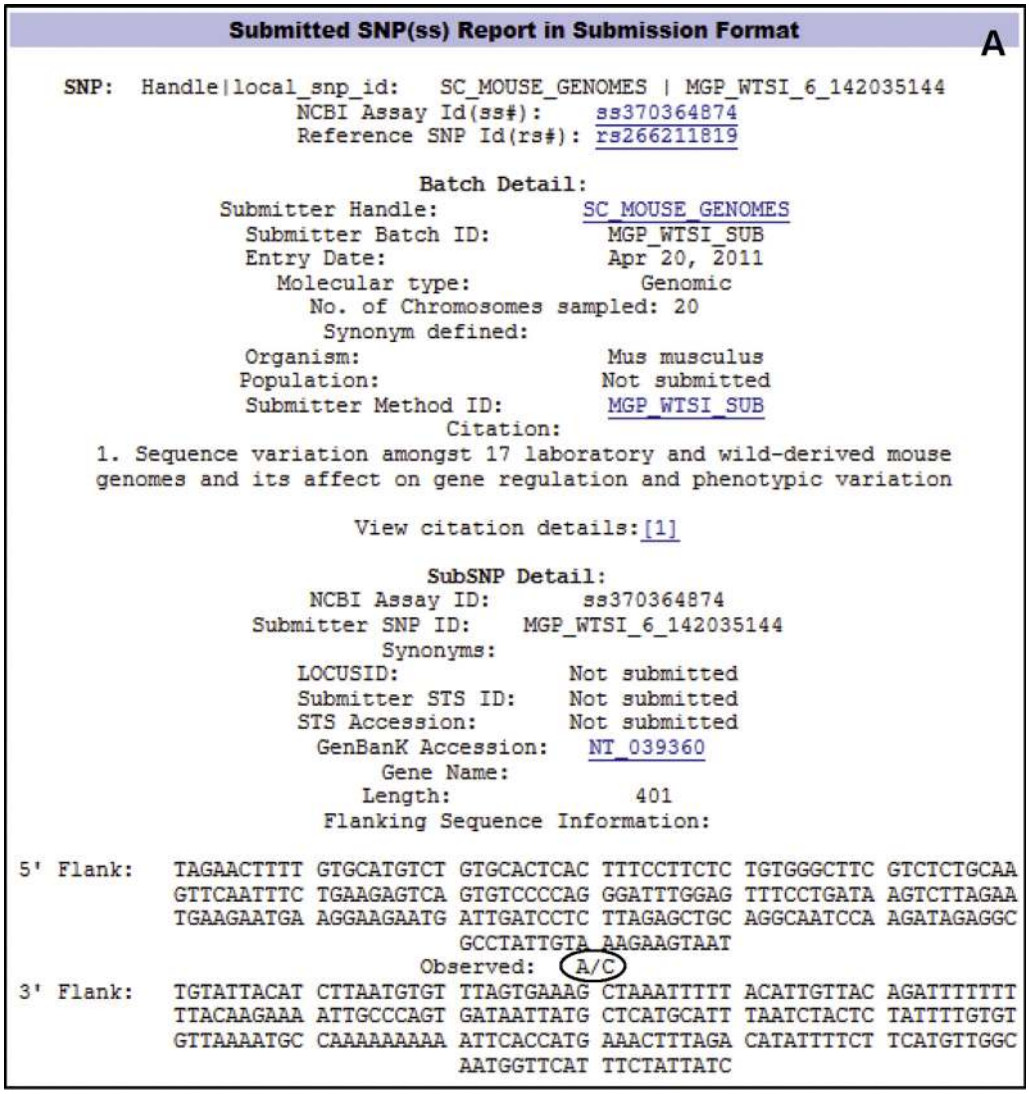
Submitted SNP(ss) Details: ss370364874

Submitter		Resource Links	
Handle	SC_MOUSE_GENOMES	GenBank Accession	NT_039360
Submitter SNP ID	MGP_WTSI_6_142035144	Submitted Gene Name	N.D
RefSNP(rs#)	rs266211819	Submitted Gene ID	N.D.
Submitted Batch ID	MGP_WTSI_SUB	Submitted Synonyms	N.D
Submitted Date	Apr 20, 2011	Submitted linkout	N.D
Publication Cited	[1] Sequence variation amongst 17 laboratory and wild-derived mouse genomes and its affect on gene regulation and phenotypic variation	Submission report	view
First entry to dbSNP	Apr 20 2011 12:00:00:000AM		

Assay		Allele	
Species	Mus musculus	Observed Allele	A/C
Molecular Type	Genomic	Ancestral Allele	N.D.
Method	MGP_WTSI_SUB	Allele Origin	N/A
Ascertainment Samplesize	20	SNP Class	SNV
Population	N.D.	CpG Code	N.D.

Validation		Variation	
Validation Status	Not Validated	Frequency Submission	N.D.
HWE Goodness of Fit	not applicable	Genotype Summary	N.D.
Homozygote Detected		Genotype Submission	N.D.
PCR Confirmed		Haplotype	N.D.
In Expressed Sequence			

FIGURE 7.14 **Submitter information for a SNP ID number.** See text for details.



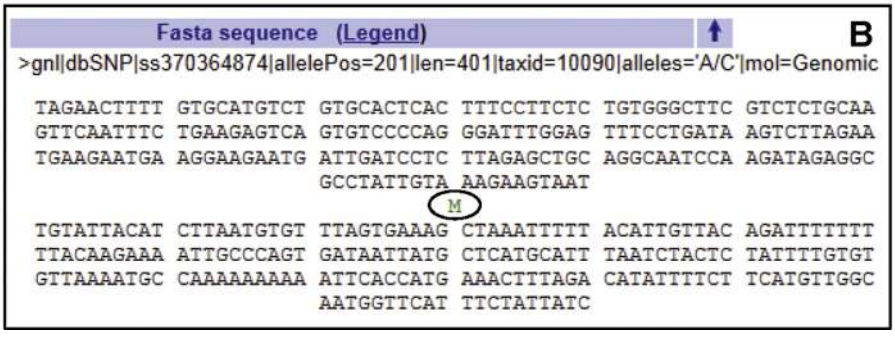


FIGURE 7.15 **IUPAC designation of the SNP in the database.** (A) The original submission showing the SNP (A/C) and the flanking sequence. (B) The substitution of A/C by M in the SNP database following the IUPAC nucleotide codes, as shown in Table 7.3.

TABLE 7.3 IUPAC Codes for Nucleotides

A = adenine	**T = thymine**	**G = guanine**	**C = cytosine**		
R = A/G	Y = C/T	S = G/C	W = A/T	K = G/T	M = A/C
B = C/G/T	D = A/G/T	H = A/C/T	V = A/C/G	N = any base	

/ means "or" (e.g. A/G means A or G)

References

1. Lander ES, Waterman MS. *Genomics* 1988;**2**:231−9.
2. Pevzner P, Shamir R, editors. *Bioinformatics for biologists*. Cambridge University Press; 2011.
3. Miller JR, et al. *Genomics* 2010;**95**:315−27.
4. Nagarajan N, Pop M. *Nat Rev Genet* 2013;**14**:157−67.
5. Compeau PEC, et al. *Nat Biotechnol* 2011;**29**:987−91.
6. Magoč T, Salzberg SL. *Bioinformatics* 2011;**27**:2957−63.
7. Baker M. *Nat Methods* 2012;**9**:333−7.
8. Yandell M, Ence D. *Nat Rev Genet* 2012;**13**:329−42.
9. Kapustin Y, et al. *Biol Direct* 2008;**3**:20.
10. Wheelan SJ, et al. *Genome Res* 2001;**11**:1952−7.
11. Taft RJ, et al. *Bioessays* 2007;**29**:288−99.
12. Tocchini-Valentini GD, et al. *Proc Natl Acad Sci USA* 2011;**108**:4782−7.
13. Yang X, et al. *Genome Res* 2011;**21**:634−41.
14. Yoon B-J. *Curr Genomics* 2009;**10**:402−15.
15. Salzberg SL, et al. *Nucl Acids Res* 1998;**26**:544−8.
16. Stanke M, et al. *Nucl Acids Res* 2006;**34**:W435−9 (Web Server issue).
17. Burge C, Karlin S. *J Mol Biol* 1997;**268**:78−94.
18. Burset M, Guigó R. *Genomics* 1996;**34**:353−67.
19. Lander ES, et al. *Nature* 2004;**431**:931−45.
20. Alexandersson M, et al. *Genome Res* 2003;**13**:496−502.
21. Dewey C, et al. *Genome Res* 2004;**14**:661−4.
22. Wang L, et al. *Nucl Acids Res* 2004;**32**:3689−702.
23. Butler JEF, Kadonaga JT. *Genes Dev* 2002;**16**:2583−92.
24. Choudhuri S, et al. *DNA Seq* 2002;**13**:103−7.
25. Hewitt SC, et al. *Mol Endocrinol* 2012;**26**:887−98.
26. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 3−48.
27. Klassen JL, Currie CR. *PLoS ONE* 2013;**8**:e58387.
28. Münch R, et al. *Bioinformatics* 2005;**21**:4187−9.
29. Reese MG. *Comput Chem* 2001;**26**:51−6.
30. Knudsen S. *Bioinformatics* 1999;**15**:356−61.
31. Prestridge DS. *CABIOS* 1991;**7**:203−6.
32. Prestridge DS. *J Mol Biol* 1995;**249**:923−32.
33. Abeel T, et al. *Genome Res* 2008;**18**:310−23.
34. Down TA, Hubbard TJ. *Genome Res* 2002;**12**:458−61.
35. Blanchette M, Tompa M. *Nucl Acids Res* 2003;**31**:3840−2.
36. Pedersen AG, Nielsen H. *ISMB* 1997;**5**:226−33.
37. Nishikawa T, et al. *Bioinformatics* 2000;**16**:960−7.
38. Vincze T, et al. *Nucl Acids Res* 2003;**31**:3688−91.
39. Chen JL, et al. *Cell* 2000;**100**:503−14.
40. Svoboda P, Di Cara A. *Cell Mol Life Sci* 2006;**63**(7-8):901−8.
41. Li F, et al. *Cell Rep* 2012;**1**:69−82.
42. Zuker M, Stiegler P. *Nucl Acid Res* 1981;**9**:133−48.
43. McCaskill JS. *Biopolymers* 1990;**29**:1105−19.
44. Hofacker IL. *Nucl Acids Res* 2003;**31**:3429−31.
45. Andronescu M, et al. *Nucl Acids Res* 2003;**31**:3416−22.
46. Do CB, et al. *Bioinformatics* 2006;**22**:e90−8.
47. Reuter JS, Mathews DH. *BMC Bioinformatics* 2010;**11**:129.
48. Sato K, et al. *Bioinformatics* 2011;**27**:i85−93.
49. Hamada M, et al. *Bioinformatics* 2009;**25**:i330−8.
50. Frith MC, et al. *BMC Bioinformatics* 2010;**11**:80.
51. Sato K, et al. *Nucl Acids Res* 2009;**37**:W277−80 (Web Server issue).
52. Reeder J, et al. *Nucl Acids Res* 2007;**35**:W320−4 (Web server issue).
53. Reeder J, Giegerich R. *BMC Bioinformatics* 2004;**5**:104.
54. Rivas E. *RNA Biol* 2013;**10**:1−12.
55. Mathews DH, et al. *J Mol Biol* 1999;**288**:911−40.
56. Saeed AI, et al. *Biotechniques* 2003;**34**:374−8.
57. Eisen MB, et al. *Proc Natl Acad Sci USA* 1998;**95**:14863−8.
58. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 49−99.
59. Pepinski W, et al. *Mol Biol Rep* 2013;**40**:4333−8.
60. Kitts A, Sherry S. The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation (chapter 5). In: McEntyre J, Ostell J, editors. *The NCBI handbook*. Bethesda (MD): National Center for Biotechnology Information; 2011.

CHAPTER

8

Additional Bioinformatic Analyses Involving Protein Sequences*

OUTLINE

8.1 PROTEIN STRUCTURE

Proteins have four levels of structure: primary, secondary, tertiary, and quaternary.

Primary structure is simply the amino-acid sequence of the polypeptide, and is determined by the sequence of codons in the gene encoding the polypeptide. Therefore, the open reading frame (ORF)-prediction programs predict the primary structure of the encoded proteins.

Secondary structure is the hydrogen (H)-bonded three-dimensional local conformation. The two most common secondary structures are the **α-helix** and **β-pleated sheet**. In addition, four other commonly occurring secondary structures are the **3_{10}-helix**, **π-helix** (**pi helix**), **β-turn**, and **Ω-loop** (**omega loop**). There are still other regions in proteins whose secondary structure can not be classified under any established categories; these have been traditionally referred to as **random coils**, but can be more appropriately referred to as **unstructured regions**.

An α-helix (radius = 2.3 Å) is a right-handed helix that has 3.6 amino acids per helical turn (100° turn/residue),

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

and the structure is stabilized by H-bonds formed between the C═O of residue n and the N—H of residue n + 4; both these groups are part of the helical backbone and not the side chains (R groups) that protrude out of the backbone. The pitch of the helix (vertical distance in one complete helical turn) is 5.4 Å; hence, the rise per residue along the helix axis is 1.5 Å. In an α-helix, the H-bonds are intrachain and parallel to the axis of the helix. The α-helix is a **3.6_{13}-helix**, where 3.6 is the number of residues per turn and 13 is the number of atoms in the H-bonded loop. *The α-helix is the most abundant secondary structure found in globular proteins, and it accounts for 32—38% of all residues. The average length of an α-helix is 10 residues.*

A less common helical secondary structure found in proteins is the **3_{10}-helix** (radius = 1.9 Å), which has 3 amino acids per turn (120° turn/residue) and 10 atoms in the H-bonded loop. In a 3_{10}-helix, H-bonds involve residues n and n + 3 (instead of n + 4 as in the α-helix), and the backbone conformational angles are slightly different from those of the α-helix. The pitch of the helix is 6.0 Å; hence, the rise per residue along the helix axis is 2.0 Å. The length of the 3_{10}-helix may vary from 3 to 10 residues. The ideal 3_{10}-helix is rare and when it occurs, it tends to be at the C- and N-termini; the 3_{10}-helix has been described in channels and membrane proteins.[1]

Like the α-helix and 3_{10}-helix, the π-helix (radius = 2.8 Å) is also a right-handed helix. There are 4.4 residues per turn (81.8° turn/residue) and 16 atoms in the H-bonded loop; hence, the π-helix is a **4.4_{16}-helix**. The structure is stabilized by H-bonds formed between the C═O of residue n and the N—H of residue n + 5 (compared to n + 4 in the α-helix, and n + 3 in the 3_{10}-helix). The pitch of the helix is 4.8 Å; hence, the rise per residue along the helix axis is 1.1 Å. A π-helix can be derived from an α-helix by the insertion of a single amino acid. Such insertion tends to destabilize the α-helix. As a result, the formation of π-helix is tolerated only if it provides some selective advantage to the protein. One such possibility involves affecting the functional site of proteins. Consistent with this hypothesis, the π-helix is typically found near the functional site of proteins. About 15% of known protein structures contain a π-helix. Naturally occurring π-helices are typically 7—10 residues in length, but are mostly composed of 7 residues; they are found at the end of a regular α-helix or within an α-helix—that is, a π-helix is flanked by α-helices.[2]

Two or more (two to seven) α-helices can wrap around each other creating **coiled coils**, which are **superhelical** (**supersecondary**) **structures**. In most coiled coils, the α-helices are wrapped around each other into a left-handed helical supercoil. The α-helical coiled coil is a common structural motif in proteins that facilitate subunit oligomerization. Coiled coils can be composed of parallel or antiparallel helices. An example of a functional protein with coiled coils is the Fos-Jun heterodimer, known to regulate gene expression. Another example is tropomyosin. Each strand of a coiled coil has a repeat of seven residues (heptads; *a-b-c-d-e-f-g*). In these heptads, the first and the fourth residues (*a* and *d*) are hydrophobic; they face the helical interface and facilitate hydrophobic interactions. Good candidate amino acids at these positions are isoleucine, leucine, and valine. The other residues are hydrophilic and exposed to the solvent. Of these, the fifth and the seventh residues (*e* and *g*) confer specificity between the two helices through electrostatic interactions. Good candidate amino acids at these positions are the charged amino acids, such as aspartic acid, glutamic acid, lysine, and arginine. Discontinuities in the heptad pattern are quite frequent. Algorithms that predict coiled coils scan the sequence for the regular patterns and heptad signatures using a window size of 14, 21, or 28 amino acids.

In contrast to the helices, a β-pleated sheet (β-sheet) involves two or more polypeptide chains and the H-bonds are formed between residues that are part of different polypeptide chains. Therefore, in a β-pleated sheet, the H-bonds are interchain and are perpendicular to the polypeptide backbones. Each polypeptide chain involved in the formation of a β-pleated sheet is a **β-strand**; a β-pleated sheet can be two stranded or multi-stranded. As the name suggests, the β-pleated sheet has a zigzag appearance. *After the α-helix, the β-sheet is the major secondary-structural element in globular proteins, accounting for 20—28% of all residues.*

In a **β-turn** (also called β-bend) the direction of the polypeptide chain is sharply reversed. The name β-turn owes its origin to the fact that they often connect antiparallel β-sheets. *A β-turn is composed of four amino acids*[a]. The **Ω loop**, as a secondary-structural motif in globular proteins, was first described in 1986.[3] These are a six-amino-acid or longer backbone motif. The polypeptide reverses its direction over the course of this six- (or more) amino-acid-long, omega-shaped loop region[b].

The **tertiary structure** of a protein is the overall folded structure in three-dimensional (3D) space. The tertiary structure is formed by the interactions between

[a]Depending on the number of amino acids involved, other tight turns are named as the **δ-turn** (involves two amino acids), **γ-turn** (involves three amino acids), **α-turn** (involves five amino acids), and **π-turn** (involves six amino acids).[4]

[b]The existence of a variety of morphologies of loops (4 to 20 residues in length) as secondary-structural motifs has been reported in proteins, such as **strap loops** (linear), **omega loops** (nonlinear and planar), **zeta loops** (nonlinear and non-planar, i.e. globular).[5]

the side-chain R-groups, such as ionic interactions, hydrophobic interactions, H-bonds, and disulfide bonds. The amino-acid sequence (the primary structure) primarily dictates how a protein should fold into a 3D tertiary structure. However, proper folding is now known to be achieved with the help of **chaperone** molecules. In folded conformation (tertiary structure), most proteins contain specific **domains** that are discrete structural and functional units of the protein (discussed later).

Quaternary structure of proteins refers to the overall structure of multimeric proteins—that is, proteins composed of two or more subunits, each subunit being a monomer. Quaternary structures are stabilized by non-covalent interactions as well as disulfide linkages. Proteins with molecular weight >100 kD mostly contain more than one polypeptide chain, and hence have a quaternary structure.

The secondary, tertiary, and quaternary structures of proteins are maintained by non-covalent forces, such as H-bonds, electrostatic interactions, and van der Waals forces.

8.2 PEPTIDE BOND, PEPTIDE PLANE, BOND ROTATION, DIHEDRAL ANGLES, AND RAMACHANDRAN PLOT

Amino acids are linked together by peptide bonds. Peptide bonds are **amide linkages** between the $-NH_2$ and $-COOH$ groups of neighboring amino acids. The peptide bond (C—N) has a partial double-bond character. Thus, it is rigid and planar and not free to rotate. The plane on which it lies is called the **peptide plane** or **amide plane**. Peptide bonds are *trans* bonds—that is, the carbonyl oxygen and amide hydrogen are in *trans* position. However, the N—C_α and C_α—C bonds are not rigid and they can freely rotate, being only limited by the size and character of the R-groups. The angle of rotation (also called **torsion angle** or **dihedral angle**) around the N—C_α bond is called **phi** (φ) and that around the C_α—C bond is called **psi** (ψ) (Figure 8.1A). These two angles largely determine the 3D shape of the polypeptide backbone of the protein.

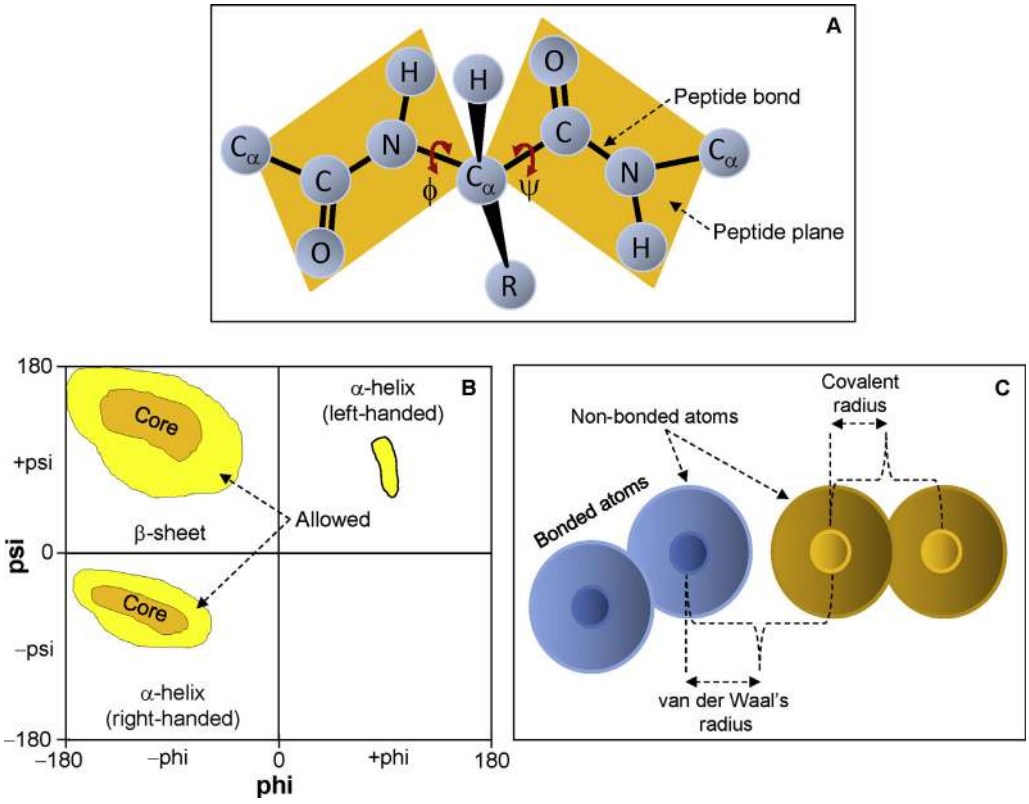


FIGURE 8.1 **Peptide bond, peptide plane, and the Ramachandran plot.** (A) Peptide bond, peptide plane, phi and psi angles, and bond rotation involving two amino acids. The N—C_α and C_α—C bonds are not rigid and can freely rotate, being only limited by the size and character of the R-groups. (B) Diagram of a typical Ramachandran plot (φ/ψ plot). The regions marked "Core" correspond to conformations that do not have any steric hindrance. The yellow areas labeled "Allowed" correspond to conformations that could be possible if the atoms could come a little closer together. The white areas represent conformations that are sterically unfavorable (see text). (C) In computing a Ramachandran plot, atoms are treated as hard spheres whose dimensions correspond to their van der Waals radii. The van der Waals radius and covalent radius are depicted for comparison.

Although φ and ψ are less restricted in terms of rotation, the bulkiness of R-groups of the amino acids tends to impose some restrictions on the rotation through steric hindrance. This makes certain combinations of φ and ψ preferred. The φ/ψ plot of the amino acid residues in a peptide is called the **Ramachandran plot**. It involves plotting the φ values on the *x*-axis and the ψ values on the *y*-axis to predict the possible conformation of the peptide. The angle spectrum in each axis is from −180° to +180°. In computing a Ramachandran plot, atoms are treated as hard spheres whose dimensions correspond to their van der Waals radii. Any angle that results in the collision of the spheres is regarded as sterically unfavorable; hence, such conformations are also sterically not allowed. Figure 8.1B shows a simplified diagram of a Ramachandran plot. The regions marked "Core" correspond to conformations that do not have any steric hindrance. The yellow areas labeled "Allowed" correspond to conformations that could be possible if slightly shorter van der Waals radii are used in the calculation. In other words, if the atoms could come a little closer together, then these conformations would be possible. The white areas represent conformations that are sterically unfavorable. The van der Waals radius and covalent radius are depicted in Figure 8.1C. The residues with a less bulky side chain or no side chain, such as glycine (no side chain), can have many possible combinations of φ and ψ (e.g. in a polyglycine backbone) resulting in a larger allowable area on the plot in all four quadrants, whereas residues with bulky side chains, such as proline or phenylalanine, have fewer possible combinations of φ and ψ, hence a smaller allowable area on the plot.

The φ and ψ angles for each residue in a helical structure are very similar, and that is what confers regularity to the helical structure. *Positive angles correspond to clockwise rotation and negative angles correspond to anticlockwise rotation*. The ideal values of φ/ψ were determined to be as follows: right-handed α-helix −57°/−47°; left-handed α-helix +57°/+47°; right-handed 3_{10} helix −74°/−4°; right-handed π-helix −57°/−70°; parallel β-sheet (uncommon) −119°/+113°; antiparallel β-sheet (common) −139°/+135°. The actual values differ somewhat from these idealized values. Recent experimental data have demonstrated that both φ and ψ can undergo large rotations, which are usually coupled. See Hovmöller, et al.[6] for more details on experimental determination of main-chain conformations in 1042 protein subunits.

Online tools are available from several sources for the analysis of Ramachandran plots of proteins. One such tool is available at the **Uppsala Ramachandran Server** (http://eds.bmc.uu.se/ramachan.html). This service is based on the Moleman2 program.[7]

8.3 PREDICTION OF PHYSICOCHEMICAL PROPERTIES OF A PROTEIN

The physicochemical properties of a protein can be deduced from its sequence. The **ExPASy** (Expert Protein Analysis System; http://www.expasy.org/) bioinformatics resource portal of the Swiss Institute of Bioinformatics (SIB) provides many protein-analysis tools. One such tool is **ProtParam**,[8] which analyzes the physicochemical properties of proteins based on the sequence. ProtParam can be accessed directly at http://web.expasy.org/protparam/, or it can be accessed by first accessing ExPASy, then clicking the "Resources A..Z" link on the left, and finding ProtParam from the resource list. Mouse Slco1a6 protein was analyzed in ProtParam; the results are presented and explained in Figure 8.2. *ProtParam analyzes the sequence as is and does not take into account any post-translational modifications*. The output parameters are explained in the "Documentation" link on the ProtParam home page (http://web.expasy.org/protparam/protparam-doc.html).

8.4 PREDICTION OF PROTEASE DIGESTIBILITY

The protease digestibility prediction tool in ExPASy is called **PeptideCutter**,[8] which can be accessed directly at http://web.expasy.org/peptide_cutter/. Alternatively, it can be accessed by first accessing ExPASy, then clicking the "Resources A..Z" link on the left, and finding PeptideCutter from the resource list. There is a list of many proteases on the PeptideCutter home page. Specific enzymes can be selected from this list to map their cleavage sites in the protein. For example, analyzing mouse Slco1a6 protein in PeptideCutter to find only the pepsin cleavage sites (at pH >2) revealed that there are a total of 179 such sites (not shown). PeptideCutter can return the output as table, as a map of cleavage sites on the sequence itself, or both. *The analysis output marks the amino acid residue; the actual cleavage occurs at the right-hand side (C-terminal side) of this marked residue*. PeptideCutter also predicts potential cleavage sites of some chemicals in a given protein sequence.

8.5 HYDROPHOBICITY, HYDROPHILICITY, AND ANTIGENICITY PREDICTION, AND THE HYDROPATHY PLOT

The R-group of an amino acid determines whether it is hydrophobic or hydrophilic. Hydropathy is a

Molecular weight: 74145.2,
Theoretical pI: 8.36

Amino acid composition:

Ala (A)	33	4.9%	Arg (R)	21	3.1%	Asn (N)	29	4.3%
Asp (D)	19	2.8%	Cys (C)	30	4.5%	Gln (Q)	13	1.9%
Glu (E)	36	5.4%	Gly (G)	57	8.5%	His (H)	8	1.2%
Ile (I)	58	8.7%	Leu (L)	73	10.9%	Lys (K)	42	6.3%
Met (M)	24	3.6%	Phe (F)	39	5.8%	Pro (P)	30	4.5%
Ser (S)	49	7.3%	Thr (T)	43	6.4%	Trp (W)	7	1.0%
Tyr (Y)	23	3.4%	Val (V)	36	5.4%			

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 74645
Abs 0.1% (=1 g/l) 1.007, assuming all pairs of Cys residues form cystines
Ext. coefficient 72770
Abs 0.1% (=1 g/l) 0.981, assuming all Cys residues are reduced

Estimated half-life:
The N-terminal of the sequence considered is M (Met)
The estimated half-life is:
30 hours (mammalian reticulocytes, in vitro)
>20 hours (yeast, in vivo)
>10 hours (Escherichia coli, in vivo)

Instability index:
The instability index (II) is computed to be 37.21
This classifies the protein as stable

Aliphatic index: 96.76

Grand average of hydropathicity (GRAVY): 0.267

FIGURE 8.2 **Partial ProtParam analysis output for Slco1a6.** The actual analysis contains more information. ProtParam analyzes the sequence as is and does not take into account any post-translational modifications. The **extinction coefficient** (E) indicates how much light a protein absorbs at a certain wavelength (e.g. 280 nm). It is useful to have an idea about the E value of a protein when purifying it. An approximate $E(Prot)_{280}$ = Tyr*E(Tyr) + Trp*E(Trp) + cystine*E(cystine); where E(Tyr) = 1490, E(Trp) = 5500, E(cystine) = 125 (cysteine does not absorb appreciably at wavelengths > 260 nm but cystine does). The approximate Abs_{280} = E(Prot)/MW (MW = molecular weight). For proteins rich in cysteines that do not form cystine (e.g. metallothionein), this calculation may have 10% or more error. ProtParam predicts an estimated **half-life** based on the "N-end rule," which relates the in vivo half-life of a protein to the identity of its N-terminal residues.[9] Note that ProtParam does not consider post-translational modifications, so the N-terminal-end-based rule does not account for any N-terminal modifications, which might significantly alter the predicted half-life. The **instability index** provides an estimate of the stability of the protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that the occurrence of certain dipeptides is significantly different in the unstable proteins compared with the stable ones.[10] Based on the statistically determined weight value of instability, an instability index can be calculated. An instability index value < 40 predicts the protein to be stable; a value > 40 predicts that the protein may be unstable. The **aliphatic index** (X) of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). X = X(Ala) + a*X(Val) + b*[X(Ile) + X(Leu)]; where X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100*mole fraction). The coefficients a and b are the volume of the valine side chain (a = 2.9) and of the Leu/Ile side chains (b = 3.9) relative to the side chain of alanine.[11] The **GRAVY** value for a peptide or protein is calculated as the sum of hydropathy values (Kyte and Doolittle) of all the amino acids, divided by the number of residues in the sequence. The hydropathy is discussed later in the chapter. A positive GRAVY value indicates that the protein is hydrophobic and a negative value indicates that it is hydrophilic.

measure of the hydrophobicity or hydrophilicity of an amino acid. Proteins are composed of both hydrophobic and hydrophilic amino acids, but the localization of these amino acids in the protein is related to the subcellular localization of the proteins (see Chapter 1 for a discussion on this subject). For example, proteins that are localized in an aqueous environment have hydrophobic amino acids (and their hydrophobic R groups) located towards the center of the molecule, away from water. In contrast, an integral membrane

protein always has a stretch of about 20 hydrophobic amino acids on the surface to enable it to pass through the membrane lipid bilayer. All hydrophilic amino acids are pushed to the outside of the membrane.

The hydropathy of amino acids is assigned specific values to create a **hydropathy scale**. There are different hydropathic scales; each scale assigns slightly different hydrophobicity or hydrophilicity values to the amino acids. Using a specific hydropathic scale the overall hydropathic character of a polypeptide can be determined, which is revealed by its **hydropathy plot**. Therefore, the hydropathy plot shows the hydrophobicity and hydrophilicity along the length of a polypeptide. Hydropathy is an important determinant of protein folding. One of the most widely used hydropathy plots is that of Kyte and Doolittle (1982).[12] The standard Kyte and Doolittle plot is a hydrophobicity plot. The plot is based on the consideration of the hydrophobic and hydrophilic properties of the 20 amino acids, shown in Table 8.1. Computation of the hydropathy plot requires setting a window size; the default is usually set at 7. The computation starts with the first window of amino acids (#1–7), the average hydrophobicity score of the first window is calculated and plotted as the midpoint of the window. Then the window moves by one amino acid, the second window spans amino acids #2–8, and the average hydrophobicity score of the second window is calculated and plotted as the midpoint of the window. This reiterative process continues until the last window at the end of the protein[c]. The averages are then plotted on a graph. The *y*-axis represents the hydrophobicity scores and the *x*-axis represents the window number/position of the amino acids. ExPASy provides **ProtScale**[8] (http://web.expasy.org/protscale/) that can be accessed to run the hydropathy plots. In addition to ExPASy, there are many more links providing online tools for the analysis of hydropathy plots of proteins. These links can be obtained by simply Googling the term.

In a hydrophobicity plot, hydrophilic amino acids receive negative values, whereas in a hydrophilicity plot, hydrophobic amino acids receive negative values.

Figure 8.3A shows the hydrophobicity plot of mouse Slco1a6 protein with a window size of 7. It is a transmembrane protein. Changing the window size to 21 clearly makes the transmembrane regions prominent (Figure 8.3B). A window size of 19 can also be used to visualize the transmembrane domains. Peaks above the line corresponding to 0 represent the hydrophobic regions and peaks below this line represent hydrophilic regions of the protein. The default window size in a Kyte and Doolittle plot is usually set at 7 or 9. An inverse Kyte and Doolittle plot will reverse these regions—that is, hydrophilic amino acids will be above the 0 axis and hydrophobic amino acids will be below the 0 axis.

Another widely used hydropathy plot, based on the Hopp and Woods hydropathy scale, is the Hopp and Woods hydrophilicity/antigenicity plot.[13] In this plot, hydrophilic amino acids get positive scores and hydrophobic amino acids get negative scores (Table 8.1). The Hopp and Woods hydropathy scale was developed for predicting potential antigenic sites in a polypeptide, which are likely to be rich in charged and polar residues. The default window size is usually set at 6 or 7; the regions of high hydrophilicity are likely to be antigenic sites. Figure 8.3C shows the Hopp and Woods plot of mouse Slco1a6 with a window size of 7.

TABLE 8.1 Hydrophobicity and Hydrophilicity Scores of Different Amino Acids

Amino Acid	Kyte–Doolittle	Hopp–Woods
Alanine	1.8	−0.5
Arginine	−4.5	3.0
Asparagine	−3.5	0.2
Aspartic acid	−3.5	3.0
Cysteine	2.5	−1.0
Glutamine	−3.5	0.2
Glutamic acid	−3.5	3.0
Glycine	−0.4	0.0
Histidine	−3.2	−0.5
Isoleucine	4.5	−1.8
Leucine	3.8	−1.8
Lysine	−3.9	3.0
Methionine	1.9	−1.3
Phenylalanine	2.8	−2.5
Proline	−1.6	0.0
Serine	−0.8	0.3
Threonine	−0.7	−0.4
Tryptophan	−0.9	−3.4
Tyrosine	−1.3	−2.3
Valine	4.2	−1.5

[c]Effective length of a polypeptide for hydropathy analysis = total # of windows of the desired size = total # of amino acids in the protein − window size + 1. For example, Slco1a6 has 670 amino acids. Hence, the effective length of Slco1a6 for hydropathy analysis = total # of windows of the desired size = 670 − 7 + 1 = 664. In other words, after the 664th amino acid, there are no more windows of 7 amino acids.

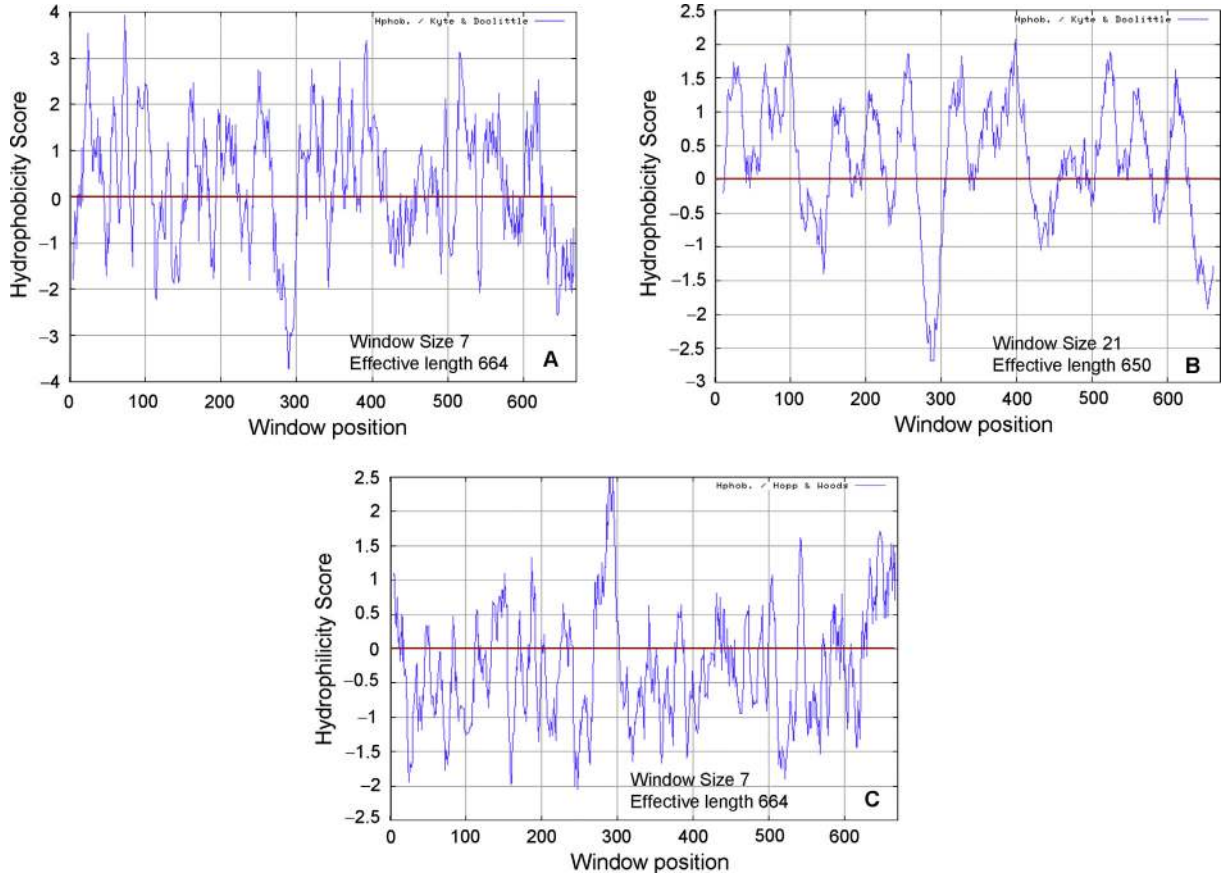


FIGURE 8.3 **Hydropathy plots.** Kyte and Doolittle plots and Hopp and Woods plot run in ProtScale at ExpaSy. (A) Kyte and Doolittle hydrophobicity plot of mouse Slco1a6 protein with a window size of 7. As a result, the effective length is 664—that is, after the 664th amino acid, another 7-amino-acid window is not available (the protein length is 670 amino acids). Peaks above the line corresponding to 0 represent the hydrophobic regions and peaks below this line represent hydrophilic regions of the protein. (B) Slco1a6 is a transmembrane protein. Thus, increasing the window size to 21 clearly makes the transmembrane regions prominent. This change makes the effective length 650. (C) Hopp and Woods hydrophilicity/antigenicity plot with a window size of 7. Peaks above the line corresponding to 0 represent the hydrophilic regions and peaks below this line represent hydrophobic regions of the protein.

When designing peptide antibodies, a Hopp and Woods hydropathy plot can be used to determine the regions of the polypeptide that are expected to have good antigenicity and thus trigger an antibody response in an animal treated with adjuvant-coupled peptide containing those sequence(s). Recently, Jääskeläinen et al. (2010)[14] investigated the prediction accuracy of 56 hydropathy scales by correlating predicted values with the accessible surface area in known 3D structures of proteins. They found that some epitopes are located among the most exposed regions, thereby reinforcing the utility of the hydropathy scales in predicting the antigenic regions of a protein.

Another metric of the overall hydrophobicity/hydrophilicity of a polypeptide is the **GRAVY** (grand average of hydropathy) score. The GRAVY value of a polypeptide is calculated by adding the hydropathy values of all the constituent amino acids and dividing the sum by the length of the sequence. *A positive GRAVY value indicates that the protein is hydrophobic and a negative value indicates that it is hydrophilic.*[12] Therefore, membrane proteins have higher GRAVY scores than globular proteins. ProtParam calculates the GRAVY score (Figure 8.2). The GRAVY score of mouse Slco1a6 is 0.267, indicating that it is a hydrophobic protein.

8.6 PREDICTION OF POST-TRANSLATIONAL MODIFICATION AND SORTING

Proteins can be post-translationally modified in many different ways, such as *N*-glycosylation, *O*-glycosylation and many other post-translational modifications. Proteins are also sorted (targeted) to various subcellular compartments either during translation (co-translational) or following translation (post-translational).

TABLE 8.2 Some Online Analysis Tools for Prediction of Post-Translational Protein Modifications, Protein Sorting, Localization Signals.

Online Tool	URL
CBS Prediction Servers (Center for Biological Sequence Analysis, Technical University of Denmark DTU)	http://www.cbs.dtu.dk/services/*
PSORT (Protein Sorting)	http://psort.hgc.jp/†
Gene Infinity	http://www.geneinfinity.org/sp/sp_proteinptmodifs.html‡

*Check CBS access policy to prediction servers at http://www.cbs.dtu.dk/cgi-bin/nph-access.
†PSORT program was coded by Kenta Nakai, Ph.D., Human Genome Center, Institute for Medical Science, University of Tokyo, Japan. Various scientists and their collaborators involved in developing different versions of the PSORT program are acknowledged on the PSORT home page.
‡Check the Terms of Service on the Gene Infinity home page.

For example, a large number of secretory proteins, membrane-bound proteins, and proteins in the endoplasmic reticulum are sorted co-translationally, whereas proteins targeted to the nucleus, mitochondria, and chloroplast are sorted post-translationally. Protein sorting requires specific signal sequences. In eukaryotic proteins, signal sequences are present at the N-terminal end of the protein. A comprehensive list of online analysis tools for the prediction of various post-translational protein modifications as well as protein sorting and localization signals can be found at the resources listed in Table 8.2.

8.7 SECONDARY-STRUCTURE PREDICTION

Efforts to predict protein secondary structures began long before the first protein structures were solved. Two of the earliest methods, the Chou–Fasman method and the GOR method, developed in the 1970s, have been widely used and are still being used.

8.7.1 The Chou–Fasman and GOR Methods

The Chou–Fasman and GOR (Garnier–Osguthorpe–Robson) methods were developed in the 1970s, and are among the oldest secondary-structure prediction methods. They are still widely used. The latest version of the GOR method is GOR V.[15] Both the Chou–Fasman and GOR methods are based on the analysis of the propensity of different amino acids to be in α-helix, β-strand, or β-turn. In these methods, the relative frequencies of amino acids in helix, strand, and turn are calculated based on known protein structures solved by X-ray crystallography. These relative frequency values are used to calculate the probability that an amino acid will appear in a helix, strand, or turn in a protein.

The application of the Chou–Fasman method is simple in principle. The sequence is scanned to identify regions of high helix or strand probability. For α-helix, a window size of six amino acids is used. If four contiguous residues out of six have P(α-helix) > 100, that segment is called as a helix. Once the helix is predicted, it is extended on both sides until at least four contiguous residues with P(α-helix) < 100 are found. That region is called as the end of the helix. For β-strand, a window size of five amino acids is used. The sequence is scanned to identify regions where at least three contiguous residues out of five have a value of P(β-strand) > 100. That region is called as a β-strand, and is extended on both sides until a set of three contiguous residues that have an average P(β-strand) < 100 is reached. That region is called as the end of the β-strand. If the residues in a region show the propensity of being in both α-helix and β-strand, the prediction is made based on the following principle: if Σ[P(α-helix)] > Σ[P(β-strand)], the region is called as a α-helix, otherwise a β-strand. Turns are also evaluated in four-residue windows, and are identified if P(β-turn) > 0.000075, where P(β-turn) = f(i)*f(i + 1)*f(i + 2)*f(i + 3). Table 8.3 shows the relative propensity values of amino acids as used by the Chou–Fasman method. Online Chou–Fasman and GOR prediction tools can be accessed from many sources (Table 8.4; see also CFSSP link in Table 8.5).

Like the Chou–Fasman method, the original GOR method also uses the propensity of amino acids to be in a helix, strand, turn, or coil. However, the GOR method uses a 17-residue window size and calculates the propensity of the residues in that window to be in each of the four states. The state with the highest score is predicted to be the state of the central residue (9th residue) of that window. Because the state of an amino acid is often influenced by the states of the neighboring amino acids, the GOR method takes into account the interactions of the neighboring residues.

With the availability of more sequences and more solved protein structures, some of the older methods have been revised and improved, such as GOR II, III, and IV.

8.7.2 Advances in Secondary-Structure Prediction

As the atomic detail of the structure of integral membrane proteins was determined in the mid-1980s, the homology-modeling method was developed as a

TABLE 8.3 Amino-Acid Relative Propensity Values Used by the Chou–Fasman Method

Amino Acid	P (α-helix)	P (β-strand)	P (β-turn)	f(i)	f(i + 1)	f(i + 2)	f(i + 3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic acid	151	037	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

TABLE 8.4 Some Online Chou–Fasman and GOR Prediction Tools

Chou–Fasman and GOR Prediction Tool	URL
University of Virginia	http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm = misc1* (select Chou–Fasman or GOR method)
ProtScale at ExPASy	http://web.expasy.org/protscale/[8] (select Chou–Fasman or GOR method)
Center for Informational Biology, Japan	http://cib.cf.ocha.ac.jp/bitool/MIX/† (select Chou–Fasman or GOR method)

*©1988, 2006, by William R. Pearson and the University of Virginia.
†The home page cites the papers based on which the method implemented in this server was developed. The Chou–Fasman and GOR papers are cited elsewhere in the text.

way of predicting secondary structures. In **homology modeling**, the secondary structure of the target protein is predicted based on the known structure of homologous proteins (template). Hence, homology modeling is based on sequence similarity/identity; obviously, the higher the sequence similarity/identity between the target and the template, the greater is the chance of accuracy of prediction. Nevertheless, homology modeling may not accurately predict the side chains and folds, making the overall predictions less accurate.

With advances in computation techniques, increase in the number of database entries, and increased knowledge of various protein folds, the concept of protein **sequence–structure threading** developed in the 1990s. In **protein threading** (**fold recognition**), target sequence is mapped to known template structures from the database. The sequence–structure compatibility is assessed by a scoring function. The method is based on the premises that, (1) there is a far lower number of unique folds among proteins than there are known proteins, and (2) information on the physicochemical properties of amino acids and knowledge of their occurrence in different structural environments provide important clues to their potential occurrence among different types of folds. Energy functions are an important consideration because energetics is very important in folding. *During computation of threading, the threading with minimum energy is assumed to represent the most likely fold structure.*

TABLE 8.5 Some Online Tools for the Analysis of Possible Secondary Structure of a Protein

Online Tool	Comments and URL
APSSP	http://imtech.res.in/raghava/apssp/*
CFSSP (Chou–Fasman[16] Secondary-Structure Prediction)	http://www.biogem.org/tool/chou-fasman/†
GOR IV	GOR IV[17]; GOR I, the original GOR[18] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html)‡[19]
HMMSTR	HMM-based[20,21] (http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php)
JPred 3	Combines the analysis from multiple prediction algorithms, such as DSC, JNET, PHD, and PREDATOR[22] (http://www.compbio.dundee.ac.uk/www-jpred/)
NPS@ (Network Protein Sequence Analysis)	This site contains links to a number of prediction tools including GOR and PHD. However, GOR and PHD are mentioned here separately as well. Pay attention to those that were developed in the late 1990s. Compare the output from these tools‡[19] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html)
PHD	Neural-network-based[23–25] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html)
PredictProtein	Meta-server that combines the analysis from multiple prediction algorithms such as Jpred, PHD, PROF, and PSIPRED. It is a good secondary-structure prediction program** (https://www.predictprotein.org/)
PROTEUS 2	Combination of HMM- and neural-network-based prediction[26] (http://wishart.biology.ualberta.ca/proteus2/)
PSIPRED	Combination of homology modeling and neural-network-based prediction. It is a good secondary-structure prediction program[27] (http://bioinf.cs.ucl.ac.uk/psipred/)
Quick2D	Provides an overview of secondary-structure features like α-helices, extended β-sheets, coiled coils, transmembrane helices, and disordered regions. Predictions by PSIPRED, JNET, Prof(Rost), Prof (Ouali), Coils, MEMSAT2, HMMTOP, DISOPRED2 and VSL2†† (http://toolkit.tuebingen.mpg.de/quick2_d)
SCRATCH Protein Predictor	The SCRATCH software suite includes predictors for a number of parameters, such as secondary structure, relative solvent accessibility, disordered regions, domains, individual residue contacts, tertiary structure, and more[28] (http://scratch.proteomics.ics.uci.edu/index.html)
SSPro 4.0	Bidirectional recurrent neural network (BRNN)-based[29,30] (http://download.igb.uci.edu/sspro4.html)
SYMPRED	SYMPRED can be run using any combination of the following programs: PHD, PROF, SSPro2.01, YASPIN, JNet, and PSIPRED. The consensus of the outputs is derived through dynamic programming to achieve a higher level of prediction accuracy[31] (http://www.ibi.vu.nl/programs/sympredwww/)
SOPMA	An improved self-optimized prediction method (SOPM)[32] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html)
YASPIN	Neural-network-based[33] (http://www.ibi.vu.nl/programs/yaspinwww/)

*An advanced version of the PSSP server.[34]

†© 2012, BioGem.Org.

‡Service supported by Ministère de la recherche (ACI IMPBio, ACC-SV13), CNRS (IMABIO, COMI, GENOME) and Région Rhône-Alpes (Programme EMERGENCE). The "Abstract" link can be clicked to obtain all the original references.

**The website provides a link to the entire PredictProtein team.

††© 2008, Dept. of Protein Evolution at the Max Planck Institute for Developmental Biology, Tübingen.

TABLE 8.6 Some Online Prediction Tools for Coiled Coils and Zippers

Online Tool	Comments and URL
ExPASy COILS	COILS compares the input sequence to a database of known parallel two-stranded coiled coils and derives a similarity score. By comparing this score to the scores in globular and coiled-coil proteins, COILS calculates the probability that the sequence will adopt a coiled-coil conformation[35] (http://embnet.vital-it.ch/software/COILS_form.html)
Paircoil2 at MIT	New version of the Paircoil program, which uses pairwise residue probabilities to detect coiled-coil motifs. Paircoil2 achieves 98% sensitivity and 97% specificity on known coiled coils[36] (http://groups.csail.mit.edu/cb/paircoil2/paircoil2.html)
2ZIP	Combines a standard coiled-coil-prediction algorithm with an approximate search for the characteristic leucine repeat. No further information from homologs is required for prediction[37] (http://2zip.molgen.mpg.de/)

Advances in protein-threading algorithms have allowed more accurate fold prediction. Secondary-structure prediction has further benefited from the introduction of methods like neural networks, hidden Markov models (HMMs), and the ability to train new models on an extensive set of sequence and structural data.

There are a number of online tools available for the analysis of possible secondary structure of a protein. ExPASy provides links to many of these tools. The links in Table 8.5 are cited because the analysis can be done in real time using most of these tools and the output is quickly obtained. There are many more online secondary-structure predictions tools that are not cited here.

These tools predict various secondary structures that different parts of the polypeptide can assume, such as the α-helix, 3_{10}-helix, π-helix, extended strand, β-turn, random coil, or ambiguous state. Analyzing a polypeptide sequence using different prediction tools may not produce the same results. For example, analyzing mouse Slco1a6 using four of these tools produces the following results: the prediction of α-helix varies between ~23 and 38%, that of extended strand varies between ~11 and 27%, and that of random coil varies between 42 and 51%. It is therefore advisable to analyze the sequence using multiple programs. Some of the standard notations in the output are as follows: α-helix (H/h), 3_{10}-helix (G/g), π-helix (I/i), extended strand (E/e), β-turn (T/t), random coil (C/c).

Online tools for the prediction of coiled coils and zippers are shown in Table 8.6. The direct link for ExPASy COILS is given in the table. It can also be accessed by first accessing ExPASy (http://www.expasy.org/), then accessing COILS by clicking "Resources A..Z".

8.7.3 Predicting the Accuracy of Secondary-Structure Prediction

A widely used metric to determine the overall accuracy of secondary-structure prediction is the **Q3 score**. A Q3 score is a measure of the quality of prediction of all three states (helix, strand, and coil), and it represents the percentage of residues that are correctly predicted (the states of the residues). The Q3 score can range from 0 to 1; 1 being the perfect prediction (100%). Currently, almost all secondary-structure-prediction algorithms achieve a Q3 score of 0.75 or higher. It should be remembered that Q3 is not an absolute measure of the prediction accuracy; there are other measures as well.

8.8 PREDICTION OF DOMAINS AND MOTIFS

A domain is part of the tertiary structure of protein. Each domain is a discrete globular unit that folds independently of the rest of the protein. Domains have specific functional roles. Domains can be composed of as few as 20–25 amino acids, but frequently much more than 25. The average number of domains in a protein is usually two to three, but can be more. By shuffling a finite number of domains, nature has created proteins with diverse functions during evolution. Thus, proteins with similar functions are expected to contain conserved regions that are associated with the function; the rest of the protein sequence may be different. Examples of some familiar domains are the **SH3 (Src-homology 3) domain**, which is around 50 amino acids and involved in protein–protein interactions; the **chromo (chromatin organization modifier) domain**, which is 30–70 amino acids and involved in the assembly of protein complexes on chromatin; and the **death domain**, which is around 80–100 amino acids and involved in apoptotic signal transduction.

As opposed to domains, a specific functional element of the protein that usually does not fold independently of the rest of the protein is called a **motif**, such as a sequence motif or a structural motif (e.g. a stretch of secondary structure). Domains contain within themselves specific motifs that are critical to domain function. Some examples of structural motifs in proteins are various loop and turns, such as omega loops, beta turns, helix–loop–helix, and helix–turn–helix. Sometimes the terms domain and motif are used interchangeably in the context of proteins, such as "coiled-coil" domain/motif, "leucine-zipper" domain/motif.

The domain analysis of Slco1a6 using **InterProScan** (http://www.ebi.ac.uk/Tools/pfa/iprscan/)[38] at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is shown in Figure 8.4 and Figure 8.5. At the default setting, all

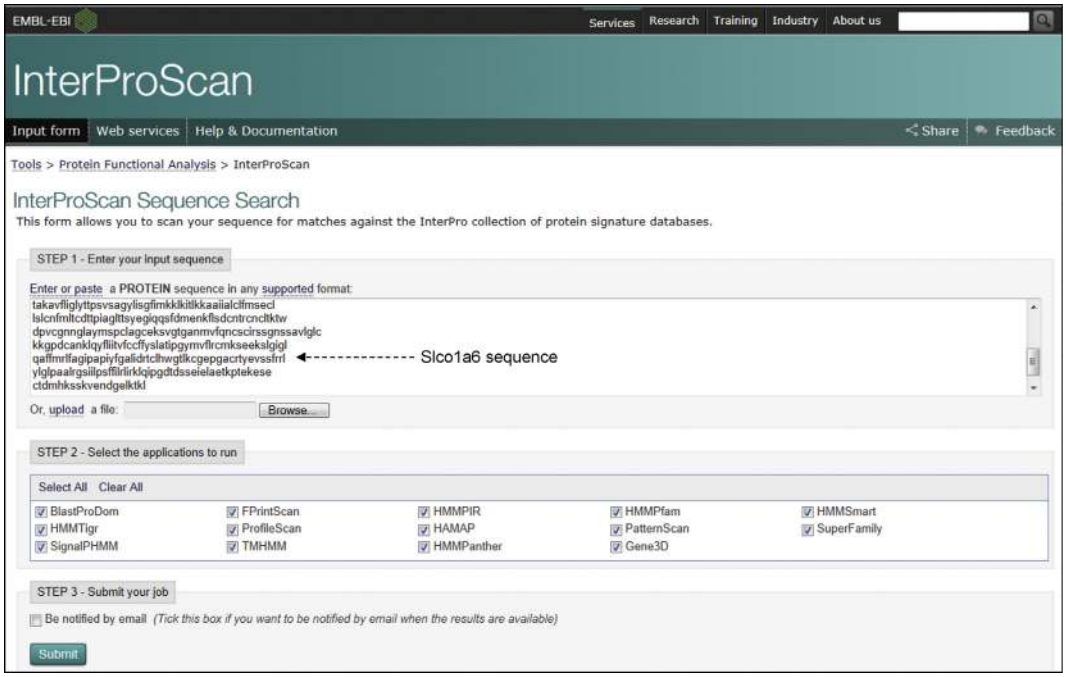


FIGURE 8.4 **InterProScan home page at EMBL-EBI from where the search and analysis can be launched.** The page shows that at the default setting all applications are checked; each one scans the input sequence against a specific database.

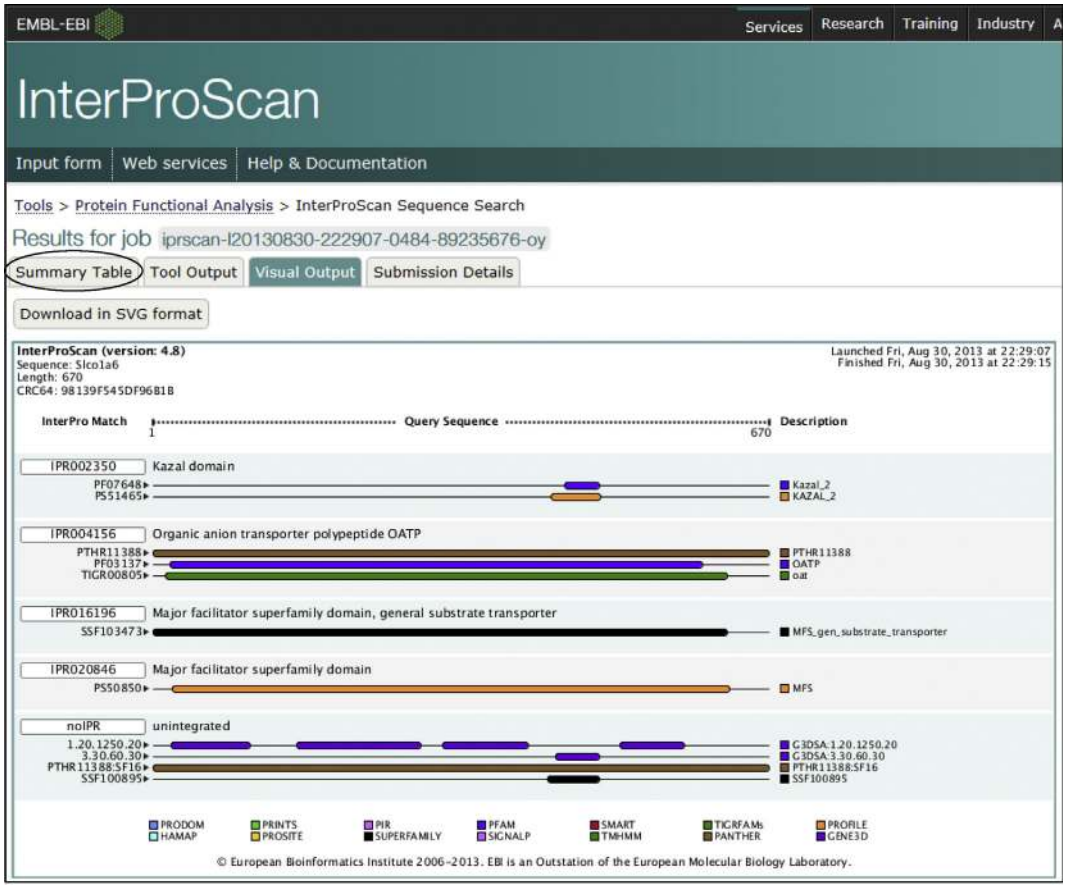


FIGURE 8.5 **The graphical display of InterProScan analysis.** Two major domains identified are Kazal and MFS. More information on these domains can be obtained from various links under the "Summary Table" tab. The predictions from different databases may not be identical (see text). Nevertheless, these tools are very important in identifying specific signatures in protein sequence.

applications are checked; each one scans the input sequence against a specific database (see "Help & Documentation" for details; Figure 8.4). The graphical display of the analysis is shown in Figure 8.5. Two major domains identified are **Kazal** and **MFS** (see Box 8.1). Clicking "Summary Table" shows various links for more information on the domains and their distribution. The predictions from different databases may not be identical; for example, PROFILE predicts the Kazal domain spanning from residue 433 to 488, whereas Pfam predicts the Kazal domain spanning from residue 447 to 486. PROFILE predicts the MFS domain spanning from residue 21 to 627, whereas SuperFamily predicts the MFS domain spanning from residue 1 to 625. Despite small differences in prediction, these tools are very important in identifying specific sequence signatures in protein sequence.

The domain analysis of Slco1a6 using the NCBI CDD is shown in Figure 8.6, Figure 8.7, and Figure 8.8. **CDD** (**Conserved Domain Database**) of NCBI provides annotation of protein sequences with the location of conserved-domain footprints and functional sites inferred from these footprints. CDD is built on NCBI-curated domains and data imported from Pfam, SMART, COG, PRK, and TIGRFAM.[39] CDD can be accessed directly at http://www.ncbi.nlm.nih.gov/cdd, or from the NCBI home page. Figure 8.6 shows the CDD home page. Clicking "**CD-Search**" (circled) takes the user to the search launch page, shown in Figure 8.7. Submitting the Slco1a6 sequence in FASTA format under default settings returns the analysis shown in Figure 8.8. The result can be displayed in a "concise format" that displays the best hits, or "full format" that displays all hits. Figure 8.8 shows the concise format. Like InterProScan, CDD analysis also shows that Slco1a6 contains **Kazal** (**Kazal_SLC21**) and **MFS** domains. However, the predicted MFS domain is shorter (21–270) than that predicted by InterProScan (PROFILE).

It should be remembered that the domain/motif prediction is predicated on sequence alignment. Just like with any other

BOX 8.1

KAZAL AND MFS DOMAINS

The activity of proteases in cells is under tight control to prevent any unintended tissue damage. Cells produce various types of proteases along with peptide protease inhibitors to regulate the protease activity. Serine protease* activities are regulated by serine protease inhibitors, which are distributed in a wide range of organisms from all kingdoms of life. Pancreatic acinar cells produce two types of serine protease inhibitors; one is the **Kunitz** inhibitors (e.g. PTI, or pancreatic trypsin inhibitor) that remain in the pancreatic cells, and the other is **Kazal** inhibitors (e.g. PSTI, or pancreatic secretory trypsin inhibitor) that are secreted with the zymogens in the pancreatic juice. Some other examples of Kazal-type inhibitors are avian ovomucoid, acrosin inhibitor, and elastase inhibitor. Kazal-type inhibitors are the most studied protease inhibitors, and they contain one or more Kazal-type domains. The typical **Kazal domain** is a small α/β fold, consisting of one α-helix surrounded by an adjacent three-stranded β-sheet and loops of peptide segments†.[40]

The major facilitator superfamily (**MFS**) is the largest known superfamily of **secondary transporters** found in living organisms. Secondary transporters do not use ATP directly for transport, but use an already-existing electrochemical gradient‡. More than 70 families are known; members of each family transport a different set of related compounds, such as simple monosaccharides, oligosaccharides, amino acids, peptides, vitamins, enzyme cofactors, drugs, nucleobases, nucleosides, nucleotides, and organic and inorganic anions and cations. MFS proteins are single-polypeptide secondary transporters, and the **MFS domain** consists of either 12 or 14 transmembrane helices connected by hydrophilic loops**.[42,43] Secondary active transport can move materials against the concentration gradient, and can also transport just one substrate (uniporter), or two substrates in the same direction (symporter), or in the opposite direction (antiporter).

*Serine proteases contain a reactive serine in their active site and this serine is crucial for their function. Trypsin, chymotrypsin, and elastase are three important eukaryotic serine proteases; subtilisin is an important bacterial serine protease. Trypsin is involved in the activation of pancreatic zymogens. Serine proteases also constitute over one-third of all proteases[41]

†http://www.ebi.ac.uk/interpro/entry/IPR002350; http://prosite.expasy.org/PDOC00254

‡An electrochemical gradient is a gradient of electrochemical potential, which is generated by the differential distribution of electrical potential and chemical concentration across the membrane. Differential distribution of ions across the membrane, for example sodium ions, generates an electrochemical gradient. It consists of two components: the electrical potential difference caused by the uneven distribution of the charge, and the concentration difference caused by the uneven distribution of sodium itself. The electrochemical gradient generates potential energy because the ions involved are ready to move across the membrane. However, the ions cannot pass through the membrane lipid bilayer without the help of an active transport mechanism. The MFS transporters convert this potential energy into kinetic energy when they transport the respective substrates

**http://www.ebi.ac.uk/interpro/entry/IPR016196;jsessionid; http://pfam.sanger.ac.uk/clan/CL0015

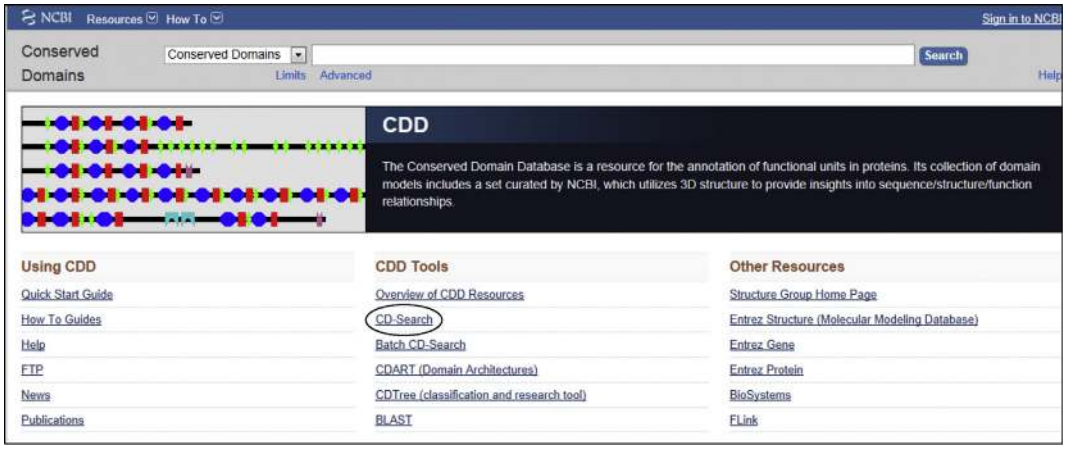


FIGURE 8.6 **The Conserved Domain Database (CDD) home page.** Clicking CD-search (circled) takes the user to the search and analysis launch page (Figure 8.7).

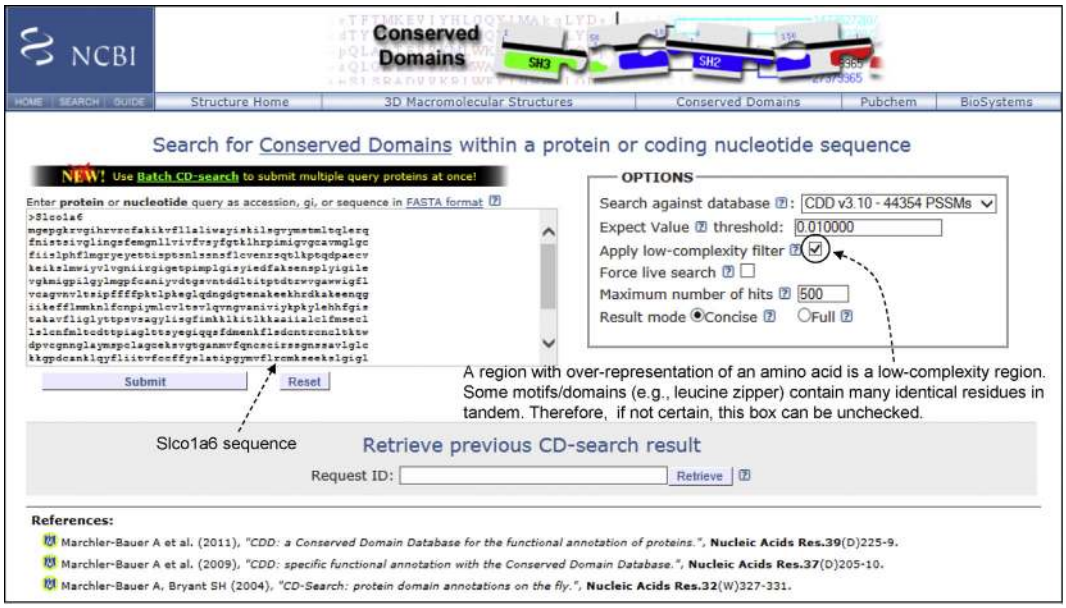


FIGURE 8.7 **The CDD search and analysis launch page.** Submitting the Slco1a6 sequence in FASTA format under default settings returns the analysis shown in Figure 8.8. In the default settings, the "low-complexity" filter is on. This can be turned off.

predictions, there is an element of uncertainty—that is, a domain may be falsely predicted or a true domain may be missed, particularly conformational domains.

Another good online tool for domain analysis is **PROSITE** (http://prosite.expasy.org/prosite.html).[44,45] PROSITE scan (**ScanProsite**) of Slco1a6 produces the following results: **Kazal** domain spanning residues 433–488 and **MFS** domain spanning residues 21–627 (not shown).

8.8.1 Transmembrane-Helix Prediction

Because domain analysis shows the existence of an MFS domain in Slco1a6, a specific search for the transmembrane (TM) helices can be done. There are a number of good online TM-helix-prediction tools, as shown in Table 8.7.

RHYTHM produces a nice graphical output of TM helices, showing the amino-acid sequence in each helix. Figure 8.9 shows the gist of TM-helix prediction by all four prediction tools. TMHMM (version 2.0) predicted 11 TM helices, whereas RHYTHM, OCTOPUS, and Phobius each predicted 12 TM helices (Figure 8.9). The graphical outputs of RHYTHM and OCTOPUS are shown in Figure 8.10. In the span of residue 110 to residue 240 (approximately), TMHMM predicted one TM helix, whereas RHYTHM, OCTOPUS, and Phobius predicted two. As a result, the assignment of inside and

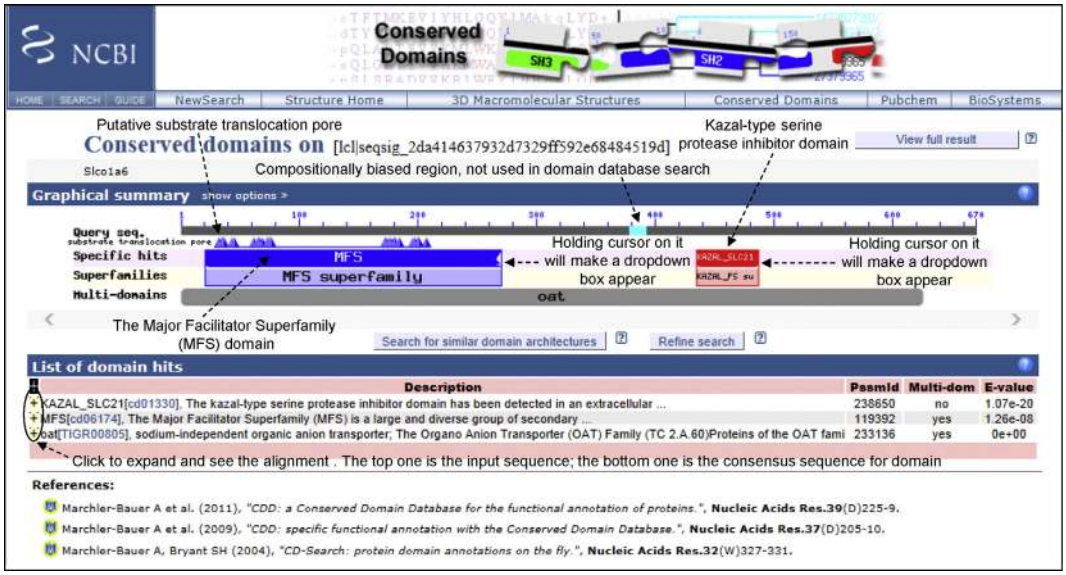


FIGURE 8.8 **Result of CDD domain analysis.** The result is displayed in the "concise format." Analysis shows that Slco1a6 contains Kazal (Kazal_SLC21) and MFS domains. The predicted MFS domain is shorter (21–270) than that predicted by InterProScan (see text). Holding the cursor over MFS or Kazal_SLC21 produces a drop-down box that contains detailed description of the specific hit.

TABLE 8.7 Some Online Tools for Transmembrane-Helix Prediction

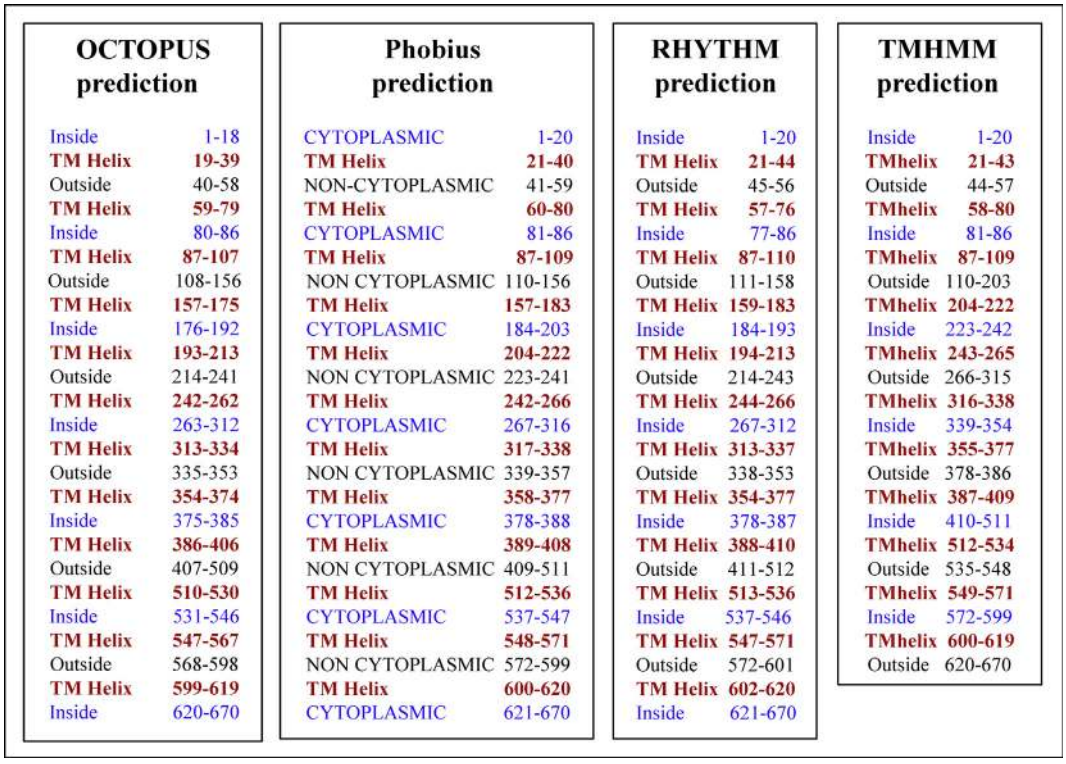
Online Tool	Comments and URL
TMHMM	Hidden-Markov-model-based[46] (http://www.cbs.dtu.dk/services/TMHMM/)
RHYTHM	Utilizes the structural information from ever-growing data sets and evolutionary information from conserved-sequence patterns in a representative data set of membrane proteins[47] (http://proteinformatics.charite.de/rhythm/)
Phobius	Hidden-Markov-model-based[48] (http://phobius.sbc.su.se/)
OCTOPUS	Artificial-neural-network-based[49] (http://octopus.cbr.su.se/)

outside segments is reversed between the TMHMM prediction and those of the other three programs from residue 214/223 onwards. However, TMHMM is a widely used, good TM-helix-prediction program, and TMHMM prediction is focused on TM helices only and not necessarily on the cytoplasmic and the extracellular segments. Overall, the TM helices were predicted correctly by all four programs. Nevertheless, this example further underscores the fact that it is a good idea to run an analysis simultaneously using multiple programs.

8.9 VIEWING THE 3D STRUCTURE OF PROTEINS (AND OTHER BIOLOGICAL MACROMOLECULES)

The 3D structures of many proteins and other biological macromolecules have been determined using various techniques of modern structural biology. These structures are deposited in the **PDB** (**Protein Data Bank**) database and are given a PDB ID. The PDB ID is a four-character unique identifier, consisting of numbers and letters, assigned to a protein or other biological macromolecule submitted to the PDB. The PDB is an archive of the structure of proteins and other biological macromolecules; the structures have been determined using techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy. After structural information is submitted to the PDB, the submission is annotated and publicly released by the **wwPDB** (http://www.wwpdb.org/). As of July 30, 2013, there were 92,689 structures in the PDB. PDB IDs are usually written in uppercase. Some examples of PDB IDs are 2HHD (human hemoglobin, deoxy form), 9INS (pig insulin), and 2VRY (mouse neuroglobin). The PDB can be searched by simply typing the description, or partial sequence, or the PDB ID (if known).

FirstGlance in Jmol (http://bioinformatics.org/firstglance/fgij/index.htm) is a user interface to the free molecular visualization program named **Jmol** (http://jmol.sourceforge.net/). **Jmol** is a free and



OCTOPUS prediction		Phobius prediction		RHYTHM prediction		TMHMM prediction	
Inside	1-18	CYTOPLASMIC	1-20	Inside	1-20	Inside	1-20
TM Helix	19-39	TM Helix	21-40	TM Helix	21-44	TMhelix	21-43
Outside	40-58	NON-CYTOPLASMIC	41-59	Outside	45-56	Outside	44-57
TM Helix	59-79	TM Helix	60-80	TM Helix	57-76	TMhelix	58-80
Inside	80-86	CYTOPLASMIC	81-86	Inside	77-86	Inside	81-86
TM Helix	87-107	TM Helix	87-109	TM Helix	87-110	TMhelix	87-109
Outside	108-156	NON CYTOPLASMIC	110-156	Outside	111-158	Outside	110-203
TM Helix	157-175	TM Helix	157-183	TM Helix	159-183	TMhelix	204-222
Inside	176-192	CYTOPLASMIC	184-203	Inside	184-193	Inside	223-242
TM Helix	193-213	TM Helix	204-222	TM Helix	194-213	TMhelix	243-265
Outside	214-241	NON CYTOPLASMIC	223-241	Outside	214-243	Outside	266-315
TM Helix	242-262	TM Helix	242-266	TM Helix	244-266	TMhelix	316-338
Inside	263-312	CYTOPLASMIC	267-316	Inside	267-312	Inside	339-354
TM Helix	313-334	TM Helix	317-338	TM Helix	313-337	TMhelix	355-377
Outside	335-353	NON CYTOPLASMIC	339-357	Outside	338-353	Outside	378-386
TM Helix	354-374	TM Helix	358-377	TM Helix	354-377	TMhelix	387-409
Inside	375-385	CYTOPLASMIC	378-388	Inside	378-387	Inside	410-511
TM Helix	386-406	TM Helix	389-408	TM Helix	388-410	TMhelix	512-534
Outside	407-509	NON CYTOPLASMIC	409-511	Outside	411-512	Outside	535-548
TM Helix	510-530	TM Helix	512-536	TM Helix	513-536	TMhelix	549-571
Inside	531-546	CYTOPLASMIC	537-547	Inside	537-546	Inside	572-599
TM Helix	547-567	TM Helix	548-571	TM Helix	547-571	TMhelix	600-619
Outside	568-598	NON CYTOPLASMIC	572-599	Outside	572-601	Outside	620-670
TM Helix	599-619	TM Helix	600-620	TM Helix	602-620		
Inside	620-670	CYTOPLASMIC	621-670	Inside	621-670		

FIGURE 8.9 **Transmembrane-helix prediction at a glance by RHYTHM, OCTOPUS, Phobius, and TMHMM.** TMHMM (version 2.0) predicted 11 TM helices, whereas RHYTHM, OCTOPUS and Phobius predicted 12 (see text for details). This example underscores the fact that it is a good idea to run an analysis simultaneously using multiple programs.

open-source software program written in Java for viewing chemical structure in 3D. It runs on various operating systems, such as Windows, MacOS, and Unix, and is also downloadable. The Jmol website has a user-friendly tutorial. **FirstGlance in Jmol** provides an easy way to look at the 3D structures of proteins, DNA, RNA, and their complexes, including with animation. In order to use **FirstGlance in Jmol**, one has to know the PDB ID of the macromolecule or have the data as PDB file format. On the **FirstGlance in Jmol** website, help is displayed automatically with links to further information about structural biology terms and concepts. The website also provides links to a "Gallery of Interactive Molecules" and a "Snapsot Gallery." Therefore, between the **Jmol** tutorial and **FirstGlance in Jmol** helpful links, the beginner will find it quite easy to understand the output.

8.10 ALLERGENIC PROTEIN DATABASES AND PROTEIN-ALLERGENICITY PREDICTION

Substances that cause allergic reactions are called **allergens**. Almost all allergens are proteins and they induce allergic response in susceptible individuals. Because allergic reactions result from complex interactions between the allergenic proteins and the immune system (see footnote on epitopes), and because allergic reactions are seen only in susceptible individuals, the allergenic potential of proteins is difficult to predict.

8.10.1 WHO/IUIS Allergen Nomenclature and Database of Allergenic Proteins

The World Health Organization/International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature Subcommittee is responsible for developing a systematic Linnaean nomenclature of allergens and maintaining a database of confirmed allergenic proteins.[50,51] A Linnaean nomenclature of an organism consists of a genus and a species term. The allergen name is normally made up of the first three letters of the genus name, first one letter from the species name, and a number that represents the order of its identification. In some instances, this rule has to be modified, such as Asp fl 13 (from *Aspergillus flavus*) and Asp f 13 (from *Aspergillus fumigatus*). Note that for *Aspergillus flavus* Asp fl 13, two letters from the species name, instead of one letter, have been used.

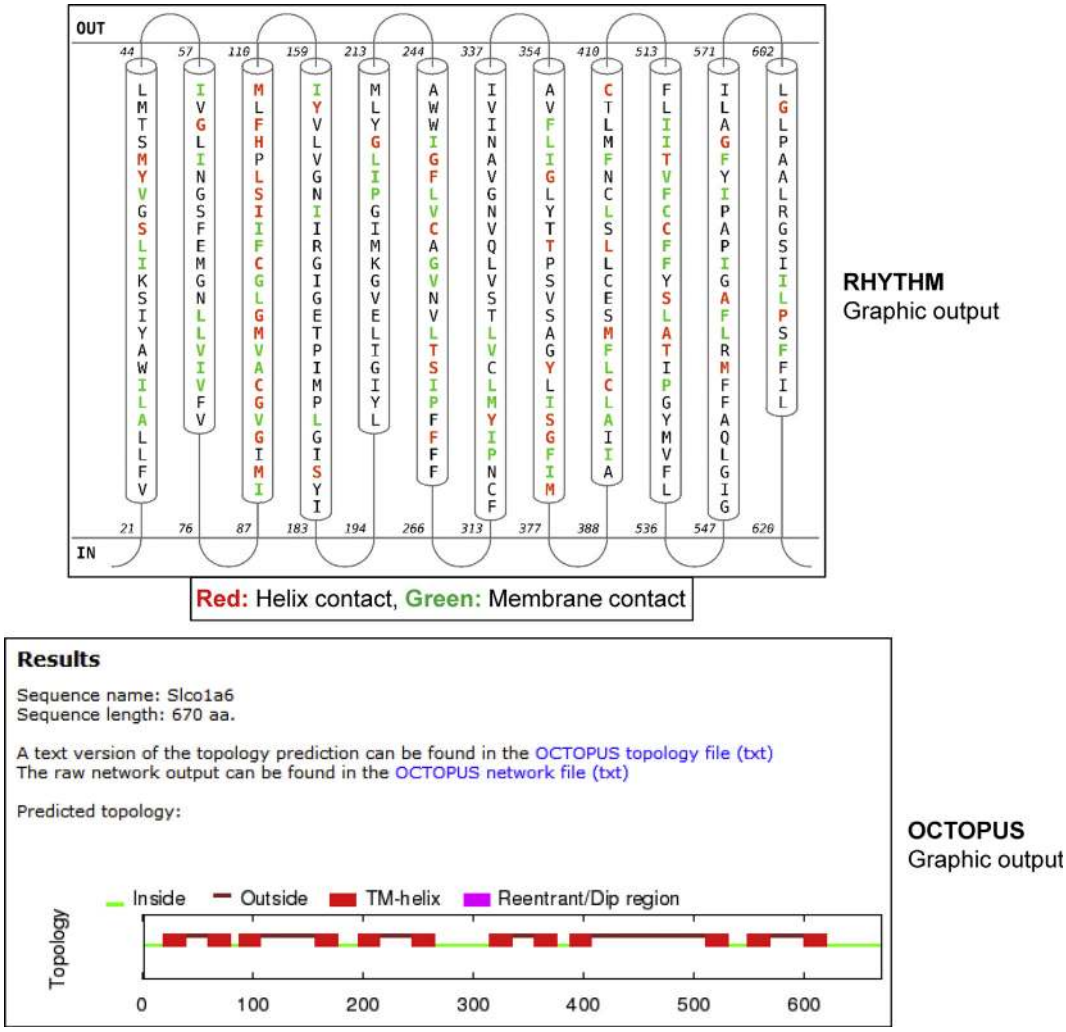


FIGURE 8.10 **The graphical outputs of RHYTHM and OCTOPUS.** The RHYTHM graphical output shows the relative length of the predicted helices and the amino-acid sequence of each predicted helix, as well as the residues that are in contact with the membrane and the residues involved in helix contact.

The WHO/IUIS allergen database contains information of approved and officially recognized allergens—that is, for a protein to be designated an allergen by WHO/IUIS, the allergenicity of the protein should be clinically documented. The database can be quickly searched for an allergen or an allergen source on the home page (http://www.allergen.org/index.php). Alternatively, an advanced search can be performed on the search page by clicking the "**Search**" tab or using the direct link http://www.allergen.org/search.php. By clicking the "**Tree View**" tab or using the direct link http://www.allergen.org/treeview.php, a list of allergens in fungi, plants, and different animal phyla can be directly obtained. An allergen record shows much important information about the allergen, such as the source, the evidence of allergenicity, allergenicity reference in PubMed, information on whether the allergen is a food allergen or not, any isoallergens and variants, and finally the sequence in both GenBank and UniProt.

8.10.2 Other Databases of Allergenic Proteins

In addition to the WHO/IUIS database, there are a number of other databases of allergenic proteins. Three of these databases are described in Chapter 5 (the **Structural Database of Allergenic Proteins (SDAP)**, **Allergenonline**, and **Allermatch**). Both the SDAP and Allergenonline databases are periodically updated; they both list more than 1500 allergenic proteins from food and non-food sources. Many allergens listed in these databases do not have IUIS designations yet. For a more comprehensive list of currently available allergen databases and allergen semantics, see Gendel[52] and other publications by the same author referenced in the paper.

8.10.3 Linear Epitopes, Conformational Epitopes, and Allergenicity

Although a protein acts as an allergen, the immune system actually recognizes smaller sections of the protein to trigger an allergic response. These small segments of the allergenic protein are called allergenic determinants, or **epitopes**[d]. The cognate antibody (IgE) binds to these allergenic epitopes to trigger the allergic response. Epitopes can be linear or conformational. In a **linear epitope**, the amino-acid sequence is continuous, whereas in a **conformational epitope**, the 3D conformation of the protein brings two separate sequence segments together to create the epitope. Conformational epitopes are usually destroyed when the protein is denatured, but linear epitopes are not affected by denaturation. Because many food allergens are stable in heat processing and digestion, it has been proposed that linear epitopes are more important than conformational epitopes for food allergens. However, the allergenicity of some foods, such as cow's milk and egg, is partly due to the IgE-binding conformational epitopes of their constituent proteins, such as α- and β-casein in cow's milk and ovomucoid in egg. Individuals whose immune system reacts to these conformational epitopes tend to grow out of the allergy as they get older, but reaction to the linear epitopes results in persistent allergy.[53–55] Conformational epitopes are also important for environmental allergens that are primarily inhaled.[56]

8.10.4 Allergenicity-Prediction Paradigm

Bioinformatics tools have been developed to identify the allergenic potential of an unknown protein by comparing its sequence to the sequences of known allergenic proteins in the database. A paradigm for assessing the allergenic potential of a protein in food was developed by the Food and Agricultural Organization/World Health Organization (FAO/WHO) as part of a multi-step safety-assessment process for foods produced through agricultural biotechnology.[57] The FAO/WHO paradigm uses two criteria: (1) an exact match of 6 contiguous amino acids, and (2) an overall sequence identity of more than 35% in a sliding window of 80 amino acids. Any protein that satisfies one or both of these criteria should trigger additional investigation to confirm whether the protein may truly have allergenic potential.

At the time the FAO/WHO paradigm was developed, it was already known that the smallest IgE-binding epitopes in an allergen could be only six-amino-acids long, as had been reported for Ara h 1 and Ara h 2.[58,59] The findings in these publications were based on epitope mapping with synthetic peptides that reacted with serum IgE from individuals with documented peanut hypersensitivity. Also, a publication by Burkhard Rost[60] had described the basis for a 35% identity cutoff and 80-amino-acid window threshold in pairwise sequence alignment. The author reported that protein pairs with similar structure (and function) are likely to have $>35\%$ sequence identity. The author analyzed more than a million sequence alignments between protein pairs of known structure. The goal was to distinguish between true and false positives for low levels of similarity. The author noted that sequence alignments could unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity was $>40\%$ for long alignments. The signal, however, became blurred when the sequence identity was between 20 and 35%; this $20-35\%$ range was termed the **twilight zone** of sequence identity. The pairwise sequence identity by itself is not meaningful without the context of a length-dependent threshold. In other words, a significant sequence identity can only be defined in the context of an optimum window of sequence length, which was determined to be around 80 amino acids. Such a requirement for a length threshold (around 80 amino acids) to determine a significant sequence identity had been described earlier by Sander and Schneider[61] and was also discussed by Rost.

8.10.5 Allergenicity-Prediction Servers

The bioinformatic tools to analyze the sequence of a protein according to FAO/WHO rules are available from multiple sources, such as SDAP, and Allermatch.

[d]An epitope, also called an antigenic determinant, is a region of the antigen (protein) that binds a secreted antibody, such as immunoglobulin G (IgG), or a membrane receptor on a lymphocyte, such as the T-cell receptor (TCR). Normally, such binding results in a humoral (antibody-mediated) immune response or a cellular (T-cell-mediated) immune response. Allergy is a special type of immune response that occurs in some individuals whose immune system overreacts to certain environmental substances that do not bother most other people. During an allergic response, IgE binds to the IgE receptor on mast cells (in tissues) and basophils (in circulation). When two or more IgEs bound to receptors on the mast cells or basophils are cross-linked by the allergen through the allergenic epitope, these cells are activated. Both mast cells and basophils contain special cytoplasmic granules that store many mediators of inflammation. The extracellular release of these mediators following activation of these cells is known as **degranulation**. A well-known mediator of inflammation released by mast cells is histamine. The released mediators of inflammation trigger allergy symptoms.

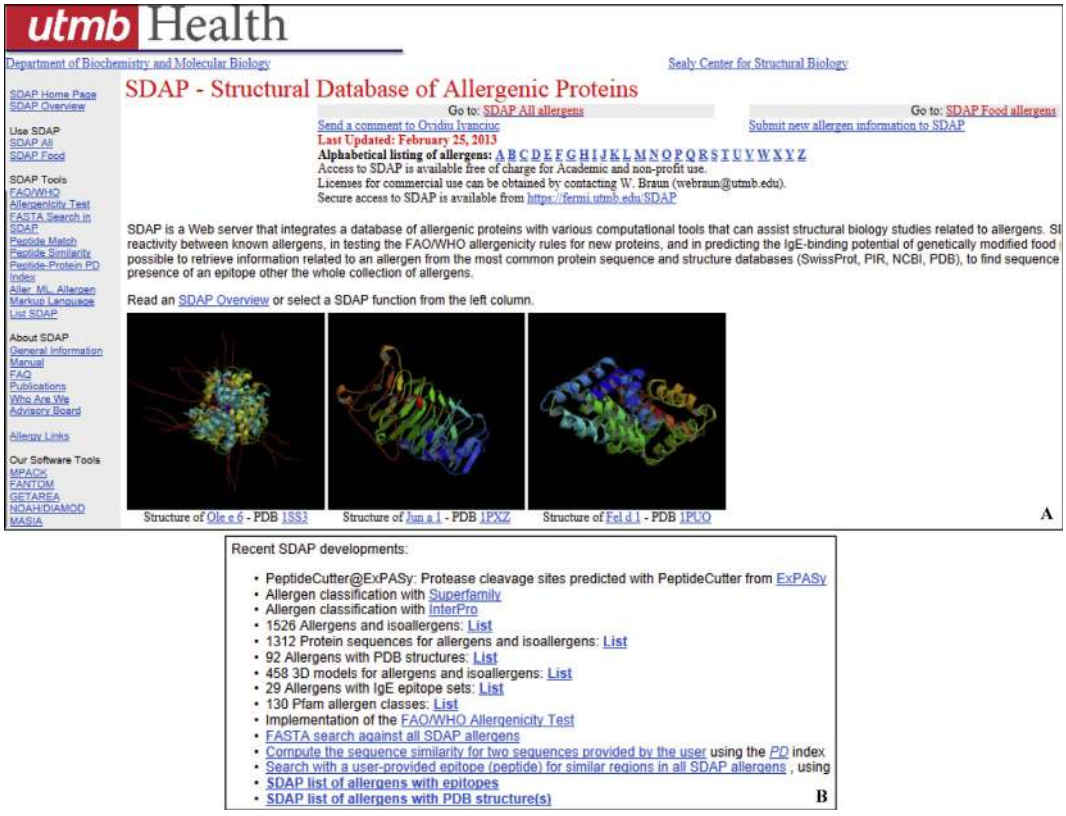


FIGURE 8.11 **The SDAP database home page.** (A) Partial (upper) screenshot of the SDAP database home page. Note the panel with links on the left-hand side, including links to SDAP tools. (B) Further down the home page is the "Recent SDAP developments" section (as of August 2013).

Allergenonline allows searching for an eight- (instead of six-) contiguous-amino-acid exact match. This change is based on the argument that searching for an exact match of six contiguous amino acids has the potential of generating many false positives.

In this section, we will focus on the information available from the SDAP database and analysis tools available on the **SDAP**[62,63] (https://fermi.utmb.edu/SDAP/) and **AlgPred**[64] (http://www.imtech.res.in/raghava/algpred/) servers. Figure 8.11A shows a partial (upper) screenshot of the SDAP database, whereas Figure 8.11B shows recent SDAP developments, as of August 2013.

On the panel on the left there are various links. One such link is "FAO/WHO Allergenicity Test." Clicking this link takes the user to the screen shown in Figure 8.12. The search for allergenicity of a protein can be launched from this page. Hitting the "Search" button returns a list of allergenic protein sequences that share one or more segments of six-contiguous-amino-acid identity with the input sequence. For demonstration, the sequence of mouse Slco1a6 has been pasted in the box (Figure 8.12) and analyzed using FAO/WHO rules. In this example, a total of six different segments of Slco1a6 (each segment is six-contiguous-amino-acids long) were found to match with segments of six different allergens from the database (Figure 8.13A and B). Figure 8.13A is a partial screenshot as displayed in the output. Figure 8.13B lists the other five hits between Slco1a6 and five different allergenic proteins. For these five hits, the screenshots of alignment are not shown, to save space. No sequence identity 35% or greater was found in a sliding window of 80 amino acids. *In practice, it is more common to have one or more six-contiguous-amino-acid sequence matches than to have >35% sequence identity in a sliding window of 80 amino acids.*

In the situation when there are six-contiguous-amino-acid segment matches between the input protein sequence and various allergenic proteins in the database, additional sequence comparison can be performed. For example, the distribution of these six-contiguous-amino-acid sequence segments can be verified using BLASTP against a curated protein database, such as UniProtKB/Swiss-Prot. The goal is to find out if these six-amino-acid sequence segments widely occur in various proteins that are not known to be allergenic. Additionally, the input sequence can be further analyzed using other prediction tools, such as **AlgPred**. Figure 8.14A shows that AlgPred offers several different approaches for predicting the allergenic potential of a protein (the input sequence). Five different

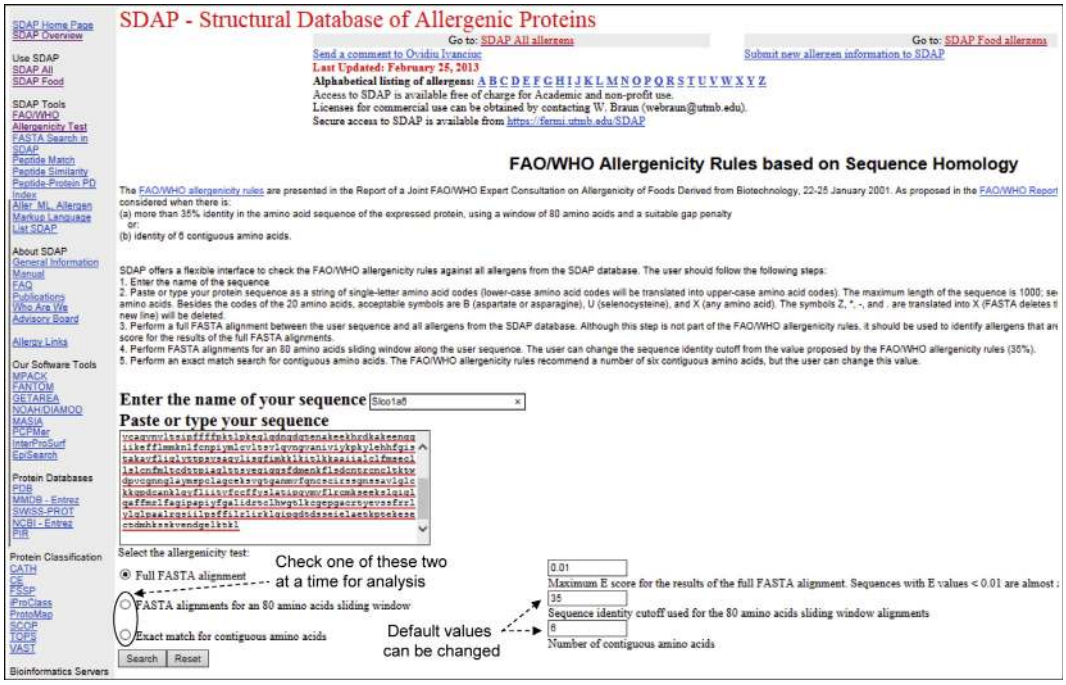
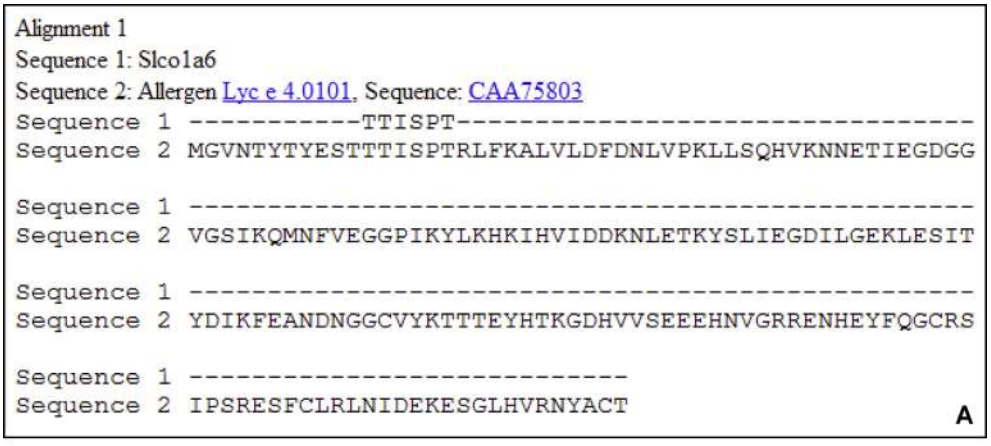


FIGURE 8.12 **FAO/WHO Rule-Based Allergenicity Prediction at the SDAP database.** The search for allergenicity of a protein according to FAO/WHO rules can be launched from this page. The default settings are 6 for contiguous amino acids, and 35 for % cutoff in a sliding window of 80 amino acids. These values can be changed by the user if needed. Selecting any one of these two options and hitting the "Search" button returns the results of the analysis. The sequence of mouse Slco1a6 has been pasted in the box for analysis according to FAO/WHO rules.



Other 5 hits

Allergen name	Sequence ID	6 amino acid exact match
Hel a 2,	O81982:	**GIIKEF**
Bos d 6,	P02769:	**AGCEKS**
Ves v 6.0101,	G8IIT0:	**NSSAVL**
Ani s 7.0101,	ABL77410:	**RCMKSE**
Hev b 10.0103,	CAC13961:	**AALRGS**

B

FIGURE 8.13 **Results of the FAO/WHO Rule-Based Allergenicity Prediction of Slco1a6.** A total of six different segments of Slco1a6, each six-contiguous-amino-acids long, were found to match with six different allergens from the database. (A) A partial screenshot of the six-contiguous-amino-acid hit, as displayed in the output. (B) The other five hits between Slco1a6 and five different allergenic proteins. No sequence identity 35% or greater was found in a sliding window of 80 amino acids.

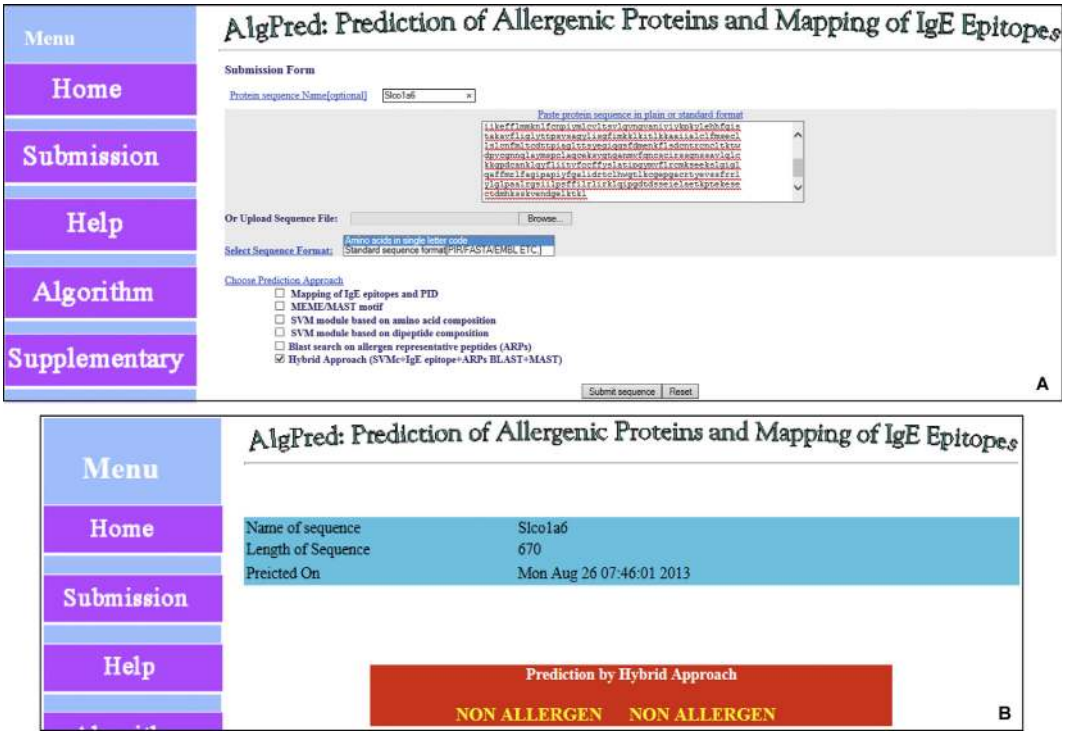


FIGURE 8.14 **Analysis of the input sequence using AlgPred.** (A) AlgPred offers several different approaches for predicting the allergenic potential of a protein (the input sequence). The hybrid approach that combines all five other approaches was chosen for the prediction (box checked). (B) The hybrid approach predicts Slco1a6 as a non-allergen. The same approach can be used to predict the potential allergenicity of a non-food protein.

approaches can be chosen for the prediction (listed on the home page), or the combination of all five in the "Hybrid Approach". Figure 8.14B shows that the hybrid approach predicts Slco1a6 as a non-allergen. The same approach can be used to predict the potential allergenicity of a non-food protein. It should be remembered that the sequence-based approach of allergenicity prediction is one of many tools utilized to assess whether a protein has the potential to be allergenic.

In addition to predicting the allergenic potential of a protein, there are a number of online T-cell and B-cell epitope-prediction tools that can be used to predict T-cell and B-cell epitopes, both continuous and discontinuous, in an input protein sequence. Such prediction methods take into account many aspects of protein structure, such as amino-acid properties (e.g. hydrophilicity and antigenicity, solvent accessibility, secondary structure, flexibility), amino-acid sequence, 3D structure wherever available, and information about the known epitopes from databases. The machine-learning prediction methods include the hidden Markov model (HMM), artificial neural network (ANN), and support vector machine (SVM). The SVM was found to be a better predictor compared to the other machine-learning prediction methods.[65] Some easily accessible online T-cell and B-cell epitope-prediction tools are available from the following sources:

http://www.imtech.res.in/raghava/
http://www.cbs.dtu.dk/services/
http://tools.immuneepitope.org/main/.

8.11 INTRINSICALLY DISORDERED PROTEIN ANALYSIS

Intrinsically disordered proteins (**IDPs**), also known as **intrinsically unstructured proteins** (**IUPs**), are characterized by the lack of a stable tertiary structure under physiological conditions. The lack of structural order in a protein goes against the traditional wisdom that protein function depends on a stable tertiary structure (the structure–function paradigm). It has long been realized that proteins possess configurational adaptability (e.g. induced fit). However, the presence of disordered segments in a functional protein became apparent when the crystal structures of various proteins became available. Techniques, such as NMR, X-ray crystallography, and circular dichroism helped uncover the disordered/unstructured state of certain proteins (e.g. missing

electron density of certain segments; hence, missing segments in X-ray crystallography). For these proteins, the intrinsically disordered state is necessary for function; some of these proteins fold only in complex with the substrate. It has been estimated that at least 50% of eukaryotic proteins possess at least one long (>40-amino-acid) loop, while this fraction is lot lower in prokaryotes and Archaea. *Protein disorder is found within loops*. Coiled coils may also assume disorder as they only assume globular structure when the coiled-coil partners interact with one another. *IDPs play an important role in signaling, recognition, and regulation*; recognition and regulation may involve processes like substrate recognition, catalysis, transport, DNA and RNA binding, and gene regulation. The presence of flexible structure and flexible structural segments helps accommodate a greater spectrum of binding targets, and also allows the IDP–target interaction to be short-lived, which is crucial for proper regulation. Because IDPs play an important role in signaling and regulation, they are much more abundant in eukaryotes than prokaryotes.[66–68]

8.11.1 IDP Databases

There are a number of databases of IDPs available; three are indicated in Table 8.8, along with their respective URLs.

Figure 8.15 shows a screenshot of the DisProt database home page. It is a curated database. The current

TABLE 8.8 IDP Databases

	URL
DisProt	http://www.disprot.org/[69]
IDEAL	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/[70]
MobiDB	http://mobidb.bio.unipd.it/[71]

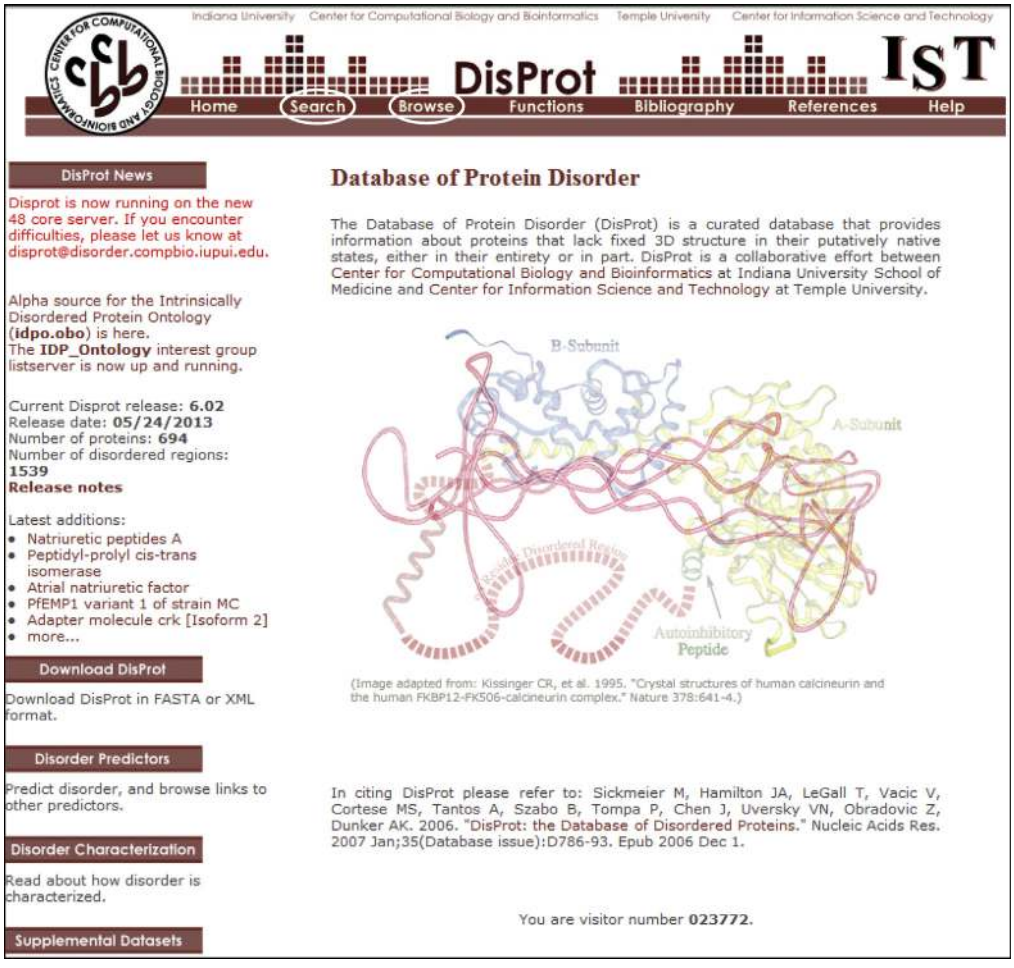


FIGURE 8.15 **Screenshot of the DisProt database home page.** On the left it displays the release number and the number of entries in the database. The entire database can be browsed by clicking the "Browse" link from the home page (circled). Alternatively, clicking the "Search" link (circled) takes the user to the search page, where a specific search can be launched (see text for details).

version (release 6.02) of the database has 694 proteins and a total of 1539 disordered regions. Clicking the "Search" link (circled) takes the user to the search page. An unknown sequence can be searched for the presence of a potential disordered segment by local-similarity search with other known disordered proteins from the database. Alternatively, a search can be launched by typing a keyword. In the absence of any specific search term, simply typing the keywords "signaling" or "regulation" will return a series of relevant entries from the database. An entry can be clicked to obtain more information, such as general information about the protein, sequence, percentage of the sequence that is disordered, map of the ordered and disordered segments, details of the disordered segments, and the references. The entire database can also be browsed by clicking the "Browse" link from the home page (circled). The other databases can also be searched/browsed in a similar fashion.

8.11.2 IDP Prediction

A number of online tools are also available to analyze a protein sequence for the existence of potentially disordered segments. Some of these tools are mentioned in Table 8.9, along with their respective URLs.

Figure 8.16 shows the DisProt disorder-prediction launch page. The sequence is pasted in the box, the desired analysis algorithm is checked, and the sequence is submitted for analysis. The Slco1a6 sequence was analyzed separately using VSL2B, VLXT, and PONDR-FIT. Because three different screenshots could not be

TABLE 8.9 Online Tools for IDP Prediction

Online Tool	Comments and URL
PONDR-FIT	Artificial-neural-network-based meta-predictor developed by combining several individual disorder predictors, such as PONDR-VLXT, PONDR-VSL2, PONDR-VL3, FoldIndex, IUPred, and TopIDP[72] (http://www.disprot.org/metapredictor.php)
DisEMBL	Artificial-neural-network-based. Trained for predicting several definitions of disorder, such as *loops/coils* as defined by DSSP*[73]; *hot loops*, i.e. the loops with a high B-factor from X-ray crystal structure†; *missing coordinates* (disordered regions) in X-ray structure as defined by REMARK465 entries in PDB, which indicate missing residues listed[74] (http://dis.embl.de/)
DISOPRED2	The link for PSIPRED analysis workbench is http://bioinf.cs.ucl.ac.uk/psipred/?disopred = 1. Check the box for DISOPRED2 in order to predict disordered protein
RONN	Bio-basis function neural network (BBFNN)-based. In BBFNN, the prediction is based on the likelihood of disorder determined by the alignment of the target sequence to a large group of sequences of known folding state (including known state of disorder)[75] (http://www.strubi.ox.ac.uk/RONN)

**DSSP (Dictionary of Secondary Structure of Proteins) is a program and database developed to standardize secondary-structure assignment for proteins of known 3D structure (hence entries in PDB database). DSSP describes eight states of protein secondary structure with single-letter codes: G (3/10 helix), H (α-helix), I (pi-helix), B (β-bridge), E (extended strand in β-sheet), S (bend), T (H-bonded turn), and C (coil).*
†In X-ray crystallography, the B-factor (temperature factor) is a measure of the extent of oscillation or vibration of an atom around the position specified in the model. So, a higher B-factor means more spread-out (lower) electron density, which indicates greater flexibility and disorder of the region.

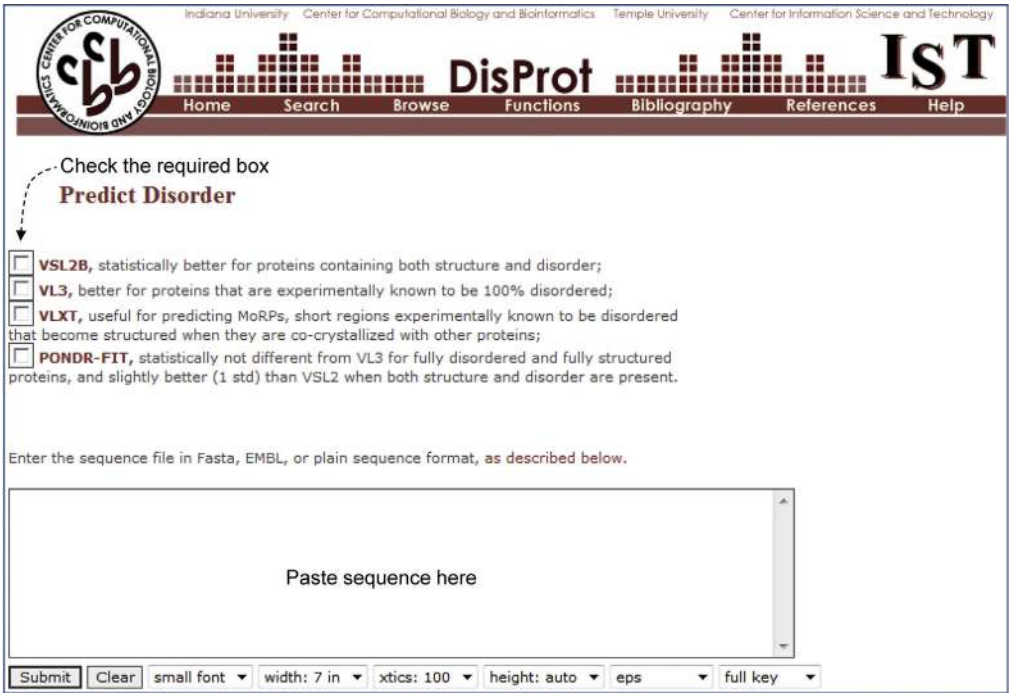


FIGURE 8.16 The DisProt disorder-prediction launch page. Providing options for analysis using PONDR-VSL2B, PONDR-VL3, PONDR-VLXT, and PONDR-FIT.

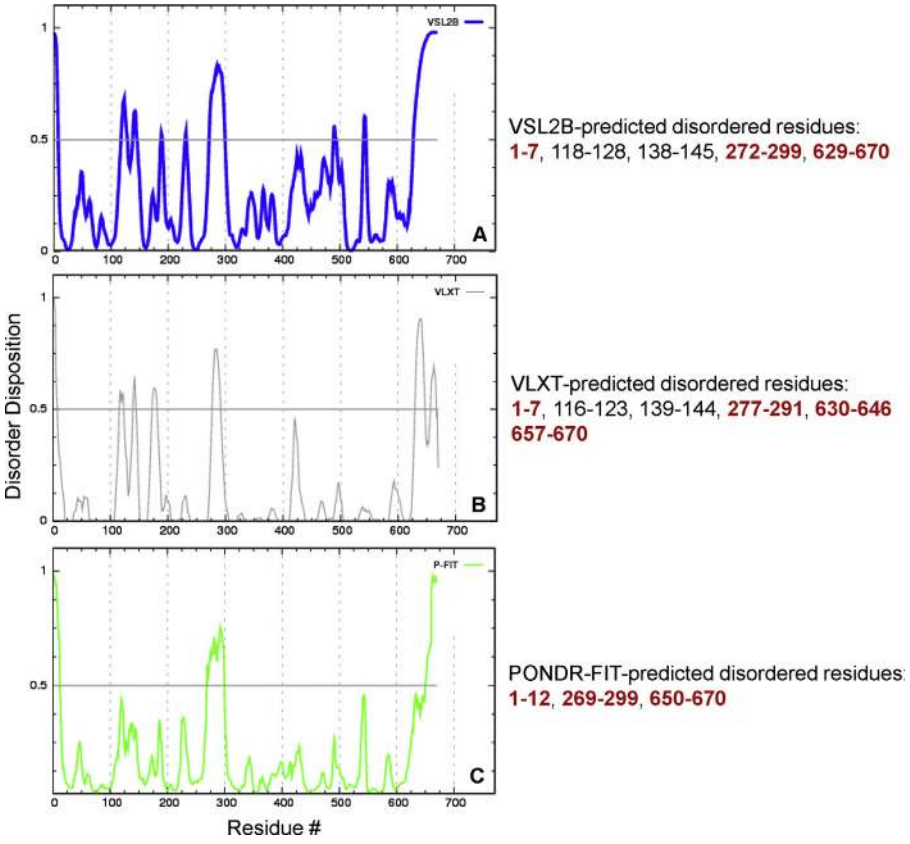


FIGURE 8.17 The Slco1a6 sequence, analyzed separately using VSL2B, VLXT, and PONDR-FIT. The graphical outputs of the analysis are shown. All three algorithms predict three regions of Slco1a6 to be potentially disordered: a very small region at the N-terminal end (around 1−10), a region in the middle (around 270−300), and at the C-terminal end (around 630−670). These predicted common residues are shown in red.

accommodated in one figure, only the graphical outputs of the analysis are shown, in Figure 8.17. All three algorithms predict three regions of Slco1a6 to be disordered. These predicted common residues are shown in red (Figure 8.17).

A separate analysis using RONN predicted three regions of disorder: 120−147, 272−299, and 630−670 (output not shown). Another analysis, using DisEMBL, predicted two regions of disorder: 279−296 and 640−670. Thus, different analysis programs consistently predicted two segments of Slco1a6 as potentially disordered regions: around 275−300 and around 635−670. Both these regions of Slco1a6 are part of the inside (cytoplasmic) segments, as predicted by RHYTHM, OCTOPUS, and Phobius (Figures 8.9 and 8.10).

References

1. Vieira-Pires RS, Morais-Cabral JH. *J Gen Physiol* 2010;**136**:585−92.
2. Cooley RB, et al. *J Mol Biol* 2010;**404**:232−46.
3. Leszczynski JF, Rose GD. *Science* 1986;**234**:849−55.
4. Chou KC. *Anal Biochem* 2000;**286**:1−16.
5. Ring CS, et al. *J Mol Biol* 1992;**224**:685−99.
6. Hovmöller S, et al. *Acta Crystallogr D Biol Crystallogr* 2002;**58** (Pt 5):768−76.
7. Kleywegt GJ, Jones TA. *Structure* 1996;**4**:1395−400.
8. Gasteiger E, et al. In: Walker JM, editor. *The proteomics protocols handbook*. Totowa, NJ: Humana Press; 2005. p. 571−607.
9. Varshavsky A. *Genes Cells* 1997;**2**:13−28.
10. Guruprasad K, et al. *Protein Eng* 1990;**4**:155−61.
11. Ikai AJ. *J Biochem* 1980;**88**:1895−8.
12. Kyte J, Doolittle RF. *J Mol Biol* 1982;**157**:105−32.
13. Hopp TP, Woods KR. *Proc Natl Acad Sci USA* 1981;**78**:3824−8.
14. Jääskeläinen S, et al. *Int J Data Min Bioinform* 2010;**4**:735−54.
15. Kloczkowski A, et al. *Proteins* 2002;**49**:154−66.
16. Chou PY, Fasman GD. *Biochemistry* 1974;**13**:222−45.
17. Garnier J, et al. *Methods Enzymol* 1996;**266**:540−53.
18. Garnier J, et al. *J Mol Biol* 1978;**120**:97−120.
19. Combet C, et al. *Trends Biochem Sci* 2000;**25**:147−50.
20. Bystroff C, Shao Y. *Bioinformatics* 2002;**18**(Suppl. 1):S54−61.
21. Bystroff C, et al. *J Mol Biol* 2000;**301**:173−90.
22. Cole C, et al. *Nucl Acids Res* 2008;**35**(Suppl. 2):W197−201.
23. Rost B, Sander C. *J Mol Biol* 1993;**232**:584−99.
24. Rost B, Sander C. *Proc Natl Acad Sci USA* 1993;**90**:7558−62.
25. Ouali M, King RD. *Protein Sci* 2000;**9**:1162−76.
26. Montgomerie S, et al. *BMC Bioinformatics* 2006;**7**:301.
27. Jones DT. *J Mol Biol* 1999;**292**:195−202.
28. Cheng J, et al. *Nucl Acids Res* 2005;**33**:72−6 (Web Server issue).
29. Baldi P, et al. *Bioinformatics* 1999;**15**:937−46.
30. Pollastri G, et al. *Bioinformatics* 2001;**17**(Suppl. 1):S234−42.
31. Simossis VA, Heringa J. *Comput Biol Chem* 2004;**28**:351−66.
32. Geourjon C, Deleage G. *Comput Appl Biosci* 1995;**11**:681−4.
33. Lin K, et al. *Bioinformatics* 2005;**21**:152−9.
34. Raghava GPS. *CASP* 2000;**4**:75−6.
35. Lupas A, et al. *Science* 1991;**252**:1162−4.
36. McDonnell AV, et al. *Bioinformatics* 2006;**22**:356−8.
37. Bornberg-Bauer E, et al. *Nucl Acids Res* 1998;**26**:2740−6.

38. Quevillon E, et al. *Nucl Acids Res* 2005;**33**:W116–20 (Web Server issue).
39. Marchler-Bauer A, et al. *Nucl Acids Res* 2013;**41**:D348–52 (Database issue).
40. Rimphanitchayakit V, Tassanakajon A. *Dev Comp Immunol* 2010;**34**:377–86.
41. Cera ED. *IUBMB Life* 2009;**61**:510–5.
42. Pao SS, et al. *Microb Mol Biol Rev* 1998;**62**:1–34.
43. Reddy VS, et al. *FEBS J* 2012;**279**:2022–35.
44. Sigrist CJ, et al. *Nucl Acids Res* 2013;**41**:D344–7 (Database issue).
45. Sigrist CJA, et al. *Brief Bioinform* 2002;**3**:265–74.
46. Krogh A, et al. *J Mol Biol* 2001;**305**:567–80.
47. Rose A, et al. *Nucl Acids Res* 2009;**37**:W575–80 (Web Server issue).
48. Käll L, et al. *J Mol Biol* 2004;**338**:1027–36.
49. Viklund H, Elofsson A. *Bioinformatics* 2008;**24**:1662–8.
50. Marsh DG, et al. *Bull World Health Org* 1986;**6**:767–70.
51. King TP, et al. *J Allergy Clin Immunol* 1995;**96**:5–14.
52. Gendel SM. *Regulat Toxicol Pharmacol* 2009;**54**:S7–10.
53. Vila L, et al. *Clin Exp Allergy* 2001;**31**:1599–606.
54. Wang J, Sampson HA. *J Clin Invest* 2011;**121**:827–35.
55. Roth-Walter F, et al. *Mol Nutr Food Res* 2013;**57**:536–44.
56. Taylor SL. *Annu Rev Pharmacol Toxicol* 2002;**42**:99–112.
57. FAO/WHO. Evaluation of allergenicity of genetically modified foods. Available online at: <http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf>; 2001.
58. Burks AW, et al. *Eur J Biochem* 1997;**245**:334–9.
59. Stanley JS, et al. *Arch Biochem Biophys* 1997;**342**:244–53.
60. Rost B. *Protein Eng* 1999;**12**:85–94.
61. Sander C, Schneider R. *Proteins* 1991;**9**:56–68.
62. Ivanciuc O, et al. *Bioinformatics* 2002;**18**:1358–64.
63. Ivanciuc O, et al. *Nucl Acids Res* 2003;**31**:359–62.
64. Saha S, Raghava GPS. *Nucl Acids Res* 2006;**34**:W202–9.
65. Bhasin M, Raghava GPS. *Vaccine* 2004;**22**:3195–204.
66. Tompa P. *Trends Biochem Sci* 2002;**27**:527–33.
67. Tompa P. *Trends Biochem Sci* 2012;**37**:509–16.
68. Uversky VN, Dunker AK. *Biochim Biophys Acta* 2010;**1804**:1231–64.
69. Sickmeier M, et al. *Nucl Acids Res* 2007;**35**:D786–93 (Database issue).
70. Fukuchi S, et al. *Nucl Acids Res* 2012;**40**:D507–11 (Database issue).
71. Di Domenico T, et al. *Bioinformatics* 2012;**28**:2080–1.
72. Xue B, et al. *Biochim Biophys Acta* 2010;**1804**:996–1010.
73. Kabsch W, Sander C. *Biopolymers* 1983;**22**:2577–637.
74. Linding R, et al. *Structure* 2003;**11**:1453–9.
75. Yang ZR, et al. *Bioinformatics* 2005;**21**:3369–76.

CHAPTER

9

Phylogenetic Analysis*

OUTLINE

9.1 PHYLOGENETICS AND THE WIDESPREAD USE OF THE PHYLOGENETIC TREE

Phylogeny refers to the evolutionary history of species. **Phylogenetics** is the study of phylogenies—that is, the study of the evolutionary relationships of species. **Phylogenetic analysis** is the means of estimating the evolutionary relationships. In molecular phylogenetic analysis, the sequence of a common gene or protein can be used to assess the evolutionary relationship of species. The evolutionary relationship obtained from phylogenetic analysis is usually depicted as branching, treelike diagram—the **phylogenetic tree**. Historically, the use of phylogenetic trees was restricted more or less to the study of evolutionary biology, and to disciplines like systematics and taxonomy. However, with the advent of sequencing and the widespread use of cladistics, the use of phylogenetic trees has pervaded many branches of biology and beyond. Construction of phylogenetic/evolutionary trees is now widespread in many areas of study where evolutionary divergence can be studied and demonstrated; be it pathogens, biological macromolecules, or languages.

Phylogenetics also provides the basis for **comparative genomics**, which is a more recent term that came into existence in the age of genomics. Comparative genomics is the study of the interrelationships of genomes of different species. Comparative genomics helps identify regions of similarity and differences among genomes. The comparison can be made at different levels, such as comparison of whole-genome sequences, comparison of genome sequences involving blocks of conserved synteny, comparison of the number of protein-coding genes, comparison of regulatory sequences, or other focused comparisons. An important application of comparative genomics is gene finding. From the standpoint of evolutionary biology, comparative genomics helps understand the evolutionary relationships among genomes.

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

A resource for comparative genomic analysis is **VISTA**, which can be accessed at http://genome.lbl.gov/vista/index.shtml.

9.2 PHYLOGENETIC TREES

A phylogenetic tree or evolutionary tree is a diagrammatic representation of the evolutionary relationships among various taxa (Figure 9.1 A–D). It is a branching diagram composed of **nodes** and **branches**. The branching pattern of a tree is called the **topology** of the tree. The nodes represent taxonomic units, such as species (or higher taxa), populations, genes, or proteins. A branch is called an edge, and represents the time estimate of the evolutionary relationships among the taxonomic units. One branch can connect only two nodes. In a phylogenetic tree, the terminal nodes represent the **operational taxonomic units** (**OTU**s) or **leaves**. The OTUs are the actual objects—such as the species, populations, or gene or protein sequences—being compared, whereas the internal nodes represent **hypothetical taxonomic units** (**HTU**s). An HTU is an inferred unit and it represents the **last common ancestor** (**LCA**) to the nodes arising from this point. Descendants (taxa) that split from the same node form **sister groups**, and a taxon that falls outside the **clade**[a] is called an **outgroup**. For example, in Figure 9.1 B, T_2 and T_3 are sister groups, and T_1 is an outgroup to T_2 and T_3.

Phylogenetic trees can be **scaled** or **unscaled**. In a scaled tree, the branch length is proportional to the amount of evolutionary divergence (e.g. the number of nucleotide substitutions) that has occurred along that branch. In an unscaled tree, the branch length is not proportional to the amount of evolutionary divergence, but usually the actual number is indicated somewhere on the branch.

Phylogenetic trees can be **rooted** (Figure 9.1 A and B) or **unrooted** (Figure 9.1 C). A rooted tree has a node (the root) from which the rest of the tree diverges.

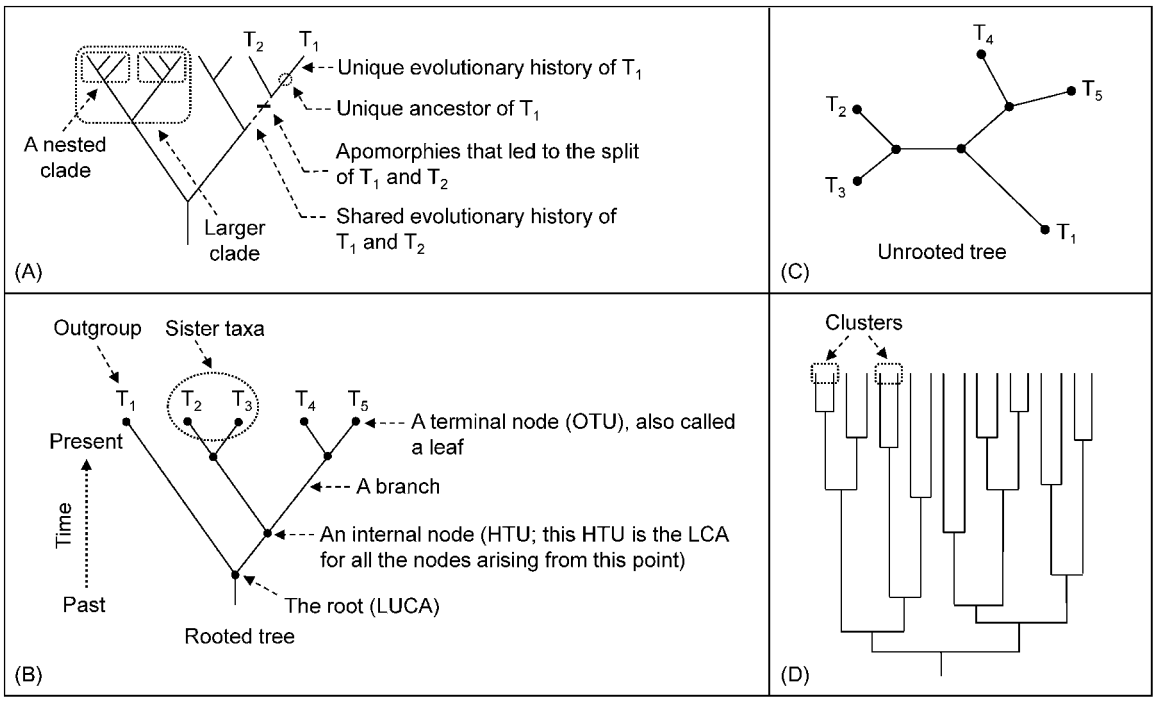


FIGURE 9.1 **Different forms of presentation of the phylogenetic tree.** The phylogenetic tree in D is a dendrogram derived from hierarchical clustering (see text). A, B, and D show rooted trees, while C shows an unrooted tree. Taxa that share specific derived characters are grouped into clades. (A) Smaller clades located within a larger clade are called nested clades. (B) The terminal nodes represent the operational taxonomic units, also called "leaves"; each terminal node could be a taxon (species or higher taxa), or a gene or protein sequence. The internal nodes represent hypothetical taxonomic units. An HTU represents the last common ancestor to the nodes arising from this point. Two descendants that split from the same node are called sister groups and a taxon that falls outside the clade is called an outgroup. Rooted trees have a node from which the rest of the tree diverges, frequently called the last universal common ancestor (LUCA).

[a]Taxa that share specific derived characters are grouped more closely together than those who do not. The groups are called **clades**; each clade consists of an ancestor and all of its descendants.

This root is frequently referred to as the **last universal common ancestor** (**LUCA**), from which the other taxonomic groups have descended and diverged over time. In molecular phylogenetics, the LUCA and LCA are represented by DNA or protein sequences. Obtaining a rooted tree is ideal, but most phylogenetic-tree-reconstruction algorithms produce unrooted trees.

9.2.1 Phylogenetic Trees, Phylograms, Cladograms, and Dendrograms

In the context of molecular phylogenetics, the expressions phylogenetic tree, phylogram, cladogram, and dendrogram are used interchangeably to mean the same thing—that is, a branching tree structure that represents the evolutionary relationships among the taxa (OTUs), which are gene/protein sequences. In the traditional evolutionary sense, the OTUs in the phylogenetic tree are represented by species. A **phylogram** is a scaled phylogenetic tree in which the branch lengths are proportional to the amount of evolutionary divergence. For example, a branch length may be determined by the number of nucleotide substitutions that have occurred between the connected branch points. A **cladogram** is a branching hierarchical tree that shows the relationships between clades; cladograms are unscaled. The word **dendrogram** means a hierarchical cluster arrangement where similar objects (based on some defined criteria) are grouped into clusters; hence, a dendrogram shows the relationships among various clusters (Figure 9.1 D). Dendrograms are also used outside the scope of phylogenetics and even outside of biology. Dendrograms are fequently used in computational molecular biology to illustrate the branching based on clustering of genes or proteins.

9.3 PHYLOGENETIC ANALYSIS TOOLS

The most convenient way to construct a phylogenetic tree is to use online tools. A good online phylogenetic analysis tool is available at **Phylogeny.fr** (http://www.phylogeny.fr/). This server provides "robust phylogenetic analysis for the non-specialist." The user can build a phylogenetic tree using the "**One Click**" option with all the default settings. Another tool for phylogenetic-tree construction is **MEGA** version 5[1] (as of October 2013). MEGA stands for **Molecular Evolutionary Genetics Analysis**, and it was developed by a group of well-known evolutionary biologists. MEGA can be downloaded from http://www.megasoftware.net/. MEGA is easy to operate, the toolbar is self-explanatory, and there are instructions provided. A recent publication by Hall[2] is also a good resource to understand MEGA. Another widely used and versatile downloadable software tool is **PHYLIP** (**Phylogenetics Inference Package**), which is a free package of programs for inferring phylogenies. It was developed by Joseph Felsenstein of the University of Washington (http://evolution.genetics.washington.edu/phylip.html). A widely used and affordable *commercial* software program for phylogenetic analysis is **PAUP** (**Phylogenetic Analysis Using Parsimony** (and Other Methods)), written by David Swofford. Another downloadable phylogenetic software tool is **MacClade** (http://macclade.org/macclade.html), written by David Maddison and Wayne Maddison. On the MacClade link, click on "Acquiring MacClade" or access the downloadable link directly at http://macclade.org/download.html.

There are several other phylogenetic analysis tools available on the web. Many of these require special formatting of data for entry, and they send the results through e-mail instead of providing real-time display of results. These tools can be checked out at the following link: http://molbiol-tools.ca/Phylogeny.htm.

9.4 PRINCIPLES OF PHYLOGENETIC-TREE CONSTRUCTION

Although a number of online resources have been mentioned above that can be used to construct/reconstruct phylogenetic trees, it is nevertheless important to understand the assumptions and steps involved in phylogenetic-tree construction for conceptual clarity.

There are certain assumptions behind making a phylogenetic tree, such as (1) the sequences are homologous—that is, the sequences share a common ancestry and they diverged through time as they evolved—and (2) each position evolved independently. The quality of multiple sequence alignment is the key to obtaining a reliable phylogenetic tree. *When using coding sequences, it is desirable to use the protein sequences to reconstruct the phylogenetic tree*.

Construction of a phylogenetic tree involves the following steps: (1) Selection of the appropriate molecular marker (genes/proteins/mitochondrial DNA), (2) Multiple sequence alignment, (3) Selection of a model of evolution, (4) Construction of the phylogenetic tree, (5) Assessment of the reliability of the tree.

9.4.1 Selection of the Appropriate Molecular Marker

The choice of nucleic acid or protein sequences as the appropriate marker depends on the need. A molecular marker in phylogenetic analysis is the

biological information that is used to infer the evolutionary relationships among taxa. In general, when coding sequences are used, it is desirable to use protein sequences to construct the phylogenetic tree. Some of the reasons why protein sequences are more appropriate are as follows:

1. There are more possible character states for amino acids (20) than nucleotides (4); the terminals may share a character state by chance simply because a given position can have only one of 4 possible character states (as opposed to 20 for amino acids).
2. Amino-acid-substitution matrices are more sophisticated than nucleotide-substitution matrices.
3. The existence of codon bias for the same amino acid in different species might artificially inflate the nucleotide sequence variation.

However, nucleotide sequences can also be used under certain circumstances to obtain a reliable tree, such as when comparing genes whose sequences are highly conserved among species, or comparing the evolution of genes in geographically separated populations within a species. Slowly evolving gene sequences can be used to assess the evolutionary relationship between distantly related species and, conversely, rapidly evolving gene sequences can be used for recently evolved species.

9.4.2 Multiple Sequence Alignment

Alignment of sequences is the most important step in constructing a reliable phylogenetic tree. Multiple sequence alignment identifies blocks of conserved residues. A good alignment should also have fewer gaps/long gaps. Gaps indicate sequences gained or lost (insertions–deletions) during evolution. The user may decide to use the entire alignment or use parts of it. There are no set rules regarding which sections of the alignment to remove; the user should apply judgment. If the alignment is ambiguous at the two ends, the ends can be removed. Such editing can also be done using **Gblocks**[3,4]. Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment to make it more suitable for phylogenetic analysis. Gblocks can be accessed at http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks, or at http://molevol.cmima.csic.es/castresana/Gblocks_server.html. The former link provides an example of how to enter the alignment data. The latter link provides an example of an output file showing the blocks selected from a protein alignment.

The "One Click" link of Phylogeny.fr (http://www.phylogeny.fr/) provides the option to utilize Gblocks to eliminate poorly aligned positions and divergent regions. This option is selected as part of the default settings. The user may choose to uncheck this option in order to use the entire sequence instead of the edited sequence.

9.4.3 Selection of a Model of Evolution

An evolutionary model of sequence data is a model of nucleotide or amino-acid substitution and consequent divergence of sequences. The evolutionary (substitution) models play an important role in the analysis of molecular sequence data. These models filter the complexity of the biological mutation process into simpler patterns that can be described and predicted using a small number of parameters. Substitution models attempt to predict the rate of substitution for nucleotides or amino acids at a given site, and also the distribution of substitutions across the entire sequence. The differential rate of substitutions across the sequence is called the **rate heterogeneity**.

Multiple alignment is followed by the selection of an appropriate evolutionary model. There are many such models. All statistical models are based on certain assumptions. One assumption is that each position in the nucleic acid or protein evolves independently. In reality, that is not the case; there are hot spots of mutation, and also some mutations are more tolerated than others.

The simplest way to determine divergence is to count the number of substitutions. However, there are caveats in such a simplistic approach. For example, an observed substitution (e.g. A→G) may not be the original substitution, but may have involved an intermediate substitution (e.g. A→T→G). Likewise, the absence of substitution at a position may also mean that an original substitution has been reversed (reverse mutation) during evolution to restore the original residue (e.g. A→G→A). Substitution models are statistical models that are supposed to correct for these biases. *Note that these methods are based on general mathematical and statistical principles that have their own set of assumptions.* The simplest substitution model *for nucleotides* is the **Jukes–Cantor (JC) one-parameter model**, which assumes that all nucleotides occur in equal frequency (25%) and are substituted with equal probability. This model requires a single parameter denoting rate. However, it is well known that transition mutations are more common than transversion mutations. **Kimura's two-parameter model** accounts for this, and proposes that transition mutations provide a better estimate of evolutionary divergence than transversion mutations. This model requires two parameters denoting rate. Like the Jukes–Cantor model, Kimura's model also assumes that all nucleotides occur in equal

frequency (25%). There are other more complex models of nucleotide substitution, such as the **Felsenstein model** and the **Hasegawa–Kishono–Yano (HKY) model**, which assume that nucleotides occur at different frequencies, and that transitions and transversions occur at different rates. The **general time reversible (GTR) model**, also known as the **general reversible (REV) model** is even more complex and assumes different rates of substitution for each pair of nucleotides, in addition to assuming different frequencies of occurrence of nucleotides. For these models, the nucleotide frequencies are estimated by the observed frequencies in the alignment. Some *amino acid substitution models* are the **Dayhoff model (PAM)**, the **Bishop–Friday model**, the **Jones–Taylor–Thornton (JTT) model**, the **Whelan and Goldman (WAG) model**, and the **Le Gascuel (LG) model**. The simplest model is the Bishop–Friday model, which assumes that all amino acids occur at equal frequency and all substitutions occur at the same rate. All other models assume different amino-acid frequencies and different substitution rates, which are experimentally determined.

The substitution model utilized for a particular data set can be displayed by the software, such as **MEGA** version 5[1] (discussed above).

9.4.4 Construction of the Phylogenetic Tree

The choice of an appropriate tree-building method for a given data set is a crucial but complex issue. Many methods have been described for reconstructing phylogenetic trees; each one has its own merits and demerits[5]. This is a highly specialized area of computation and statistics. Therefore, only some overall principles are discussed here. The methods to construct phylogenetic trees can be classified into two major types: (1) **distance-based** and (2) **character-based**, also called the **discrete method**.

9.4.4.1 Distance-Based (Distance-Matrix) Methods

In distance-based methods, the distance between each pair of sequences is calculated, and a distance matrix is computed. This distance matrix is used for tree construction. Distance-based methods use substitution models; hence, they are model based. Figure 9.2 A shows a simple distance matrix of four 10-nt-long sequences that differ from one another by 1, 2, 3, or 4 nucleotides. These nucleotide differences are used to compute the evolutionary distances among these sequences. There are two popular distance-based methods, the **unweighted pair group method with arithmetic mean (UPGMA)** and **neighbor joining (NJ)**.

The **UPGMA** is the simplest distance-matrix method, and it employs sequential clustering to build a rooted phylogenetic tree. First, all sequences are compared through pairwise alignment to compute the distance matrix. Using this matrix, the two sequences with minimum distance are identified and clustered as a single pair. Next, the distance between this pair and all other sequences is recalculated to form a new matrix. Using this new matrix, the sequence that is closest to the first pair is identified and clustered. This process is repeated until all sequences have been incorporated in the cluster. Figure 9.2 B shows how an UPGMA tree is computed. *Because the process is "unweighted," all pairwise distance are assumed to contribute equally.*

The **neighbor-joining (NJ) method**[6] is the most widely used distance-matrix method. It starts with a star tree—that is, it is assumed that the branches leading to the respective OTUs (the sequences) radiate from one internal node forming a star-like pattern. Next, a pair of sequences is chosen at random, removed from the star, and attached to a second internal node which is connected by a branch to the center of the star-like pattern (Figure 9.3). The branch lengths are calculated. These two sequences are then returned to their original positions and another pair is selected to repeat the same operation. The goal of these repetitive operations until all possible pairs have been examined is to find out the combination of neighbors that minimizes the total length of the phylogenetic tree.

9.4.4.2 Character-Based Methods

In contrast to the distance-matrix methods, the character-based methods utilize the sequence itself rather than the pairwise distance obtained from the sequence features. A character is a site (position) in the alignment. There are two popular character-based methods, **maximum parsimony (MP) and maximum likelihood (ML)**.

The **maximum parsimony** method computes many trees from the given data set and assigns a cost to each tree. The assumption of maximum parsimony is that the simplest tree is the most plausible tree. The simplest tree is the one that requires the fewest number of changes to explain the data in the alignment. Thus, parsimony uses the data and does not attempt to use any model to estimate the total number of changes. The tree score is the sum of character lengths over all sites. If more than one tree with a smallest number of changes can be obtained, then the trees are said to be equally parsimonious. In maximum parsimony, the site (position of the sequence) that has *at least two different kinds of nucleotides (bases) represented in at least two of the sequences* is considered to be an informative site (Figure 9.4 A). Figure 9.4 B shows the principle of tree

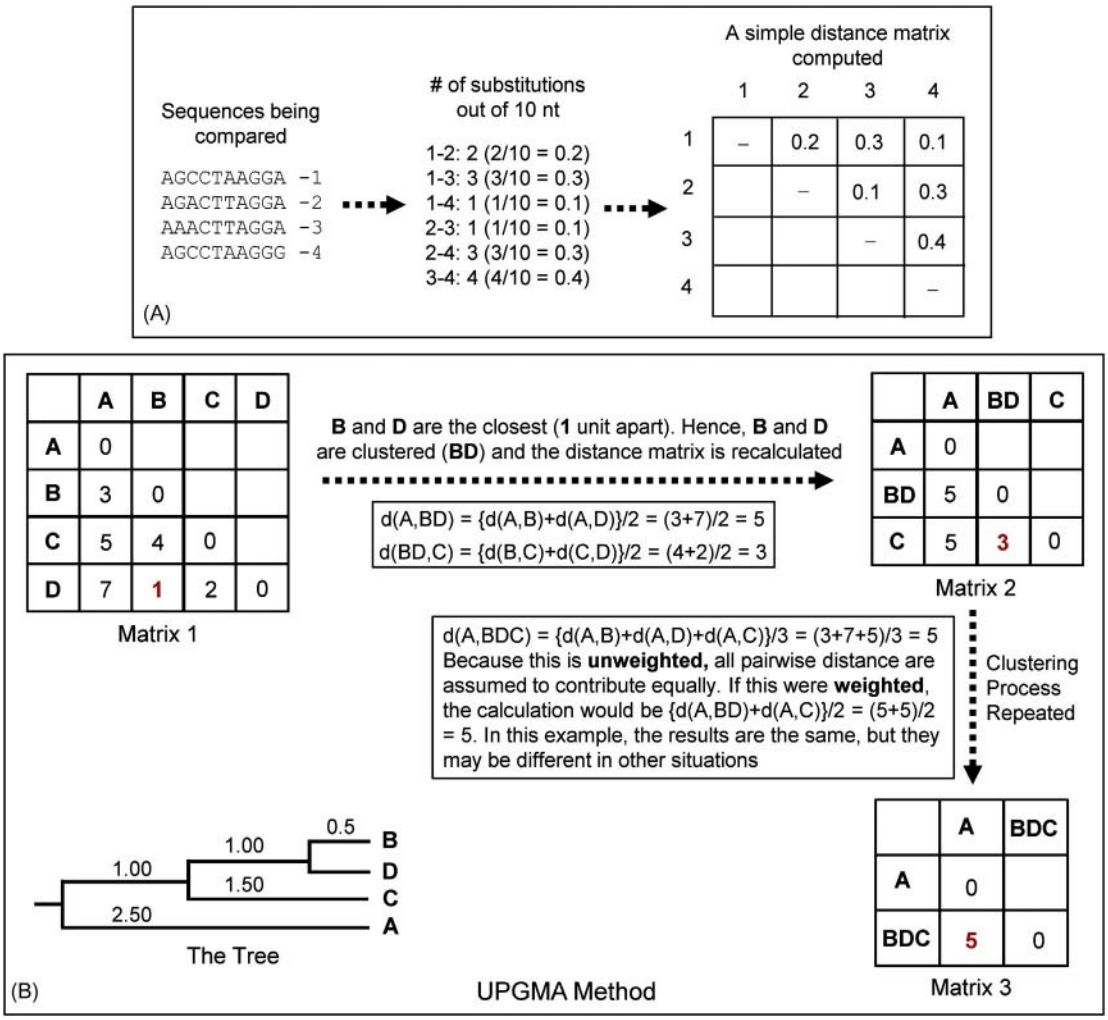


FIGURE 9.2 **Construction of phylogenetic tree using the distance-matrix method.** (A) A simple distance matrix of four 10-nt-long sequences is shown; the sequences differ from one another by 1, 2, 3, or 4 nucleotides. (B) The UPGMA method involves sequential clustering, with calculation of a new distance matrix at each step (see text).

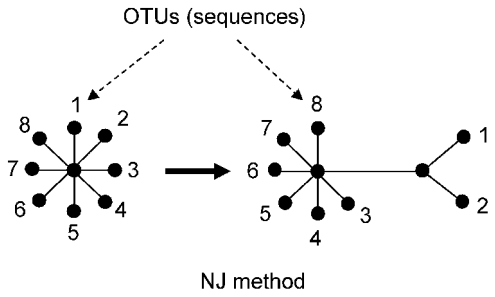


FIGURE 9.3 **Construction of phylogenetic tree using Saitou and Nei's neighbor-joining method.** See text for details.

construction by maximum parsimony using the informative sites (positions 7 and 9) of the sequences shown in Figure 9.4 A. The figure shows that tree 1 is the most parsimonious tree because its topology is based on the minimum number of mutations.

Maximum likelihood is a statistical method that estimates the unknown parameters of a probability model. The maximum-likelihood method is currently widely used for the construction of phylogenetic trees because of increased computational ability. Maximum likelihood evaluates the probability that the selected evolutionary model predicts the observed sequences. In other words, the topology of the phylogenetic trees constructed using maximum likelihood should yield the highest probability of producing the observed sequences.

The use of **Bayesian** phylogenetic analysis is far more recent than the maximum-parsimony and maximum-likelihood methods. The Bayesian phylogenetic method has gained considerable ground ever since the use of Bayesian statistics in phylogenetics was proposed in the mid-1990s. The Bayesian method draws inference on the probability of an unknown event by deriving a "posterior probability." Unlike

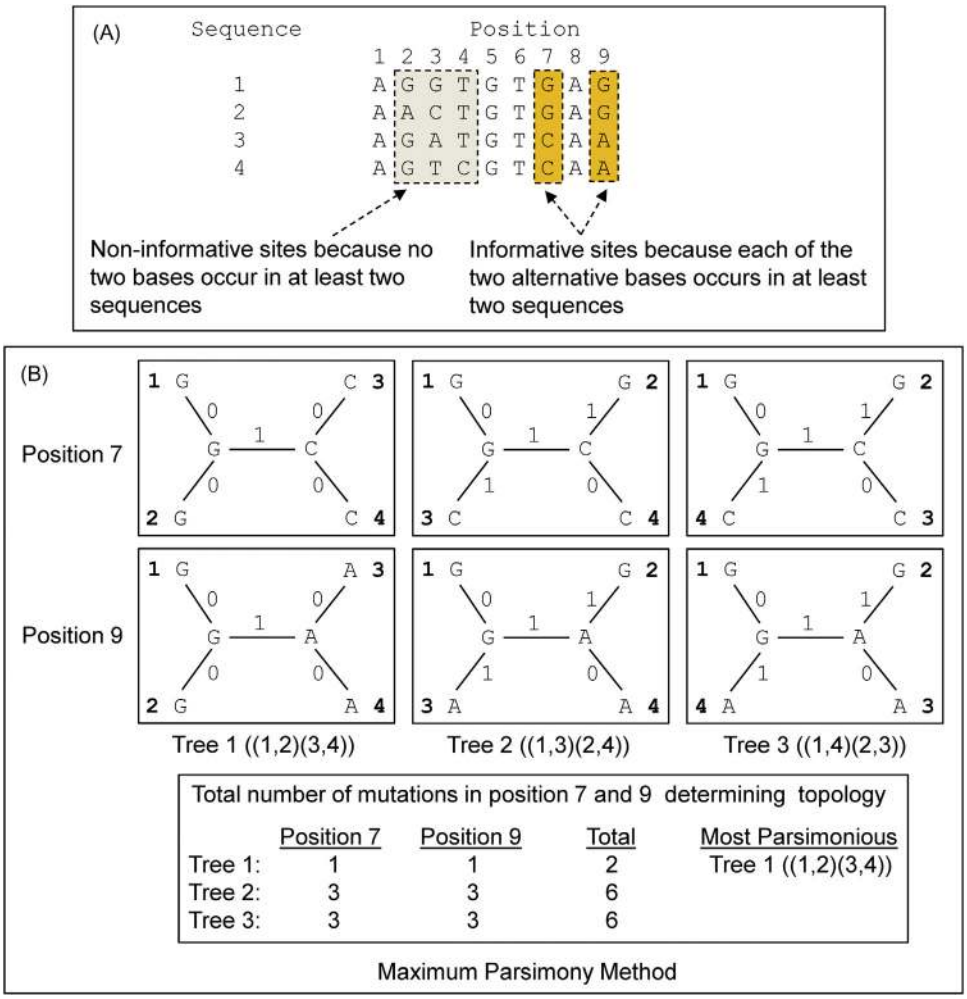


FIGURE 9.4 **The maximum parsimony method.** (A) Informative and non-informative sites considered in maximum parsimony. Non-informative sites do not have each of the alternative bases occurring in at least two sequences. In contrast, in an informative site, each of the alternative bases occurs in at least two sequences. (B) Principles of tree construction by the maximum parsimony method. Tree 1 is the most parsimonious tree because its topology is based on the minimum number of mutations (see text).

standard statistical tests, in which the existing data are used to test a hypothesis, Bayesian statistics uses prior knowledge, in addition to the existing data, to test a hypothesis. The prior knowledge/data provide an estimate of the prior probability of an event, whereas integrating the existing data with the prior probability helps estimate the posterior probability of the event. A prior probability might be derived based on a set of known principles or experimental results. Tree construction in the Bayesian method utilizes repetitive random sampling using a Markov chain Monte Carlo (MCMC) process, which seeks the tree topology with increasingly higher score with each repetitive sampling. Finally, the consensus tree with the highest posterior probability is built from a set of high-scoring tree topologies. The Bayesian method is faster than the ML method, and hence can handle large data sets. **MrBayes** is a Bayesian phylogenetic analysis tool. An online version is available at http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type = mrbayes. This link also shows the format of data entry. Alternatively, MrBayes can be downloaded from http://mrbayes.sourceforge.net/. MrBayes was written by John Huelsenbeck, Bret Larget, Paul van der Mark, Fredrik Ronquist, Donald Simon, and Maxim Teslenko (http://mrbayes.sourceforge.net/authors.php).

9.4.5 Assessment of the Reliability of a Phylogenetic Tree

Construction of a phylogenetic tree is followed by an assessment of the reliability of the tree. Determining the reliability of the tree means determining whether the topology of the tree is accurate or whether a better tree can be obtained. These questions are answered by **bootstrapping** the reconstructed tree.

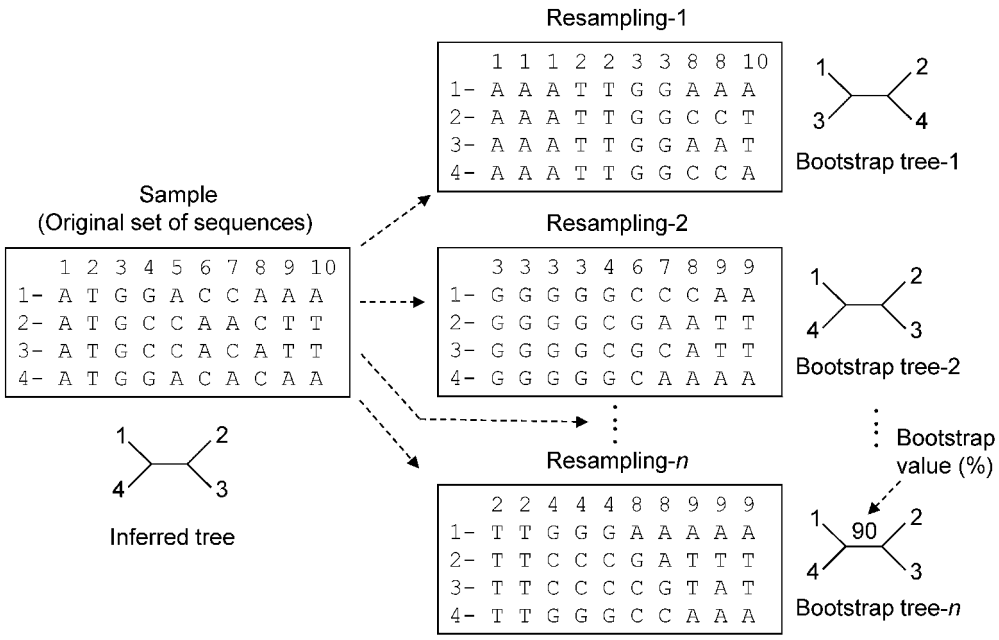


FIGURE 9.5 **Principles of bootstrapping the phylogenetic tree.** The bootstrap method involves repeated resampling (with replacement) from the original sample to create many new subsets of pseudosamples that are subjected to the same analysis as the original sample to obtain many bootstrap trees. The topology of these bootstrap trees is compared with that of the original tree to statistically assess the reliability of the original phylogenetic tree.

Felsenstein[7] first applied the bootstrap method to phylogenetic analysis to assess the reliability of the tree. (Phylogenetic) tree bootstrapping is a computationally performed statistical analysis, which is based on Efron's original bootstrap technique of resampling one's own data to infer the variability of the estimate. The bootstrap method involves repeated resampling (with replacement) from the original samples to create many new subsets of pseudosamples that are subjected to the same analysis as the original samples. The resampling with replacement means that some of the characters/data of the original samples will be in the bootstrap sample multiple times, whereas others will not appear at all. The statistical concept behind such resampling is that if a parameter can be estimated from samples drawn from a population, then the reliability of the estimate of that parameter can be verified by drawing new samples from the same population. The higher the number of resamplings, the greater is the confidence level of the estimate.

In the case of the bootstrap method using sequences, once the phylogenetic tree is constructed after aligning the original set of sequences, the sequences are repeatedly resampled to create many new subsets of derived sequences, i.e. the bootstrap samples. Each round of resampling (with replacement) of the original set of sequences creates a new subset of bootstrap samples of derived sequences. In each derived sequence, some of the bases from the original sequence will be represented multiple times, whereas other bases will not appear at all. One bootstrapping may perform 500–1000 such resamplings from the original sequences.

The derived sequences of each subset are then aligned and a new phylogenetic tree (bootstrap tree) is constructed using the same tree-construction method used to construct the original tree (e.g. neighbor-joining method, maximum-parsimony method, etc.). When the splitting pattern of an interior branch (branch topology) in the original tree is reproduced in the bootstrap tree, that branch is given a value of 1 (identity value). In other words, when an interior branch is given a value of 1, it is assumed to accurately predict the clade and the sister taxa, as reflected not only in the original tree but also in the bootstrap tree. Conversely, when the splitting pattern of an interior branch in the original tree is not reproduced in the bootstrap tree, that branch is given a value of 0. This process is repeated hundreds of times, and the percentage of times each interior branch is given a value of 1 is computed. This is known as the **bootstrap value** or **bootstrap confidence value**. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered accurate. Bootstrap values, expressed as percentages, are indicated on the branches. Therefore, a bootstrap value of 95 indicated on a branch means that 95% of the bootstrap trees support the topology at the branch obtained in the original phylogenetic tree. Figure 9.5 shows the principle of bootstrapping.

It should be remembered that, despite the rigor, the construction of phylogenetic trees is not exact and it involves general mathematical and statistical principles that have their own set of assumptions. As a result, many phylogenetic trees reconstructed from molecular sequences may conflict with common sense; they may be partially correct or even be incorrect[8].

9.5 MONOPHYLY, POLYPHYLY, AND PARAPHYLY

This concept relates to the groupings of organisms. If the classification is performed based on synapomorphic characters (shared derived characters), monophyletic groups are obtained. A monophyletic group includes the last common ancestor (LCA) plus all the descendants of the LCA. Monophyly can be assigned based on nodes as well as apomorphies (Figure 9.6). For example, mammals form a monophyletic group; so do birds, fish, etc. Monophyletic groups form clades and provide accurate information about the evolutionary history.

If the classification is performed based on homoplastic characters (similar characters that evolved independently in different groups through convergent evolution), polyphyletic groups are obtained. A polyphyletic group includes the descendants only and excludes the LCA, and the taxa are grouped based on superficial similarities (Figure 9.6). Thus, polyphyletic taxa could be evolutionarily very distant but linked by homoplasy. Polyphyletic groups do not provide any accurate information about the evolutionary history. In fact, once it is realized that a group of taxa are polyphyletic, they are reclassified. For example, birds and bats could form a polyphyletic group based on homeothermy and the ability to fly. Similarly, sharks and dolphins could form a polyphyletic group based on the ability to swim and other aquatic adaptations.

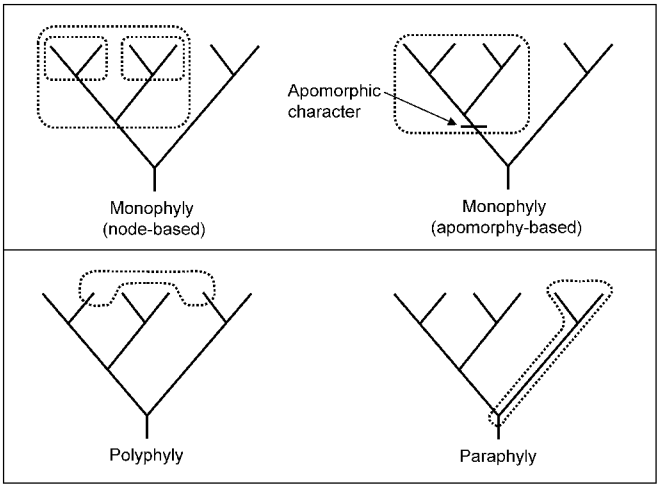


FIGURE 9.6 **Character-based classification to obtain monophyletic, polyphyletic, and paraphyletic groups.** A monophyletic group includes the last common ancestor (LCA) plus all the descendants of the LCA. A polyphyletic group includes the descendants only and excludes the LCA. A paraphyletic group includes the LCA but does not include one or more descendants.

If the classification is performed based on symplesiomorphic characters (shared ancestral characters), paraphyletic groups are obtained. A paraphyletic group includes the LCA but does not include one or more descendants. Therefore, a paraphyletic group is an incomplete clade and does not provide much information about the recent evolutionary history of the taxa concerned (Figure 9.6).

The terms polyphyly and paraphyly are of academic and historical interest. From the phylogenetic perspective, only monophyletic groups are important.

9.6 SPECIES TREES VERSUS GENE TREES

Phylogenetic trees can be constructed to depict the evolutionary history of species/populations or genes. A phylogenetic tree that shows the evolutionary history of species/populations is called a **species tree**. **Speciation** involves the splitting of an ancestral population into two populations that diverge and become reproductively isolated, giving rise to two species. Therefore, the branching in a species tree shows the time when the two species descended from the ancestral population and became reproductively isolated.

In contrast, when the phylogenetic tree is constructed based on a group of homologous gene sequences, where each sequence is sampled from a different species, then a gene tree is obtained. The general assumption is that gene trees are less ambiguous than species trees because gene trees are constructed based on definitive molecular data. However, the event that drives divergence between two populations leading to speciation is reproductive isolation, whereas the event that drives divergence between two homologous gene sequences is mutation. Mutations in genes and speciation do not necessarily happen at the same rate. Genetic polymorphism and multigene families add additional twists to the problem of gene tree to species tree extrapolation. When there is allelic polymorphism within species, a gene tree constructed from DNA sequences for a given gene can be quite different from the species tree, and this is particularly so when the time of divergence between different species is

short[9]. When the gene whose evolutionary history is being studied belongs to a multigene family, it may be difficult to correctly assign the homology of the sequences under study.

Therefore, inferring species trees from gene trees requires a great deal of caution. In general, gene trees are useful in studying the evolutionary history of the members a gene family, and inferring the evolutionary relatedness of the species from which the genes are obtained.

References

1. Tamura K, et al. *Mol Biol Evol* 2011;**28**:2731–9.
2. Hall BG. *Mol Biol Evol* 2013;**30** 1229-1135
3. Castresana J. *Mol Biol Evol* 2000;**17**:540–52.
4. Talavera G, Castresana J. *Syst Biol* 2007;**56**:564–77.
5. Yang Z, Rannala B. *Nat Rev Genet* 2012;**13**:303–14.
6. Saitou N, Nei M. *Mol Biol Evol* 1987;**4**:406–25.
7. Felsenstein J. *Evolution* 1985;**39**:783–91.
8. Lake JA, Moore JE. Trends guide to bioinformatics. *Trends J Suppl* 1998;**1998**:22–3.
9. Pamilo P, Nei M. *Mol Biol Evol* 1988;**5**:568–83.

Index

Note: Page numbers followed by "*f*", "*t*" and "*b*" refers to figures, tables and boxes respectively.