Chapter 1

Introduction to Bioinformatics

1.1 Introduction

To understand the functions of the human body, it is first necessary to understand the function of the basic unit of the body—the cell. The human body consists of trillions of cells that perform independent functions and are synchronized to carry out complex bodily functions. Scientists have dug into the functionality of cells, investigating how and why cells perform the tasks that they do. The study of the principles that govern these functions using modeling and computational techniques is the foundation of computational biology.

The human cell possesses hereditary material that is vital for cell replication and duplication and contains several parts, including a plasma membrane and various organelles, which are each designed to render both structure and function for the body (U.S. National Library of Medicine 2011) (Figure 1.1).

Typically, the plasma membrane, also called the lipid bilayer in animal cells, forms an outer lining called the plasma membrane of a cell. This membrane separates the cell from the rest of the environment and selectively allows materials to enter and leave the cell. It is also the characteristic difference between animal and plant cells, as the animal lipid bilayer is characteristically flexible, unlike the rigid plant plasma membrane. The flexibility of the plasma membrane in an animal cell membrane is brought about by its composition of lipid molecules that are characteristically polar, hydrophilic, or hydrophobic in nature. This diversity in composition allows the cell membrane to form various shapes, depending on changes in environmental conditions. The membrane of a cell is coated with

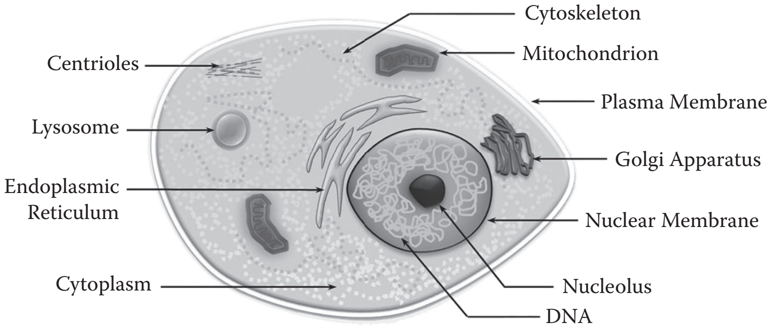


Figure 1.1 A schematic representation of the anatomy of the cell.

surface proteins, such as cell surface receptors, surface antigens, enzymes, and transporters, that bring about the functions of the membrane (Schlessinger and Rost 2005; Tompa 2005). These surface proteins are highly sensitive to the environment, as they are highly hydrophobic or hydrophilic. Research in identifying the structure and function of these membrane proteins has generated interest in recent times (Schlessinger et al. 2006).

The plasma membrane encases the cytoplasm and various organelles of the cell. The bulk of the cell is composed of cytoplasm, which is composed of cytosol (a jelly-like fluid), the nucleus, and other organelle structures. The largest organelle is the cytoskeleton, which is composed of long fibers that spread over the entire cell. Thus, the cytoskeleton provides the vital structure of the cell. Apart from providing the structure and shape of the cell, the cytoskeleton provides several critical functions, including the cell division and movement of the cell.

The endoplasmic reticulum is an organelle of the cell that is a collection of vesicles and tubules held together by the cytoskeleton. Also referred to as the lacey membrane, the endoplasmic reticulum can be one of three types: the rough endoplasmic reticulum (RER), the smooth endoplasmic reticulum (SER), or the sarcoplasmic reticulum (SR). Each of these types of endoplasmic reticulum has specific functions. The RER manufactures proteins through embedded structures known as ribosomes. Ribosomes are organelles that help create proteins by processing genetic instructions coded in the DNA of the nucleus. The ribosomes characteristically attach to the endoplasmic reticulum but, at times, float freely in the cytoplasm. The SER enables the synthesis of lipids and the metabolism of steroids. It is also responsible for regulating the calcium concentration throughout the cell. The SR, which is similar to the SER, functions as a calcium pump. Overall, the endoplasmic reticulum facilitates protein creation, folding, and the transport of the molecules that are in the form of sacs, referred to as the cisternae.

Other organelles in the cell, such as the Golgi apparatus, aid in the packaging of the processed molecules (proteins) from the endoplasmic reticulum for excretion from the cell; this is better known as the recycling center of the cell. Similarly, lysosomes are organelles that break down and digest toxic substances, engulfed bacteria, and viruses in a cell. They also maintain the proper functioning of the cell by recycling worn-out organelles. The organelle responsible for cell function is the mitochondrion, which is responsible for converting food to energy that can be used by the cell. The mitochondrion is a complex organelle that has its own genetic material (deoxyribonucleic acid (DNA)), which is different from the genetic material in the nucleus. This material is known as mitochondrial deoxyribonucleic acid (mtDNA) and enables the mitochondria to self-replicate.

The most important central command center of the cell is the nucleus that houses DNA, the heredity material of the cell. The DNA found in the nucleus is known as the nuclear DNA. Nuclear DNA stores genetic information in the form of a code consisting of four chemical bases, adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, more than 99% of which are the same in all people. Moreover, nearly every cell in the human body has the same DNA. The nucleus is enveloped by a membrane called the nuclear envelope that protects and separates the DNA from the rest of the cell organelles.

A closer inspection of the DNA sequence shows the existence of an order of the bases in the DNA sequence. This order determines the coded instructions for the cell to grow, mature, divide, or die. In the DNA, the bases A, C, T, and G combine to form base pairs, such as A and T or C and G. A nucleotide consists of an ensemble of these base pairs attached to a sugar molecule and a phosphate molecule (refer to Figure 1.2 for examples of these molecules). The nucleotides in a DNA molecule are arranged in two long strands to form a spiral called the double helix. The structure of DNA is analogous to that of a ladder, where the ladder rungs correspond to the base pairs while the sugar and phosphate molecules correspond to the vertical side pieces of the ladder. This double helix structure of the DNA molecule facilitates replication, and each strand serves as a pattern template for the duplication of sequence bases during cell division, as the resultant child cells should possess the exact copy of the DNA in the parent cell (Figure 1.2).

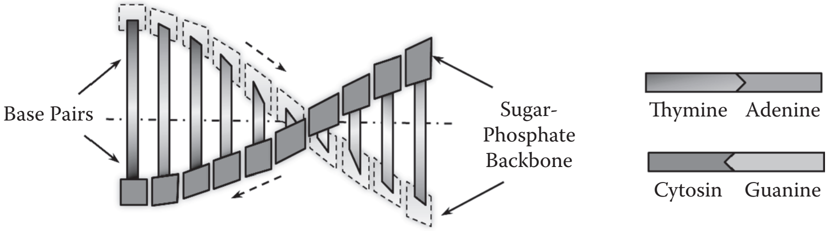


Figure 1.2 Schematic representation of the DNA double helix formed by base pairs attached to a sugar-phosphate backbone. (From http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure.jpg.)

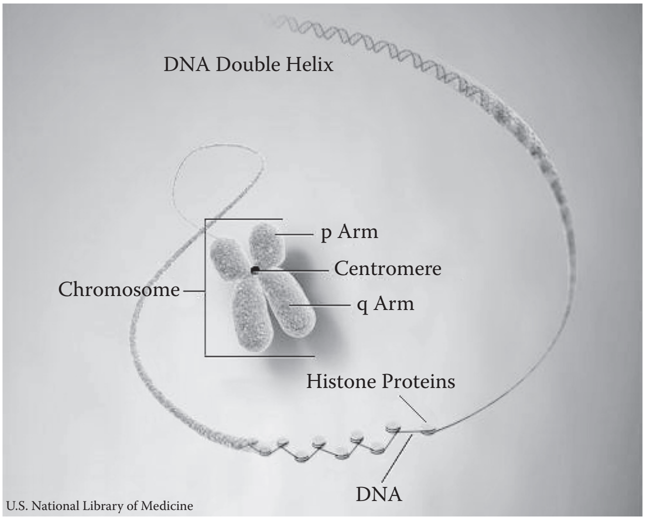
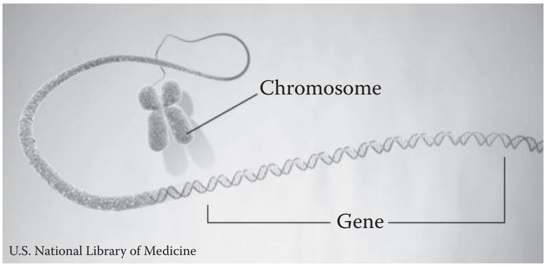


Figure 1.3 DNA and histone proteins are packaged into structures called chromosomes. (From http://ghr.nlm.nih.gov/handbook/illustrations/chromosomestructure.jpg.)

Chromosomes are thread-like structures that contain multiple, tightly packed DNA molecules. These tightly packed units are coiled multiple times around proteins called histones. These histone molecules are believed to provide the necessary structural reinforcement for the chromosome and help in the analysis of the structure of chromosomes. Typically, the structure of a chromosome consists of a central point called the centromere (refer to Figure 1.3), which divides the chromosome into sections called arms. The location of the centromere over the entire chromosome renders the characteristic shape of a chromosome, and acts as the point of reference in locating genes throughout the chromosome. Typically, a chromosome consists of two arms of different lengths. The shorter arm is referred to as the *p*-arm, and the longer is called the *q*-arm.

Genes are best known as the basic physical and functional units of heredity. They are found at characteristic locations over the chromosome; these locations are called loci. The coded information (i.e., the DNA) found in genes is translated and transcribed to create protein molecules.

Most humans share the same genes; however, a small number of genes vary from individual to individual. These genes provide individuals their unique characteristics, like hair, eye color, body shape, and skin pigmentation. A particular gene with two or more forms is called an allele. The difference in the gene is exhibited as changes in the DNA bases that contribute to an individual's unique physical features (Figure 1.4).



U.S. National Library of Medicine

Figure 1.4 **Genes are made up of DNA. Each chromosome contains many genes. (From http://ghr.nlm.nih.gov/handbook/illustrations/geneinchromosome.jpg.)**

Genes contain codes that are translated into proteins. During translation, the gene codes consisting of trinucleotide units called codons provide the necessary coding for an amino acid. Table 1.1 shows the triplet combinations of nucleotides that result in the creation of 20 known amino acids. The translation is initiated by a START codon (along with nearby initiation factors) and is terminated by a STOP codon. A sequence of amino acids forms a protein, which is a complex molecule that carries out critical functions in the human body. The function of the

Table 1.1 **All Amino Acids and Their Corresponding Codons**

Amino Acid	*Codon*	*Amino Acid*	*Codon*
Ala/A	GCU, GCC, GCA, GCG	Lys/K	AAA, AAG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	Met/M	AUG
Asn/N	AAU, AAC	Phe/F	UUU, UUC
Asp/D	GAU, GAC	Pro/P	CCU, CCC, CCA, CCG
Cys/C	UGU, UGC	Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Gln/Q	CAA, CAG	Thr/T	ACU, ACC, ACA, ACG
Glu/E	GAA, GAG	Trp/W	UGG
Gly/G	GGU, GGC, GGA, GGG	Tyr/Y	UAU, UAC
His/H	CAU, CAC	Val/V	GUU, GUC, GUA, GUG
Lle/I	AUU, AUC, AUA	START	AUG
Leu/L	UUA, UUG, CUU, CUC, CUA, CUG	STOP	UAA, UGA, UAG

complex protein molecule is determined by its sequence and its three-dimensional (3D) structure, which has direct bearings on the function of the associated gene.

The function of genes is, at times, affected by random changes to naturally occurring sequences. These changes are called mutations. Mutations are random changes in the structure or composition of DNA, which can be caused by mistakes in reproduction or external environmental events, like UV damage. While evolutionary changes in species are caused by beneficial mutations that enable organisms to adapt over time, not all mutations are beneficial. Certain mutations cause diseases such as cancer and could affect the survival of organisms and species over time.

A significant amount of biomedical research has been carried out to determine the functions of protein complexes for medical use. This research has resulted in breakthroughs in drug development.

Section 1.2 contains a description of transcription and translation, closely followed by an introduction to the Human Genome Project (HGP) in Section 1.3, which resulted in an estimate of between 20,000 and 25,000 genes reported in humans.

1.2 Transcription and Translation

The creation of proteins from a gene is complex and consists of two integral steps: transcription and translation. Though most genes contain the information needed to generate proteins, some genes help the cell assemble proteins. Transcription and translation are part of the central dogma of molecular biology, which is the fundamental principle that governs the conversion of information from DNA to RNA to protein (refer to Figure 1.5). The following section provides an overview of the two-stage process of transcription and translation.

The first step of transcription occurs in the nucleus of the cell where the information stored in the DNA (of a gene) is transferred to the mRNA (messenger ribonucleic acid). Typically, both RNA and DNA are composed of nucleotide base chains; however, they differ in properties and chemical composition. The mRNA is a type of RNA that holds the chemical blueprint of the protein product. The resultant protein product carries the encoded information from the DNA within the nucleus to the DNA within the cytoplasm of the cell for the production of the protein complex.

The second step of translation occurs outside the walls of the nucleus, in which the ribosomes present on the rough endoplasmic reticulum read the encoded information from the mRNA to produce the protein. The mRNA sequence consists of a string of codons, three bases that represent independent amino acids. The assembly of amino acids into the corresponding protein sequence is brought about by the transfer RNA (tRNA) one amino acid at a time. This process of assembly continues until the stop codon in the mRNA is encountered. This two-step process is called the central dogma of molecular biology (refer to Figure 1.5).

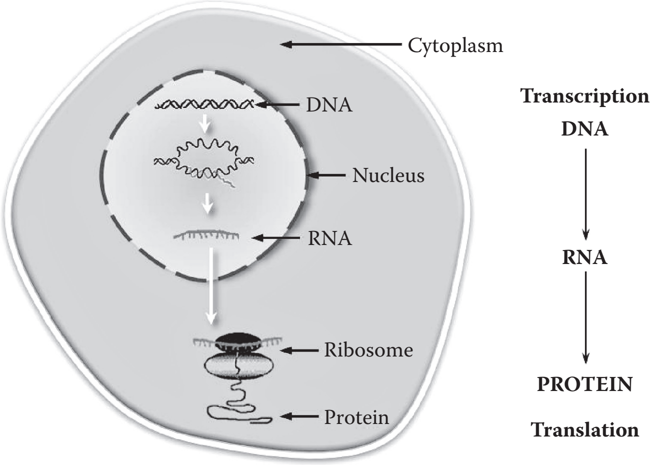


Figure 1.5 The central dogma of molecular biology. The processes of transcription and translation of information from genes are used to make proteins. (From http://ghr.nlm.nih.gov/handbook/illustrations/proteinsyn.jpg.)

1.2.1 The Central Dogma of Molecular Biology

As described previously, each gene contains the genetic makeup of an individual and the coded information required to manufacture both noncoding RNA and proteins. The expression of a gene is carried out by the two-stage process of translation and transcription (refer to Figure 1.6).

The first step in this process is called transcription, which involves the replication of gene content by copying the content of the DNA to an equivalent RNA molecule also known as the primary transcript. The primary transcript is essentially the same sequence as the gene, except that it is complementary in its base pair content. This similarity enables the sequence to convert from DNA and RNA and vice versa, in the presence of certain enzymes. The resultant RNA sequence reflecting the transcribed DNA is called a transcription unit encoding one gene. The nucleotide composition of the resultant RNA includes uracil (U) in place of thymine (T) in the DNA complement. DNA transcription is regulated and directed by regulatory sequences. The DNA sequence before the coding sequence is called the five prime untranslated region (5’UTR); similarly, the sequence following the coding sequence is called the three prime untranslated region (3’UTR). The direction of transcription moves from the 5’ to the 3’. Each gene is further divided into intermediate regions called exons and introns. The exons carry information required for protein synthesis. As shown in Figure 1.6, the messenger RNA (mRNA) contains information from the exons. The process of splicing filters out the intron sequence from the primary transcripts.

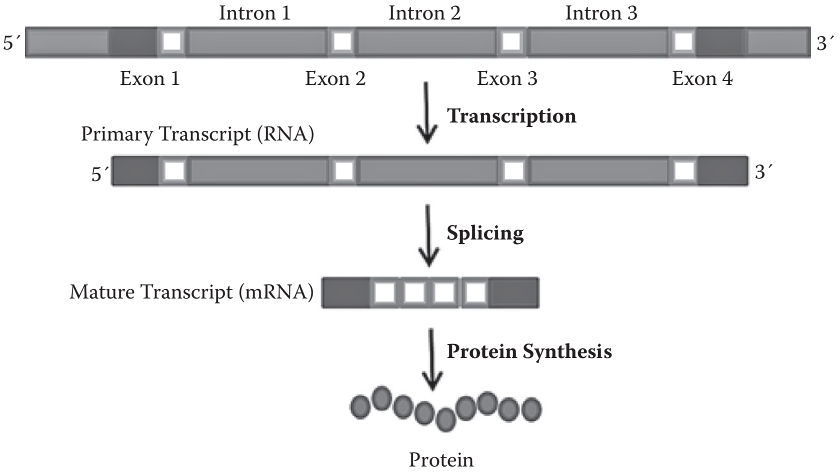


Figure 1.6 An overview of the transcription to translation. The gene is first transcribed to yield a primary transcript, which is processed to remove the introns. The mature transcript (mRNA) is then translated into a sequence of amino acids, which defines the protein. (From http://genome.wellcome.ac.uk/assets/GEN10000676.jpg.)

The second step is translation, also known as protein synthesis. In this step, the resultant mRNA from transcription is translated to the resultant protein complex with the help of ribosomes. Translation occurs in the cytoplasm of the cell, outside the nuclear wall. The decoding of mRNA is initiated when the ribosome binds to the mRNA with the help of tRNAs, which transfer specific amino acids from the cytoplasm to the ribosome. The ribosome helps build the protein complex as it reads the information encoded in the mRNA.

The process of translation begins when the ribosome binds to the 5' end of the mRNA. The codons of the mRNA specify which amino acid needs to be appended to create the polypeptide chain. This process is terminated when the ribosome encounters the 3' (stop codon) of the mRNA. The resultant chain of amino acids folds to form the structure of the protein. This process is called translation, as there is no direct correspondence between the nucleotide sequence of the DNA and the resultant protein complex.

Transcription and translation is a regulated process that enables the controlled expression of genes. With evolution and differences in species, it is known that all genes are not expressed in the same way. With the exception of the housekeeping genes, genes that are always expressed in all cells (performing the basic functions) are expressed differently during different phases of development. Proteins known as transcription factors (TFs) regulate genes. These proteins bind to DNA sequences, preventing them from being transcribed and translated, and thereby switching them on or off as desired. Thus, the gene expression can be a controlled process based on the activity of transcription factors.

Transcription factors, being proteins themselves, require genes to produce them. This requirement opens a conundrum in which one gene expression affects the expression of the other genes. In this manner, genes and proteins are linked in a regulatory hierarchy. This process of turning genes on and off is called gene regulation. Gene regulation is an important part of normal development; however, a number of human diseases are the result of the absence or malfunction of transcription factors and the resultant disruption of gene expression. Considering the importance of gene regulation, a significant amount of research should be performed to understand how genes regulate each other (Figure 1.6) (Baumbach et al. 2008; Cao and Zhao 2008).

1.3 The Human Genome Project

The Human Genome Project (HGP) was initiated as a joint endeavor and sponsored by the Office of Biological and Environmental Research at the Department of Energy (DOE) and the National Human Genome Research Institute at the National Institutes of Health (NIH), with the goal of sequencing the human genome within 15 years (Collins 1998). More than 2,000 scientists from over 20 institutions in 6 countries collaborated to produce the first working draft of the human genome, a landmark in scientific research. The final phase of the HGP (1993–2003) has fulfilled its promise as the single most important project in biology and the biomedical sciences. Although the initial sequence had ~150,000 gaps, and the order and orientation of many of the smaller segments had yet to be established, the finished sequence contained 2.85 billion nucleotide base pairs (bp) and just 341 gaps (Figure 1.7).

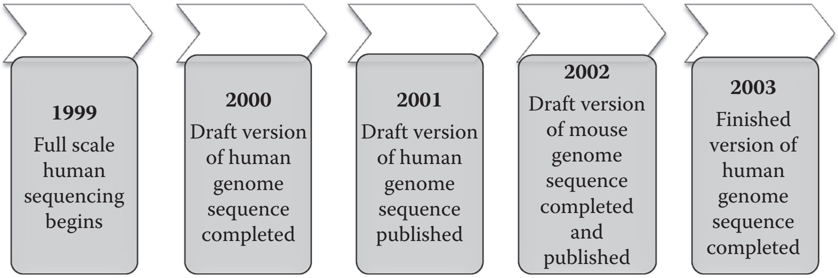

1999
Full scale human sequencing begins
2000
Draft version of human genome sequence completed
2001
Draft version of human genome sequence published
2002
Draft version of mouse genome sequence completed and published
2003
Finished version of human genome sequence completed

Figure 1.7 Key milestones achieved in the last 5 years of the HGP (1999–2003) (Constructed based on information from http://www.genome.gov/Images/press_photos/highres/38-300.jpg.)

The comprehensive human genome sequence made available through this project has increased our ability to analyze genomes, and has aided research in areas such as large-scale biology, biomedical research, biotechnology, and health care. Though researchers involved with the project have proclaimed it to be complete, certain aspects of the project have yet to be fully implemented. The methods and outcomes of this project are constantly evolving and can lead to a better understanding of gene environment interactions, structures, and functions, thereby eventually leading to the creation of accurate DNA-based medical diagnostics and therapeutics that would be important to the biomedical research community (Collins 1998).

Genetic sequence variation is necessary for the study of evolution. The HGP provides a comprehensive availability of the human genome sequence, thereby presenting unique scientific and research avenues for collaborative research. Apart from providing a means to understand numerous medically important and genetically complex human diseases, the HGP is also focused on delivering (1) genetic tests, (2) a better understanding of inherited diseases, and (3) patient-specific therapies.

Bioinformatics and computational biology are important components of making these goals a reality. The HGP (along with the other genome projects) has provided us with a description of the complete sequences of all the genes in more than a dozen organisms, and continuously provides more complete genome sequences as research continues. With technological innovations, the data generated have been growing at an exponential rate and are stored in distributed databases across the world. These databases provide challenges and opportunities for the analysis and exploitation of genes and protein sequences. In order to reap the intellectual and commercial benefits of this genetic information, researchers must be able to find the function of individual gene products. In the following section, we highlight the goals laid by the HGP and the corresponding strides made thereof in achieving the goals.

1.4 Beyond the Human Genome Project

With the completion of the sequencing of the human genome, the HGP focus switched to making the sequence publicly available to its mapping. The extraction of 3 billion base pairs was in itself a humongous task, and the analysis of this magnitude of data presented its own set of challenges and opportunities requiring a huge number of resources. Researchers from around the world realized the importance and the significant scientific contributions that could be made in the areas of human health and participated in the global endeavor to map the entire human genome (Figure 1.8).

The following sections describe the technological strides made thus far in five key areas: (1) sequencing technologies, (2) sequence variation studies, (3) functional genomics, (4) comparative genomics, and (5) functional annotation.

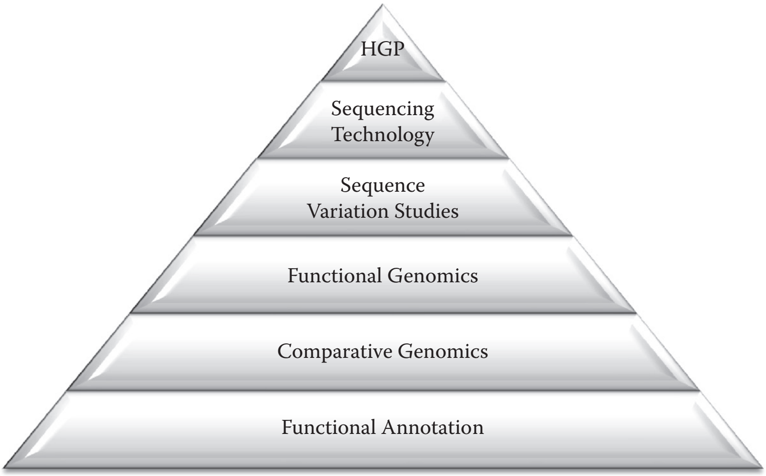


Figure 1.8 The five key areas that have been formed since the completion of the human genome project (HGP).

1.4.1 Sequencing Technology

With technological innovations, DNA sequencing technology continues to improve dramatically. Since the HGP began, the growth in data generated from sequencing projects has been exponential. This growth is caused by the emphasis given to sequencing technologies, due to:

1. Reduced costs and increased throughput of current sequencing technology
2. Support for novel technologies that can significantly improve sequencing technologies
3. Newly developed effective methods that introduce new sequencing technologies

The consequent technological innovations in the recent past have brought about a decline in the per-base cost of DNA sequencing at an exponential rate. These innovations are attributed to the improvement in the read length and accuracy of sequencing traces and have resulted in the consequent exponential growth of the genome databases (Shendure et al. 2008). The introduction of instruments capable of producing millions of DNA sequences read in a single run provides the ability to answer questions with unimaginable speed. These technologies are aimed at providing inexpensive, genome-wide sequence readouts as endpoints to applications.

There are six distinct techniques for DNA sequencing: (1) dideoxy sequencing, (2) cyclic array sequencing, (3) sequencing by hybridization, (4) microelectrophoresis,

(5) mass spectrometry, and (6) nanopore sequencing. The primary objective of these sequencing technologies is to identify the primary nucleotides, such as adenine (A), guanine (G), cytosine (C), and thymine (T), in the content of the DNA strands. The following sections provide an overview of these various sequencing strategies used.

1.4.1.1 Dideoxy Sequencing

Dideoxy sequencing was initially proposed by the Sanger Institute. The process proceeds by primer-initiated, polymerase-driven synthesis of DNA strands complementary to the template with the determined sequence. Numerous identical copies of the sequencing template undergo the primer extension reaction within a single microliter-scale volume.

Generating sufficient quantities of a template for a sequencing reaction is typically achieved by either (1) miniprep of a plasmid vector into which the fragment of interest has been cloned, or (2) polymerase chain reaction (PCR) followed by a cleanup step.

In the sequencing reaction, both the natural deoxynucleotides (dNTPs) and the chain-terminating dideoxynucleotides (ddNTPs) are present at a specific ratio. The ratio determines the relative probability of incorporation of dNTPs and ddNTPs during the primer extension. Incorporation of a ddNTP instead of a dNTP results in the termination of a given strand. Therefore, for any given template molecule, or strand, elongation will begin at the 3' end of the primer and will terminate upon the incorporation of a ddNTP. In older protocols for dideoxy sequencing, four separate primer extension reactions are carried out, each containing only one of the four possible ddNTP species (ddATP, ddGTP, ddCTP, or ddTTP), along with template, polymerase, dNTPs, and a radioactively labeled primer. The result is a collection of many terminated strands of different lengths within each reaction. As each reaction contains only one ddNTP species, fragments with only a subset of possible lengths will be generated, corresponding to the positions of that nucleotide in the template sequence. The four reactions are then electrophoresed in four lanes of a denaturing polyacrylamide gel to yield size separation with single nucleotide resolution. The pattern of bands (with each band consisting of terminated fragments of a single length) across the four lanes allows researchers to directly interpret the primary sequence of the template under analysis.

Current implementations of dideoxy sequencing differ in several key ways from the protocol described above. Only a single primer extension reaction is performed. This reaction includes all four species of ddNTP, which are labeled with fluorescent dyes that have the same excitation wavelength but different emission spectra, allowing for identification by fluorescent energy resonance transfer (FRET).

To minimize the required amount of template DNA, a cycle sequencing reaction is performed, in which multiple cycles of denaturation, primer annealing, and primer extension are performed to linearly increase the number of terminated strands.

1.4.1.2 Cyclic Array Sequencing

All of the recently released or soon-to-be-released non-Sanger commercial sequencing platforms, including systems from 454/Roche, Solexa/Illumina, Agencourt/Applied Biosystems, and Helicos BioSystems, fall under the rubric of a single paradigm, called cyclic array sequencing. Cyclic array platforms are cheap because they simultaneously decode a 2D array bearing millions (potentially billions) of distinct sequencing features. The sequencing features are "clonal," in that each resolvable unit contains only one species of DNA (as a single molecule or in multiple copies) physically immobilized on the array. The features may be arranged in an ordered fashion or randomly dispersed. Each DNA feature generally includes an unknown sequence of interest (distinct from the unknown sequence of other DNA features on the array) flanked by universal adaptor sequences. A key point in this approach is that the features are not necessarily separated into individual wells. Rather, because they are immobilized on a single surface, a single reagent volume is applied to simultaneously access and manipulate all features in parallel. The sequencing process is cyclic because in each cycle an enzymatic process is applied to interrogate the identity of a single base position for all features in parallel. The enzymatic process is coupled to either the production of light or the incorporation of a fluorescent group. At the conclusion of each cycle, data are acquired by charge-coupled device (CCD)-based imaging of the array. Subsequent cycles are aimed at interrogating different base positions within the template. After multiple cycles of enzymatic manipulation, position-specific interrogation, and array imaging, a contiguous sequence for each feature can be derived from an analysis of the full series of imaging data covering its position.

1.4.1.3 Sequencing by Hybridization

The principle of sequencing by hybridization (SBH) is that the differential hybridization of target DNA to an array of oligonucleotide probes can be used to decode the target's primary DNA sequence. The most successful implementations of this approach rely on probe sequences based on the reference of a genome sequence of a given species, such that genomic DNA derived from individuals of that species can be hybridized to reveal differences relative to the reference genome (i.e., resequencing, rather than *de novo* sequencing). The difference between SBH and other genotyping array platforms that use similar methods is that SBH attempts to query all bases, rather than only bases at which common polymorphisms have been defined. In resequencing arrays developed by Affymetrix and Perlegen, each feature consists of a 25 bp oligonucleotide of a defined sequence. For each base pair to be resequenced, there are four features on the chip that differ only at their central position (dA, dG, dC, or dT), while the flanking sequence is constant and is based on the reference genome. After hybridization of the labeled target DNA to the chip and the imaging of the array, the relative intensities at each set of four features targeting a given position can be used to infer the target DNA's identity.

1.4.1.4 Microelectrophoresis

As mentioned above, conventional dideoxy sequencing is performed with microliter-scale reagent volumes, with most instruments running 96 or 384 reactions simultaneously in separate reaction vessels. The goal of microelectrophoretic methods is to make use of microfabrication techniques developed in the semiconductor industry to enable significant miniaturization of conventional dideoxy sequencing. A key advantage of this approach is the retention of the dideoxy biochemistry, which has proven robustness for >1,011 bases of sequencing. Until alternative methods achieve significantly longer read lengths than they can today, there will continue to be an important role for Sanger sequencing. Microelectrophoretic methods may prove critical to continue to reduce costs for this well-proven chemical process. There may also be a key role for lab-on-a-chip integrated sequencing devices that will provide cost-effective, clinical point-of-care molecular diagnostics.

1.4.1.5 Mass Spectrometry

Mass spectrometry (MS) has established itself as the key data acquisition platform for the emerging field of proteomics. There are also applications for MS in genomics, including methods for genotyping, quantitative DNA analysis, gene expression analysis, analysis of indels and DNA methylation, and DNA/RNA sequencing.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) is an MS sequencing technique that relies on the precise measurement of the masses of DNA fragments present within a mixture of nucleic acids. *De novo* sequencing using MALDI-TOF-MS read lengths are limited to <100 bp. Applications of MS sequencing include:

1. Deciphering sequences that appear as compression zones by gel electrophoresis
2. Direct sequencing of RNA (including for identification of posttranslational modifications of ribosomal RNA)
3. Robust discovery of heterozygous frameshift and substitution mutations within PCR products in resequencing projects
4. DNA methylation analysis

1.4.1.6 Nanopore Sequencing

Nanopore sequencing is an approach for single-molecule sequencing that involves passing single-stranded DNA through a nanopore. The nanopore is a biological membrane protein or a synthetic solid-state device. As individual nucleotides are expected to obstruct the pore to varying degrees in a base-specific manner, the resulting fluctuations in electrical conductance through the pore can, in principle, be measured and used to infer the primary DNA sequence. Published examples of the nanopore-based characterization of single nucleic acid molecules include:

1. The measurement of duplex stem length, base pair mismatches, and loop length within DNA hairpins (Vercoutere et al. 2001)
2. The classification of the terminal base pair of a DNA hairpin, with approximately 60 to 90% accuracy with a single observation, and >99% accuracy with 15 observations of the same species (Winters-Hilt et al. 2003)
3. Reasonably accurate (93 to 98%) discrimination of deoxynucleotide monophosphates from one another with an engineered protein nanopore sensor (Astier et al. 2006)

Significant pore engineering and technology development may be necessary to accurately decode a complex mixture of DNA polymers with single-base pair resolution and useful read lengths. Provided these challenges can be met, nanopore sequencing has the potential to enable rapid and cost-effective sequencing of populations of DNA molecules with comparatively simple sample preparation.

1.4.2 Next-Generation Sequencing

With the advancements made in sequencing technologies, there has also been recent advancement in the form of a new generation of sequencing instruments. These instruments cost less than the techniques described in the previous section and promise faster sequence readings, as they require only a few iterations to complete an experiment. These faster reads foster the potential to add to the exponential increase of sequence data. The expected increase of data is also attributed to the next-generation sequence technology's ability to process millions of reads in parallel, rather than the traditional 96 reads. Thus, with the introduction of next-generation sequencing technology, large-scale production gene sequence data may require specialized use of robotics and high-tech instruments, computer databases for storage of the huge data, and bioinformatics software for analysis.

An added advantage of the proposed next-generation sequence reads is that they are generated from fragment libraries that have not been subjected to conventional vector-based cloning and *Escherichia coli*-based amplification stages used in capillary sequencing rendering the sequences of any prevalent biases caused by cloning.

Three commercially used and commonly cited next-generation sequencing platforms include the Roche (454) GS FLX Sequencer, the Illumina Genome Analyzer, and the Applied Biosystems SOLiD Sequencer (refer to Table 1.2 for a detailed comparison). The generic work flow for creating a next-generation sequence library is simple. Fragments of DNA are prepared for sequencing by ligating specific adaptor oligos to both ends of each DNA fragment. Typically, only a few micrograms of DNA are needed to produce a library. Each of these platforms applies a unique or

Table 1.2 Comparison of Metrics and Performance of Next-Generation DNA Sequencers

	Platform		
	Roche (454)	*Illumina*	*AB SOLiD*
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing by synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Mb/run	100 Mb	1,300 Mb	3,000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32–40 bp	35 bp

Source: Mardis, E.R., *Trends Genet* 24, no. 3 (2008): 133–141.

modified approach to sequence the paired ends of a fragment, the scope of which is not covered in this book. For details refer to Mardis (2008).

Since next-generation sequencing technology is relatively new, there is little insight on the accuracy of the reads, and the quality of the results obtained have yet to be understood. When compared to the more traditional capillary sequencers, next-generation sequencers produce shorter reads, ranging from 35 to 250 base pairs (bp), than the traditional 650 to 800 bp created by other methods. The length of the reads could impact the utilization of the generated data. Efforts are being pursued currently to benchmark the reads with the traditional capillary electrophoresis.

Although next-generation sequence technology provides many advantages over traditional methods, it also poses several computational challenges. Many storage and data management systems cannot handle the amount of data generated. The data storage must be scalable, dense, and inexpensive to handle the exponential growth. Various centers of bioinformatics around the globe are investing heavily in high-performance disk systems and data pipelines to overcome the challenge of handling the large number of files that are expected to be accessed when the demand arises.

Software pipelines are also required to provide the necessary analysis and visualization of the data generated. More importantly, software has to be in place to provide annotations of the sequences generated.

1.4.2.1 Challenges of Handling NGS Data

The challenges of handling the deluge of NGS data stem from two key concepts that are used to analyze the sequence reads. These concepts focus on *de novo* assembly and alignment. The following sections describe the computational algorithms used to handle the massive amounts of Illumina sequencing data for both *de novo* assembly and alignment of reads (Paszkiewicz and Studholme 2010).

1.4.2.1.1 *De Novo* Assembly

De novo sequence assembly is the process whereby we merge individual sequence reads to form long contigs (continuous sequences) that share the same nucleotide sequence as the original template DNA from which the sequence reads were derived.

Two algorithms are prominently used to assemble sequence reads: (1) algorithms based on the overlap-layout-consensus (OLC) approach (Huang and Madan 1999) and (2) algorithms based on a de Bruijn graph (Simpson et al. 2009). These have been well-reviewed techniques and have been implemented in effective genome-assembly software packages. However, these genome sequence assembly programs are not well suited to short sequence reads generated by Illumina and AB SOLiD platforms (Paszkiewicz and Studholme 2010).

1.4.2.1.2 Alignment

Once the assembly is performed, the contigs are subject to alignment algorithms (Li and Homer 2010), which focus on the creation of auxiliary data structures called indices for the sequence reads and the reference sequence. We can categorize these structures into three algorithms: (1) hash table-based algorithms, (2) suffix tree-based algorithms and (3) algorithms based on merge sorting.

1.4.2.1.2.1 Hash Table-Based Algorithms — These algorithms create a hash table index that can be used to trace back to specific basic local alignment search tool (BLAST) matches as they rely on a seed-and-extend paradigm. In the first phase of the algorithm, BLAST maintains the position of each *k*-mer subsequence of the query in a hash table with the *k*-mer sequence being the key, and scans the database sequences for *k*-mer exact matches called seeds. Once this phase is complete, BLAST extends and joins the seeds without gaps. Further refinements are carried out using Smith-Waterman alignment to refine the seeds, which achieves statistically significant results. The tools that are prominently using the hash table-based algorithms are MAQ, the SOAP family of alignment tools, viz., SOAP, SOAP2, and SOAP3/GPU, and Abyss (Simpson et al. 2009).

1.4.2.1.2.2 Suffix-Based Trees — With the short sequence reads it is a challenge to obtain the exact matches of the reads using BLAST. Thus researchers tend to favor inexact matches of sequence for alignments. The suffix-based approaches aim to essentially reduce the inexact matching problem to the exact matching problem using two steps: (1) identifying exact matches and (2) building inexact alignments supported by exact matches. To find exact matches, these algorithms use a certain representation of suffix trees. The advantage of using suffix trees is that alignment to multiple identical copies of a substring in the reference is only needed once because these identical copies collapse on a single path in the tree, whereas with a typical hash table index, an alignment must be performed for each copy. The tools that prominently use the suffix-based trees for alignment of sequences are MUMmer and REPuter (Paszkiewicz and Studholme 2010).

1.4.3 Sequence Variation Studies

Nature retains diversity in a population of organisms living in varied environmental conditions. This diversity is the result of genetic variations: traits that vary and are coded in the genes of the population. Since the inception of the HGP, several studies have been conducted to understand the effect of genetic variations between individuals.

Natural sequence variation is the fundamental property of all genomes. It is believed that any two haploids exhibit multiple kinds of genetic variations and polymorphisms (see Figure 1.9). There are three basic forms of genetic variations: mutations, gene flow, and sex. Not all of these genetic variations have functional implications. Sequence polymorphisms also include duplications, rearrangement, insertions, and deletions. The most common polymorphism in the human genome

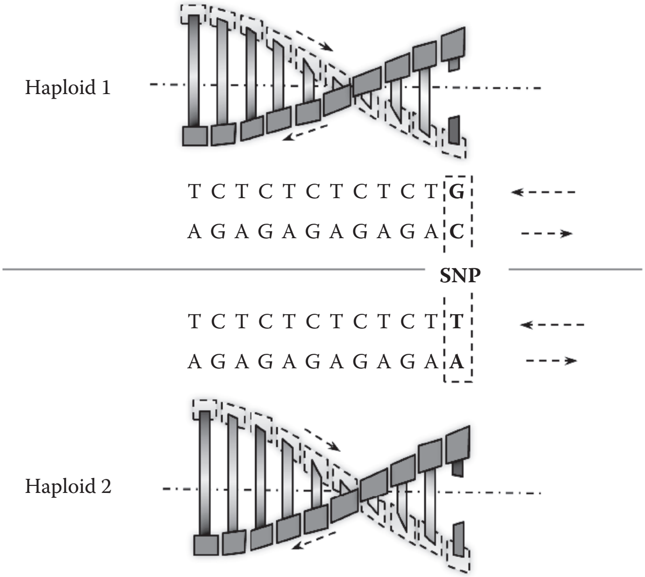


Figure 1.9 A schematic representation of a single nucleotide polymorphism between two haploids.

is the single-base pair difference, better known as a single nucleotide polymorphism (SNP). When two haploid human genomes are compared, it is observed that SNPs occur at every kilobase of the gene sequence. SNPs are abundant, stable, and widely distributed across the genome. Because of these properties, SNPs can be used for the mapping of complex traits such as cancer, diabetes, and mental illness. However, the occurrences of these variations across the entire genome are rare, making it a challenge to challenge to identify and understand these variations (Figure 1.9).

Keeping this challenge in mind, the objective of sequence variation studies is to provide dense maps of SNPs that will make genome-wide association studies possible. These maps are powerful means for identifying genes that contribute to disease risk. They will also permit the prediction of individual differences in drug responses. When the maps are made available to the public, maps of a large number of SNPs distributed across the entire genome come together with technology for rapid, large-scale identification. The scoring of SNPs must be developed to facilitate this research. The HGP envisioned the following goals concerning genetic variation analysis. First, the goal is to develop technologies for rapid, large-scale identification or scoring, or both, of SNPs and other DNA sequence variants. In order to achieve this goal, the following objectives had to be met:

1. The creation of an SNP map of at least 100,000 markers
2. The development of concepts and methods to study multigene traits and map DNA sequence variations to phenotypic variations such as complex disease
3. The creation of public resources containing DNA samples and cell lines to enable SNP discovery using the public resources

To this end, large bodies of works have been conducted through primary data sources that contain SNP data, including the dbSNP (current build 134) containing approximately 6,961,883 human reference SNP clusters, the Human Gene Mutation Database (HGMD) containing 113,247 entries (professional release 2011.2), and the disease-specific Online Mendelian Inheritance in Man (OMIM) (September 2011) that contains approximately 2,648 genes with disease-causing mutations. Several tools are available for the analysis of SNPs, of which SNPper is prominently used. Furthermore, BioPerl provides an API for the analysis of SNPs and Genewindow provides visualization technology. Other online resources that enable effective visualization of SNP data include the UCSC Genome Browser (see Figure 1.10) and the Ensembl Genome Browser (Table 1.3).

1.4.3.1 Kinds of Genomic Variations

HGP focuses on the creation of a repository of all known SNPs derived from a diverse population across the United States and the creation of appropriate tools to

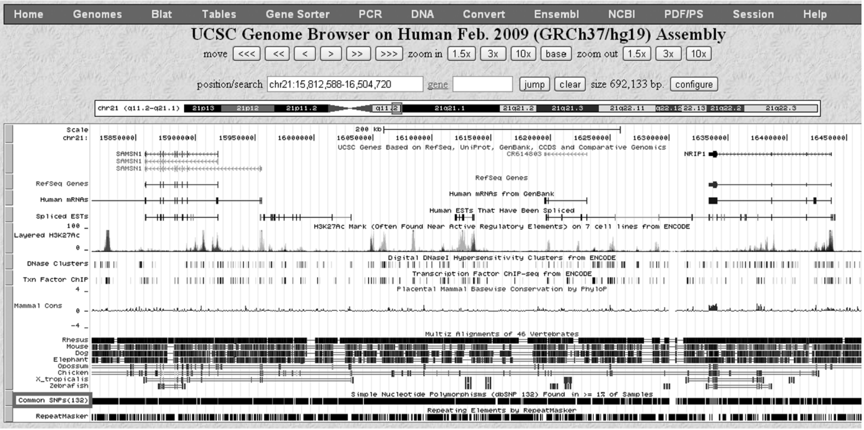

Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
Ensembl
NCBI
PDF/PS
Session
Help
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr21:15,812,588-16,504,720 gene jump clear size 692,133 bp. configure
chr21 (q11.2-q21.1)
21p13
21p12
21p11.2
q11.2
21q21.1
21q21.2
21q21.3
21q22.11
21q22.2
21q22.3
Scale
chr21:
200 kb
15850000
15900000
15950000
16000000
16050000
16100000
16150000
16200000
16250000
16300000
16350000
16400000
16450000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
SAMSN1
CR614803
NRIP1
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
0
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0
-4
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Common SNPs(132)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 1.10 A screenshot of the UCSC Genome Browser, a tool to visualize SNP data.

analyze SNPs. The HGP suggests that approximately 95% of the discovered SNPs belong to the noncoding regions of the genome. Furthermore, it is still an open challenge to determine the functional aspect of the SNPs found near or in genes. However, it is still believed that based on their location on the genome, SNPs can potentially alter the functions of DNA, RNA, and proteins alike. A general categorization of SNPs based on their location is shown in Table 1.4 (Mooney 2005; Rebbeck et al. 2004).

Generally, nonsynonymous SNPs (nsSNPs) cause a change in the amino acid sequence of the resultant protein sequence, either by substituting amino acids or introducing a nonsense/truncation mutation (Ng and Henikoff, 2006). Table 1.4 shows variants that affect the expression of a gene translation by interrupting a regulatory region known as a regulatory SNP. Similarly, those variants that interfere with normal splicing and mRNA functions are categorized as intronic SNPs or synonymous SNPs. Due to increasing research efforts, the molecular effects of variations are becoming better understood, which allows us to shed more light on genetic diseases.

1.4.3.2 SNP Characterization

To understand the patterns of sequence variations in coding regions of genes, bioinformatics strategies have been focused on analyzing disease-associated mutations that focus precisely on where diseased alleles occur with respect to their corresponding protein structures. It is important to understand the underpinnings of these mutations and what properties guide such mutations.

Table 1.3 SNP Resources Widely Used

Description	
Genome Resources	
dbSNP	The primary repository for SNP data
Ensembl	Genome database
GoldenPath	Genome database
HapMap Consortium	Haplotype block information
JSNP	Japanese SNP database
Mutation Repositories	
HGVBase	Public genotype-phenotype database
HGMD	Mutation database with many annotations
Swiss-Prot	Protein database with extensive variant annotations
List of Locus-Specific Databases	
CGAP-GAI	Cancer Gene Anatomy Project at the National Cancer Institute
Other databases and tools	Tools for SNP analysis and gene characterization
Tools	
SNPper	Novel software for SNP analysis
BioPerl	A programming application program interface (API) for bioinformatics analysis
Genewindow	Interactive tool for visualization of variants

Table 1.4 Existing SNP Categorization

Coding SNPs	cSNP	Positions that fall within the coding regions of genes
Regulatory SNPs	rSNP	Positions that fall in regulatory regions of genes
Synonymous SNPs	sSNP	Positions in exons that do not change the codon to substitute an amino acid
Nonsynonymous SNPs	nsSNP	Positions that incur an amino acid substitution
Intronic SNPs	iSNP	Positions that fall within introns

It is hypothesized that mutations on the gene sequence (position specific) are conserved through evolution and are reflected to the protein structure (Ng and Henikoff 2002; Krishnan and Westhead 2003).

One of the tasks of SNP analysis is to gauge the impact of each nsSNP on protein function. Due to the size of the SNP data, this task is experimentally infeasible. Thus, researchers have looked into computational methods to predict changes in protein function if an amino acid changes. This technique, also known as amino acid substitution (AAS), focuses on disease-causing mutations that are likely to occur at positions that are conserved through evolution. It is further believed that disease-causing AASs affect the structural characteristics of the resulting protein, suggesting that protein structural information can be used to analyze these mutations (Table 1.5).

1.4.4 Functional Genomics

With the entire human genome sequence publicly available, a new approach to address biological challenges has taken form. This approach, called functional genomics, entails the functional understanding of the human DNA on a genome scale. Functional genomics is viewed as an intermediate step that brings biological research to being applied in medicine (from bench-side to bedside). Based on successes of previous studies of sequences within organisms, it is inferred that the function of genes and other functional elements of the genome can be inferred more accurately only when the genome is studied in its entirety.

Table 1.5 Strategies That Have Been Used for Analysis of AAS

Method	*Algorithm*
SIFT (Ng and Henikoff 2002)	Sequence homology and position-specific scoring matrices
PolyPhen (Stitziel et al. 2004)	Sequence conservation, structural information modeling
SNPs3D (Yue and Moult 2005)	Structure-based support vector machines (SVMs) and sequence conservation-based SVMs
PANTHER PSEC (Thomas et al. 2003)	Sequence homology and scores obtained from PANTHER hidden Markov models of protein families
TopoSNP (Stitziel et al. 2004)	Characterization of residues based on topological information such as buried, on-surface, or pocket information

At the end of the HGP, knowledge about a gene's structure and other elements was only the tip of the iceberg. Further insights about the function of a gene can be derived from its interaction with the environment.

Existing methods for analyzing DNA function at a genomic scale include the comparison and analysis of sequence patterns, large-scale analysis of mRNA, various approaches of gene distribution, and the analysis of protein complexes (for gene products). Despite these methods, there is still a need for novel strategies to elucidate the function of genes. Thus, functional genomics focuses on the development of technology that can be used for the large-scale analysis of the human genome in its entirety rather than in parts. In functional genomics, emphasis is given to gene transcripts and their protein products, including the identification and sequencing of full-length cDNAs that represent the entire human genome. Thus, the following were the objectives of functional genomics:

1. Extend support for the creation of global approaches, improved technologies, and the creation of relevant libraries for the comparative and computational analysis of noncoding sequences: It is imperative to understand these sequences, as they are noncoding and carry out other functions, such as RNA splicing, sequences that are responsible for the formation of chromatin domains, sequences that maintain chromosome structure, sequences that are responsible for recombination and replication, and sequences that specify numerous functional untranslated RNAs.
2. Enable and support the creation of technology for the comprehensive analysis of gene expression so that it is possible to analyze spatial and temporal patterns of gene expression in both human and model organisms, thereby providing a means to understand the expression of genes: To make this analysis possible, cost-effective and efficient technology that measures the parameters of gene expression in a reliable manner and can be easily reproduced must be developed. In addition to the required technological innovations, complementary DNA (cDNA) sequences and validated sets of clones with unique identifiers are also needed to analyze gene expression data. Other required developments include novel methods to quantify, represent, analyze, and archive the resulting gene expression data.
3. Investigate alternate means of studying functions, like methods for genome-wide mutagenesis: This step includes the creation of mutations that cause loss or alteration in gene functions. Associated technologies for large-scale *in vivo* and *in vitro* are also required to generate and find mutations in each gene and phenotype.
4. Understand protein functions on a genome-wide scale to develop technology for global protein analysis to provide a comprehensive understanding of genome functions: The development of computational and experimental models to analyze both spatial and temporal patterns of protein expression, protein-ligand interactions, and protein modification is required.

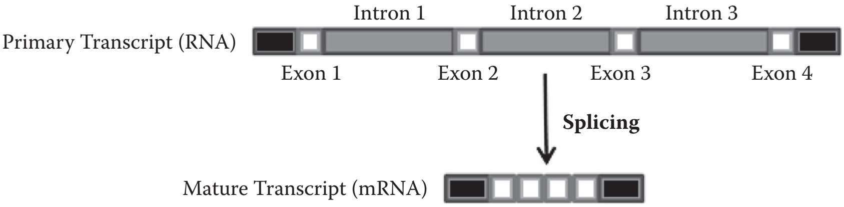


Figure 1.11 The process of splicing, in which the introns are removed from the primary transcript (RNA) and the exons are combined to form the mature transcript (mRNA).

1.4.4.1 Splicing and Alternative Splicing

Splicing, the first step to understanding the functions of genes and the roles they play in an organism, is the alteration of the primary transcript RNA after transcription. In this process, introns are removed, and the remaining exons are joined (see Figure 1.11). It is necessary for the mature transcript (of the mRNA) to be subject to splicing, as it enables the production of the correct protein during translation. However, it is commonly observed that a set of unique proteins can be created by varying the exon composition of the mRNA through the process of splicing. This process is referred to as alternative splicing. Alternative splicing can occur in many ways using different combinations of exon units. Moreover, exons can be skipped, or introns can be retained, creating a complex system requiring the need for computational modeling and interpretation.

The sequencing of the human genome has raised the importance of alternative splicing as an RNA regulatory mechanism. Furthermore, alternative splicing has provided a means for researchers to explain why there is such a large repertoire of proteins. It has also potentially helped identify and explain defects that occur in the splicing mechanism and that result in complex diseases such as cancer.

Bioinformatics has played a key role in cataloguing splice variations in humans and other eukaryotic genomes (Modrek et al. 2001). Tools and algorithms have also been developed to characterize splice regulatory elements that control the expressions of genes (Florea 2006). Instead of focusing on an organism’s total number of genes to explain its functional and behavioral complexity, researchers are now interested in determining how each gene can be “reused” to create multiple functions and new modes of regulation. To this end, studies on both human and mouse sequence data have resulted in algorithms that have clustered genes and samples based on their alternative splicing patterns, indicating the importance of alternative splicing to differentiate between genes (Lee and Wang 2005).

1.4.4.1.1 Types of Alternative Splices

Alternative splicing of pre-mRNA is an important regulatory mechanism to modulate genes and their corresponding protein complexes within a cell. It is believed

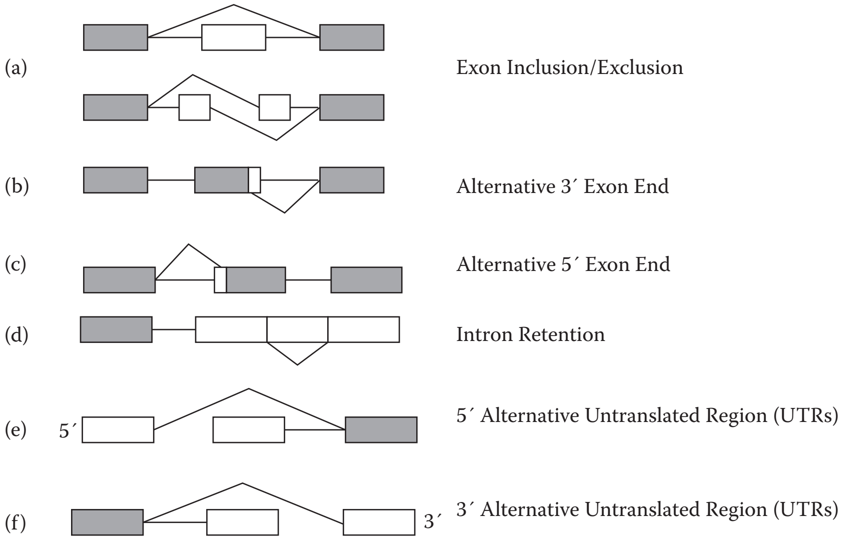


Figure 1.12 Schematic representation of the types of alternative splicing events. Alternatively, spliced elements (exons or portions of exons) are shown in red, and those constitutively spliced are shown in blue. The exons are represented as boxes, and the introns by straight lines connecting the exons. (From Florea, L., *Briefings Bioinformatics* 7, no. 1 (2006): 55–69. With permission.)

that the proteins obtained from alternative splices can be used to regulate a gene expression within a cell. It is therefore necessary to understand and catalog all possible combination of exons obtained from a gene.

With the perspective of gene structure, alternative splicing is categorized into four types of events (see Figure 1.12). It should be noted that due to the data's intrinsic property of being noisy, the identification of gene boundaries is difficult. Therefore, it is an open challenge to identify and characterize the 5' and 3' alternative untranslated regions (UTRs), as shown in Figure 1.12e–f.

1.4.4.1.2 Alternative Splicing for Gene Annotation

The role of bioinformatics in alternate splicing is prevalent in areas of gene annotation and splice regulation (Lee and Wang 2005). Traditional gene discovery, better known as gene prediction (Birney et al. 2004), has been performed through a combination of *ab initio* and comparative methods for the identification of linear exon-intron models of genes. With the completion of the HGP and the resultant large-scale annotation projects such as the Ensembl (Hubbard et al. 2002) and UCSC Genome Browser database (Fujita et al. 2010) with different data, dependent models came into existence. These models are based on different prediction methods that create the "evidence" of the existence of a gene and use a combiner algorithm to associate the collected evidences into a unified representative model of a gene. With the inclusion of alternative spliced transcripts or alternative splicing events as part of the annotation process through manual curation, these databases improve the quality of their datasets.

There are four prominent approaches used in gene prediction:

1. ***Ab initio* programs:** These programs do not require any prior or additional information to predict a gene for a given DNA sequence. They rely on the hidden Markov model (HMM) framework to provide the parameterization and decoding of a probabilistic model of gene structure (Zhang 2002).
2. **Evidence-based techniques:** There are two classes of evidence-based techniques for gene prediction. The first class uses the well-known pairwise HMM methods. The second class uses external evidence to score potential exons (Parra et al. 2003; Birney et al. 1996; Alexandersson et al. 2003).
3. **Informant approach:** This technique predicts a gene based on information of exons derived from two or more sample genomes (Pedersen and Hein 2003).
4. **Feature-based approaches:** These approaches do not rely on a probabilistic model or prior knowledge from the underpinning DNA. However, the framework facilitates the integration of multiple component features derived from the DNA sequence (Howe et al. 2002).

1.4.4.1.3 Regulation of Alternative Splicing

To regulate splicing, it is important to identify what causes or controls the variation in splicing. The control of alternative splicing affects the abundance, structure, and function of transcripts and encoded proteins from a gene through the modification of their properties, such as its binding affinity, intracellular localization, stability, and enzymatic activity (Stamm et al. 2005). Furthermore, exon selection in alternative splicing is tissue specific, and is determined based on the developmental stage, or disease specific (Florea 2006). Thus, the regulation of alternative splicing is more specific and case driven than transcriptional regulation.

Though little is known about splicing regulation through regulatory proteins, there is an alternative form of regulation that focuses on splice regulation that is not part of the basal spliceosome function. The basal spliceosome function is regulated by families of splicing regulatory proteins. These proteins bind to the RNA in the surrounding regions of exons, thereby catalyzing the exon's inclusion or exclusion by activating or inhibiting the function of the splice site. Little is known about the characteristics of regulatory proteins and the corresponding RNA binding sites, and these issues are being actively investigated.

1.4.4.1.4 Splice Variants

Gene annotation using alternative splicing and the regulation of alternative splicing form the crux of research that relies on computational methods. The resulting bioinformatics techniques focus mainly on cataloging the various splice variants.

Despite the tendency of genomes to remain the same for different tissues or cell types in an organism, their transcriptomes (set of all RNAs of the tissues/cell) can be significantly different.

The motivation for using splice annotation is to identify and catalog all mRNA transcripts of a cell at different stages, using both spatial and temporal expression, along with functional information of the splices. This objective is difficult, if not impossible to achieve, considering the incomplete and fragmented nature of the data along with insufficient experimental characterization.

Several computational approaches have been suggested to overcome these limitations and identify splice variants. These techniques rely on characterizing splicing patterns obtained from partial cDNA or protein sequences, or exon-level alternative splicing events to analyze and characterize transcriptomes. The varying splice patterns can be applied in the design of diagnostic markers that can be validated using either *in vitro* microarray and proteomic experiments or *in silico* via the identification and annotation of splice forms.

Bioinformatics techniques used for the annotation of full-length alternative spliced transcripts include:

1. **Gene indices:** Gene indices refer to gene- or transcript-oriented collections of express sequence tags (ESTs) and micro-RNA (mRNA) sequences grouped by sequence similarity (Lee et al. 2005; Liang et al. 2000). This method employs a pairwise sequence similarity for comparison between two sequences. Here, all the EST and mRNA sequences are subject to a one-to-one comparison to identify overlaps between each other. These sequences are then grouped and assembled into disjointed clusters (a consensus sequence) based on a threshold of overlaps.

   The creation of a gene index is complex and suffers from two drawbacks: (a) overclustering, in which different paralogs (similar sequences belonging to different genes) are put into the same cluster creating a false overlap, and (b) underclustering, in which several clusters are produced for a single gene.
2. **Genome-based methods for clustering spliced alignments:** In this approach, unlike gene indices, the spliced alignments of cDNA or protein sequences are clustered at a point of reference along the reference genome sequence (loci) (Florea et al. 2005). Splice graphs are one such technique that is prominently used in alternative splicing annotation for capturing splice variants in a gene (Kim 2005). Using the concept of directed acyclic graphs, with each node representing an exon and edge that connects two exons representing an intron, a splice variant corresponds to the paths obtained

through the graph traversal from a predetermined source vertex (vertex with no incoming edges) to a sink vertex (vertex with no outgoing edges). The advantage of this technique is that it results in all possible combinations of exon-intron combinations. However, not all of the combinations are biologically significant. Several filtering strategies prioritize the combinations and rank splice variants that are biologically significantly higher.

1.4.4.2 Microarray-Based Functional Genomics

Microarray technology has been an important contribution to functional genomics as it provides a means to analyze the expressions of hundreds of thousands of genes that belong to an organism for a specific reaction at a given instance, simultaneously. This technology has facilitated an understanding of the fundamental aspects of growth and development. Moreover, it has aided in the exploration of the genetic causes of complex genetic diseases such as cancer. Typically, microarray data are classified into three categories, based on the types of the samples used to construct the microarrays (see Table 1.6, Figure 1.13).

Gene regulatory network analysis (Huang et al. 2007) is an analytic technique that is used to extract gene regulatory features (i.e., activation and inhibition) from gene expression patterns. Changes of gene expression levels across samples provide information that allows reverse engineering techniques to construct the network of regulatory relations among those genes (Lockwood et al., 2006).

For instance, the expression of a gene is regulated by a transcriptional control mediated by a complex cis-regulatory system. Transcriptional factors activate or repress gene expression by binding to their respective binding sites: comparatively short sequences (several hundred to several thousand base pairs, depending on the species) upstream, downstream, or far away from the transcriptional start sites. Specific sites within such regions, which are generally composed of dense clusters, are recognized by the regulatory proteins (transcription factors (TFs)) that control the rate of gene transcription.

Table 1.6 The Categorization of Microarrays and Their Associated Applications

Microarray Type	*Application*
CGH	Tumor classification, risk assessment, and prognosis prediction
Expression analysis	Drug development, drug response, and therapy development
Mutation/polymorphism analysis	Drug development, therapy development, and tracking disease prognosis

Source: NCBI, *NCBI: A Science Primer*, July 27, 2007, http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html#ref1 (accessed September 13, 2011).

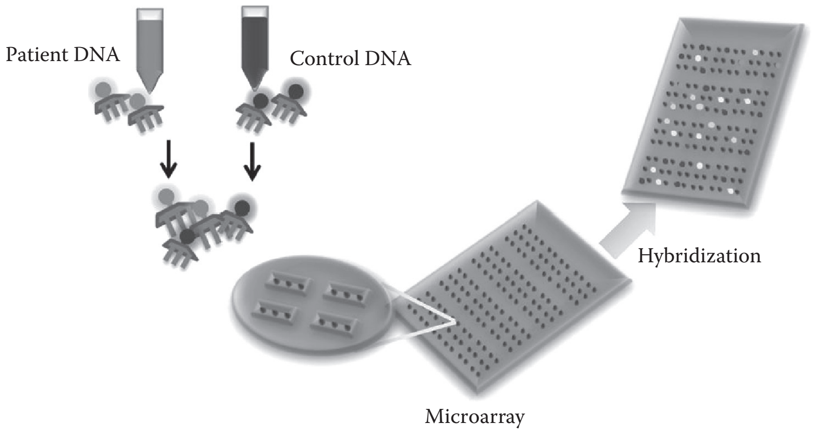


Figure 1.13 Schematic representation of the microarray-based comparative genomic hybridization (CGH) process.

1.4.4.2.1 Types of Regulatory Regions

Regulatory regions of higher eukaryotes can be subdivided into proximal regulatory units—promoters—which are located close to the 5’ end of the gene, and distal transcription regulatory units called enhancers or cis-regulatory modules (CRMs). CRMs may be located far upstream or downstream of the target gene, and are much more difficult to recognize because they lack proximal specific transcriptional signals, such as position relative to coding sequence, the TATA box, the CAAT box, the transcription start site consensus, etc. Therefore, recognition of CRMs is even more difficult than recognition of promoters (Abnizova and Gilks 2006).

1.4.4.2.2 Experimental Determination of Regulatory Region Function

Biochemical characteristics can identify binding sites precisely and are the only way to determine whether consensus sequences differ among species. There are several methods available for producing DNA-protein interaction data. Nitrocellulose binding assays, electrophoretic mobility shift assay (EMSA), enzyme-linked immunosorbent assay (ELISA), DNase footprinting assays, DNA-protein cross-linking (DPC), and reported conducts are examples of *in vitro* techniques that are used to determine DNA binding sites and analyze the difference in binding specificity for different protein-DNA complexes. The major disadvantage of these methods is that they are not suited to high-throughput experiments.

A microarray-based assay called chromatin immunoprecipitation (ChiP) was developed for genome-wide determination of protein binding sites on DNA.

Other types of experiments are systemic evolution of ligands by exponential enrichment (SELEX) and phage display (PD), which offer a high-throughput possibility to select high-affinity binders, DNA and protein targets, respectively. Both SELEX and PD suffer the same disadvantage: most sequences obtained from these experiments are good binders, but it is hard to say anything about their relative affinities. It is assumed that the best binders occur more frequently.

In dsDNA microarrays are presented for exploring sequence-specific protein-DNA binding. The major advantage over the aforementioned methods is that it is a high-throughput method resulting in data with associated relative binding affinities.

Finally, x-ray crystallographic and NMR spectroscopic data provide a base for studying the structural details of protein-DNA interactions. Protein-DNA complexes have successfully been co-crystallized, and the data have been deposited into the protein data bank and nucleic acid database (NDB). However, these experiments are time-consuming.

Unfortunately, for technical reasons, the numbers experimentally verified, binding sites are nearly always underestimated, and the physical length of regulatory regions is rarely well defined.

1.4.5 Comparative Genomics

Due to evolution, all organisms are believed to be related. Thus, the study of one species could lead to valuable information about another species. Molecular genetics enables researchers to understand the genes of one species based on the genetic makeup of related genes in other species. To this end, several experiments provide insights into the universality of biological mechanisms, through comparisons between genomes. Thus, valuable insights relating to the gene structure and function of closely related species are brought to the forefront using comparative genomics.

The comparative analysis of the human genome with a variety of modeled organisms is advantageous and is an important field of research. The underpinning rationale that governs cross-species sequence comparative genomics, as stated above and in Pennacchio and Rubin (2003), is based on the observation that sequences and functions are conserved across evolutionary distant species. This conservation enables researchers to identify and distinguish between functional and nonfunctional genetic sequences in both gene sequence data and protein sequence data. This rationale lays the impetus for gene expression, regulation, and control experiments.

It has also been shown that the inverse also holds true in orthologous genomic sequences from different vertebrates. Thus, the comparative analysis of evolutionary conserved sequences is a viable strategy to identify biologically active regions over the human genome.

Various genomic visualization, annotation tools, and databases are available to the biomedical research community and are publicly available. These tools have been successfully used to identify biologically important genes and sequences involved in gene regulation.

1.4.6 Functional Annotation

Supporting the above genomic research is one of the keystones of the HGP. This support includes the effective recoding, distribution, and analysis of all results and discoveries. Bioinformatics and computational biology are core components that are targeted toward satisfying this goal. Thus, the services that bioinformatics offers can be categorized into two areas: (1) databases and (2) analytical tools.

This section is devoted to the effective collection, analysis, annotation, and storage of sequence data that is exponentially growing. For effective use of the data generated in the public domain, it is important to provide effective mapping of all gene sequence data to expression data and protein sequence data. User-friendly interfaces and user-friendly databases are imperative to the success of the genome project. Additionally, a range of computational algorithms that allow researchers to extract, view, and annotate gene and protein sequences effectively will benefit the research community. Such algorithms address the following objectives:

1. Improve the content and utility of databases.
2. Develop better tools for data generation, capturing, and annotation.
3. Develop and improve tools and databases for comprehensive functional studies.
4. Develop and improve tools for representing and analyzing sequence similarity and variation.
5. Create mechanisms to support effective approaches for explorative and robust software that can be widely used in different applications.

The successes of these objectives have been documented primarily in the creation and maintenance of large databases such as the PDB, Ensembl, and SwissProt. However, bioinformatics and computational biology has been actively pursued as an area of research for the creation of better analysis techniques, algorithms, and tools in fields like gene sequence analysis, microarray analysis, protein sequence and structural analysis, and functional annotation.

1.4.6.1 Function Prediction Aspects

One of the problems arising from the completion of the HGP was the functional annotation of generated sequences. Biologists were then and are now faced with the challenge of analyzing the functional significance of genes with traditional statistical techniques. Not only is the volume of sequence and structure data growing, but the diversity of the sources that generate the data also poses significant challenges that require computational expertise and has led to a disproportionate growth in the number of uncharacterized gene sequences.

The established and traditionally used method for gene and protein annotation is based on homology modeling in which new sequences are assigned functions based on the similarity they share with sequences of known annotations. However, homology modeling amplifies existing erroneous annotations. Because of this problem, the efficacy of this method is questionable considering the constant growth of sequence information. Thus, there is a need for standardized, large-scale sequence annotation tools that use machine learning and are free of manual curation. This automated function prediction of sequences could be incorporated into larger work flows. This section explains some computational protein function prediction techniques (Friedberg 2006).

The definition of biological function is ambiguous, and the exact meaning of the term varies based on the context in which it is used. Further, there is a multiperspective view of protein function that is categorized into three classes:

1. **The biochemical aspect:** In this class, the protein function is derived from the specific substrate information. This definition requires only a disembodied protein performing alone *in vitro*.
2. **The physicological aspect:** In this class, function is defined in respect to the function of a protein within an organism from the subcellular level to the whole organism. Here, sequences could derive functional information from the signal pathways that the protein is a part of or from their interacting partners.
3. **The phenotypic (medical) aspect:** In this class, the functional information is derived from the mutations that occur in the sequence of the protein.

Keeping these aspects in mind, there are several methods proposed in the automated functional annotation of sequences, and the following section enumerates them.

1.4.6.1.1 Computational Functional Annotation

The basic challenge faced in the computational annotation of sequences is determining what constitutes functional information and how that function should be described in a computationally interpretable manner. Two forms of information can be adopted to define protein function. From a data mining perspective, these forms of information include protein sequence information and protein structure information that can be included as features of interest in the algorithm.

Protein sequences are represented as character strings that are used in an array of tasks: pairwise and multiple sequence alignments and motifs, all of which can easily be included as features for analysis using computational algorithms. Protein structural information, on the other hand, is more complex. Here, the PDB files (.pdb) have vast amounts of information, in the form of 3D coordinates, which can be exploited to find similarities between two proteins.

Apart from features of interest, there is also a need for controlled vocabulary, or keywords that can be used to annotate functionally significant regions of a protein, and well-defined relationships in describing functions. The Enzyme Commission Classification (EC) (Webb 1992) is one such annotation system that classifies reactions based on a four-level hierarchy (represented using a four-position identifier) that moves from a general to a specific categorization.

For example, the hierarchy starts with a generalized lyase (4.-.-.-), in the first position and moves through a more specific nitrogen lyase (4.3.-.-) or to ammonia lyases (4.3.1.-) to the more specific histidine-ammonia lyase (4.3.1.3) in the fourth position.

Several other such annotation schemes provide a controlled vocabulary to annotate sequences; the most prominently used annotation scheme is that of Gene Ontology (GO). The adopted controlled vocabulary is based on three aspects of gene product function: molecular function, biological process, and cellular location.

The primary purpose of the GO Annotation (GOA) project is to annotate genomes and their by-products using GO terms. When GO terms are assigned to a gene product, an evidence code stating how the annotation was obtained is assigned as well, so that the source of the annotation is noted. Thus, GO provides a standard means for programs to describe their functional predictions.

1.4.6.1.1.1 Sequence Homology-Based Functional Annotation — Traditional means of predicting the function of sequences rely on homology. These techniques are also known as the homology transfer technique, as they traverse databases of sequences to find matches between sequences and the query sequence. From the reported matches, a transfer of relevant functional information takes place in the query sequence.

A commonly adopted tool, basic local alignment search tool (BLAST), matches significant sequence similarity to other sequences in a database of experimentally annotated sequences. The biological rationale for using homology transfer is that if two sequences have a high degree of similarity, then they have evolved from a common ancestor and hence have similar, if not identical, functions.

However, this rational does not seem to hold with growing databases and fails in three conditions:

1. High sequence similarity does not guarantee accurate annotation transfer: When two sequences share functional similarity, it is observed that only certain regions of the sequences (subregions) contribute to the functional characterization of sequences. Thus, if two sequences share a higher degree of similarity, it does not imply that the subregions contribute to the exact matches or are being conserved. Moreover, it has been shown that enzymes that are supposedly analogous due to undetectable sequence similarity are in fact similar. It is believed that 35% sequence identity and 60% aligned enzymes share four EC numbers.

   Domain shuffling also contributes to the failure of homology transfer by adding, deleting, or redistributing domains of the sequence between homologous sequences.

2. Growing databases exhibit greater diversity in sequences that affect sequence-based tools to discover similarity between proteins: Here, with the evolution of databases, categorization of sequences is constantly changing. These changes make homology transfer more challenging, as the number of clustered similar proteins for which there is no reference sequence is also growing at the same rate.
3. Chances of propagating erroneous annotations throughout the database: As more sequences enter the database, errors in annotation are often propagated and amplified based on a single erroneous annotation.

The Pfam database is the most commonly used database for protein sequence analysis. A slew of other databases, such as InterPro, SMART, CDD, and PRODOM, use the annotations at the domain level derived from Pfam and provide the user multiple alignments of protein domains. Users of these programs need to take into consideration that Pfam does not address domain shuffling, and thus the results obtained could not be as accurate as anticipated.

1.4.6.1.1.2 Structure-Based Functional Annotation — Protein structural information is represented by a collection of 3D coordinates that correspond to the amino acids that make up the protein. This representation is computationally expensive; thus, algorithms have been designed to find ways of reducing this 3D representation while preserving the spatial and physicochemical information.

The functional annotation of proteins using the 3D structural information of proteins is built on the pretext that more information can be extracted from the structure than just the sequence information. That is, knowing the structure can yield better insight into the biochemical mechanism of how proteins function. The underlying hypothesis in structural methods is that if the 3D structure is of a known fold, then that protein may possess the function of proteins processing the corresponding fold. Moreover, structure is better conserved than sequence; thus, proteins with little or no sequence similarity still have structural similarity.

Traditional structural methods are dependent on structural alignment, which entails aligning a novel protein with other proteins from its fold. Functional transfer is performed by verifying whether the aligned proteins share the same catalytic sites that are believed to be conserved by amino acid content and side-chain orientations.

With proteins of unknown structural folds and low similarity to any known fold, functional annotation is still possible by analyzing structural patterns of the protein. Here, just like sequence-based patterns, the program looks for shared structural patterns between a novel protein and a protein of a known function.

Structural patterns are best described as 3D shapes completely dissociated from the amino acids or a string of characters representing amino acids and their physical environment. For example, one can look for 3D motifs to describe the function of a protein. Here, an algorithm creates a library of 3D motifs with associated functions. A search algorithm scans the library, attempting to match extracted 3D motifs from the protein molecule. The result is a map of potential functional sites for a given protein to a library of existing function sites.

1.5 Conclusion

In this chapter we highlight the accomplishments made after the completion of the HGP that have led to the formation of key areas of research. Though bioinformatics and computational biology has created a niche for itself, its applications can be felt in other areas, such as comparative genomics, functional genomics, and sequence variation analysis. With new technological innovations being made in these areas, there has been a volume of data that require analysis. To this end, this book is dedicated to understanding the principles of data mining and its applications in the area of bioinformatics.

References

Abnizova, I., and W.R. Gilks. Studying statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the eukaryotic genomes. *Briefings Bioinformatics* 7, no. 1(2006): 48–54.

Alexandersson, M., S. Cawley, and L. Pachter. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res* 13 (2003): 496–502.

Astier, Y., O. Braha, and H. Bayley. Toward single molecule DNA sequencing: Direct identification of ribonucleoside 5'-monophosphate by using an engineered protein nanopore equipped with a molecular adapter. *J Am Chem Soc* 128 (2006): 1705–1710.

Baumbach, J., A. Tauch, and S. Rahmann. Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Briefings Bioinformatics* 10, no. 1 (2008): 75–83.

Birney, E., M. Clamp, and R. Dirbin. GeneWise and Genomewise. *Genome Res* 14 (2004): 988–995.

Birney, E., J.D. Thompson, and T.J. Gibson. Pairwise and searchwise: Comparison of a protein profile to all three translation frames simultaneously. *Nucl Acids Res* 24 (1996): 2730–2739.

Cao, J., and H. Zhao. Estimating dynamic models for gene regulation networks. *Bioinformatics* 20, no. 14 (2008): 1619–1624.

Collins, F. S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282 (1998): 682–689.

Florea, L. Bioinformatics of alternative splicing and its regulation. *Briefings Bioinformatics* 7, no. 1 (2006): 55–69.

Florea, L., et al. Gene and alternative splicing annotation with AIR. *Genome Res* 15, no. 1 (2005): 54–66.

Friedberg, I. Automated protein function prediction—The genomic challenge. *Briefings Bioinformatics* 7, no. 3 (2006): 225–242.

Fujita, P.A., et al. The UCSC Genome Browser database: Update 2011. *Nucl Acids Res* 39 (2010): D876–D882.

Howe, K.L., T. Chothia, and R. Durbin. GAZE: A generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res* 12 (2002): 1418–1427.

Huang, X., and A. Madan. CAP3: A DNA sequence assembly program. *Genome Res 9, no. 9* (1999): 868–877.

Huang, Z., J. Li, H. Su, G.S. Watts, and H. Chen. Large-scale regulatory network analysis from microarray data: Modified Bayesian network learning and association rule mining. *Decision Support Syst* 43 (2007): 1207–1225.

Hubbard, T., et al. The Ensembl genome database project. *Nucl Acids Res* 30, no. 1 (2002): 38–41.

Kim, N., S. Shin, and S. Lee. ECgene: Genome-based EST clustering and gene modeling for alternative splicing. *Genome Res* 15, no. 4 (2005): 566–576.

Krishnan, V.G., and D.R. Westhead. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19, no. 17 (2003): 2199–2209.

Lee, C., and Q. Wang. Bioinformatics analysis of alternative splicing. *Briefings Bioinformatics* 6, no. 1 (2005): 23–33.

Lee, Y., et al. The TIGR Gene Indices: Clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucl Acids Res* 3 (2005): D71–D74.

Li, H., and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings Bioinformatics* 11, no. 5 (2010): 473–483.

Liang, F., I. Holt, G. Pertea, S. Karamycheva, S.L. Salzberg, and J. Quackenbush. An optimized protocol for analysis of EST sequences. *Nucl Acids Res* 28, no. 18 (2000): 3657–3665.

Lockwood, W.W., R. Chari, B. Chi, and W.L. Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur J Hum Genet* 14 (2006): 139–148.

Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24, no. 3 (2008): 133–141.

Modrek, B., A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucl Acids Res* 29, no. 13 (2001): 2850–2859.

Mooney, S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings Bioinformatics* 6, no. 1 (2005): 44–56.

NCBI. *NCBI: A science primer.* July 27, 2007. http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html#ref1 (accessed September 13, 2011).

Ng, P.C., and S. Henikoff. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12, no. 3 (2002): 436–446.

Ng, P.C., and S. Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genom Human Genet* 7 (2006): 61–80.

Parra, G., P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, and R. Guigo. Comparative gene prediction in human and mouse. *Genome Res* 13 (2003): 108–117.

Paszkiewicz, K., and D.J. Studholme. *De novo* assembly of short sequence reads. *Briefings Bioinformatics* 11, no. 5 (2010): 457–472.

Pedersen, J.S., and J. Hein. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* 19 (2003): 219–227.

Pennacchio, L.A., and E.M. Rubin. Comparative genomic tools and databases: Providing insights into the human genome. *J Clin Invest* 111 (2003): 1099–1106.

Rebbeck, T.R., M. Spitz, and X. Wu. Assessing the function of genetic variants in candidate gene association studies. *Genetics* 5, no. 8(2004): 589–597.

Schlessinger, A., and B. Rost. Protein flexibility and rigidity predicted from sequence. *Proteins* 61 (2005): 115–126.

Schlessinger, A., G. Yachdav, and B. Rost. PROFbval: Predict flexible and rigid residues in proteins. *Bioinformatics* 22, no. 7 (2006): 891–893.

Shendure, J.A., G.J. Porreca, and G.M. Church. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* 81, no. 7.1.1–7.1.11 (2008): 1–11.

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J.M. Jones, and I. Birol. ABySS: A parallel assembler for short read sequence. *Genome Res* 19, no. 6 (2009): 1117–1123.

Stamm, S., et al. Function of alternative splicing. *Gene* 344 (2005): 1–20.

Stitziel, N.O., T.A. Binkowski, Y.Y. Tseng, S. Kasif, and J. Liang. topoSNP: A topographic database of non-synonymous single nucleotide polymorphisms with and without known disease associations. *Nucl Acids* 32 (2004): D520–522.

Thomas, P.D., M.J. Campbell, A. Kejariwal, H. Mi, and B. Karlak. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13 (2003): 2129–2141.

Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579 (2005): 3346–3354.

U.S. National Library of Medicine. *Genetics home reference: Your guide to understanding genetic conditions.* September 5, 2011. http://ghr.nlm.nih.gov/ (accessed September 13, 2011).

Vercoutere, W., S. Winter-Hilt, H. Olsen, D. Deamer, D. Haussler, and M. Akeson. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nat Biotechnol* 19, no. 3 (2001): 248–252.

Webb, E.C. *Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* San Diego: Academic Press, 1992.

Winters-Hilt, S., W. Vercoutere, V.S. DeGuzman, D. Deamer, M. Akeson, and D. Haussler. Highly accurate classification of Watson-Crick basepair on termini of single DNA molecules. *J Biophys* 84, no. 2 (2003): 967–976.

Yue, P., and J. Moult. Identification and analysis of deleterious human SNPs. *J Mol Biol* 356 (2005): 1263–1274.

Zhang, M.Q. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* 3 (2002): 698–709.