CHAPTER 1

INTRODUCTION

Data mining for genomics and proteomics belongs to an interdisciplinary and relatively new field of bioinformatics, which evolves so rapidly that it is difficult to predict the extent and pace of the changes. Biology, or more generally life sciences, can now be considered information sciences. They are changing from disciplines that deal with relatively small data sets to research fields overwhelmed by a large number of huge data sets. Two main triggers are the source of these changes. The first was *the Human Genome Project*. As the result of research sparked by this project, we now have a large and growing library of organisms with already sequenced genomes. The second was a new technology—genomic microarrays—that allows for the quick and inexpensive measurement of gene expression level for thousands of genes simultaneously.

These and other changes have occurred during the last ten years or so. Before then, biologists and biomedical researchers were dealing with data sets typically consisting of dozens or perhaps hundreds of biological samples (patients, for example) and dozens of variables. The number of samples was typically greater than the number of variables. “Traditional” statistical methods were used and researchers did not have to think about *heuristic approaches* to overcome *the curse of dimensionality* as we do today.

Typical data sets generated with the use of current microarray technologies include many thousands of variables and only dozens or hundreds of biological samples. When exon arrays are more widely used or when protein chip technologies allow for direct quantification of the protein expression level on the whole-human-proteome scale, we may routinely analyze data sets with more than a million variables. The traditional univariate approach—*one-gene or one-protein-at-a-time*—is no longer sufficient. Different approaches are necessary and multivariate analysis has to become a standard one. There is nothing wrong with using the univariate analysis, but if research stops at that point, as is the case in some studies, a huge amount of generated data may be heavily underused, and potentially important biomedical knowledge not extracted. Here is where data miners should be involved.

Today, hardly any study involving high throughput gene or protein expression data is performed exclusively by biologists or biomedical scientists. Although few research groups realize the importance of including data miners in their studies, the role of a relatively new breed of scientists called *bioinformaticians* is indisputable. The bioinformaticians are not necessarily data miners, although data mining should

Data Mining for Genomics and Proteomics. By Darius M. Dziuda

be one of their required skills. There is a small but growing population of scientists who were majoring in bioinformatics. However, most of the experienced bioinformaticians are still either biologists who learned computer science and statistics, or computer scientists familiar with biology. The interdisciplinary nature of this field is expansive and involves many other disciplines, like physiology, clinical sciences, mathematics, physics, and chemistry.

Since this book is written for students and practitioners of data mining and bioinformatics as well as biomedical sciences, there may be some terms that are well known to some readers but new for others. We will start with a short explanation of some basic, mostly biological, terms that are relevant for our data mining focus.

1.1 BASIC TERMINOLOGY

1.1.1 The Central Dogma of Molecular Biology

The central dogma of molecular biology was originally introduced by Francis Crick in 1957 at a symposium of the Society for Experimental Biology in London and then published in 1958 (Crick 1958). The dogma[1] states that once the detailed sequence information has passed into protein it cannot be transferred to nucleid acid or protein. Crick's original description of the dogma (Crick 1958) was:

> ***The Central Dogma***
>
> *This states that once 'information' has passed into protein* ***it cannot get out again****. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the* ***precise*** *determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.*

Please note the qualification that 'information' here means the *sequential* information. Understood in its original meaning, the central dogma is still one of the fundamental ideas of molecular biology. Although introduced as a speculative idea, the central dogma holds true as well as there are plausible arguments that it is rather unlikely for it to be reversed (Crick 1970; Crick 1988).

The central dogma is quite often confused with the standard pathway of information flow from DNA to RNA to protein. To address misunderstandings about the dogma, Crick explained it in relation to three classes of transfers of sequential information: general transfers (ones that commonly occur), special transfers (may occur in special situations), and unknown transfers. The central dogma is about the *unknown transfers*—protein to protein, protein to DNA, and protein to RNA—and it postulates that these transfers never occur (Crick 1970).

[1] Francis Crick admits in his autobiography (Crick 1988) that he was criticized for using the word *dogma* in the situation where the word *hypothesis* would be more appropriate. Crick argues, however, that he already used the term hypothesis in the *sequence hypothesis* introduced at the same time (Crick 1958) and wanted to emphasize the more powerful and central position of the "dogma."

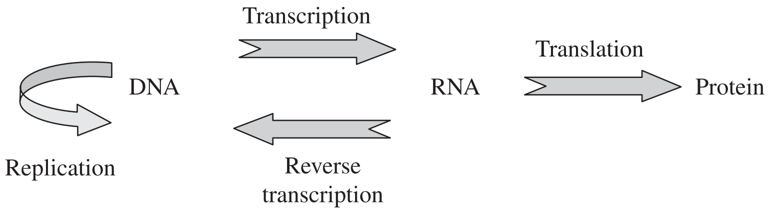

Figure 1.1: The basic flow of sequential information: DNA to DNA (replication), DNA to RNA (transcription), RNA to protein (translation), and RNA to DNA (reverse transcription). The central dogma of molecular biology states that "*once (sequential) information has passed into protein it cannot get out again*" (Crick 1970).

The current knowledge of information transfer is consistent with the central dogma. The standard pathway of information flow describes the process, in which proteins are synthesized based on DNA information (Fig. 1.1): DNA is ***transcribed*** into RNA, and then RNA—or more precisely, mRNA (messenger RNA)[2]—is ***translated*** into protein. These are two of the Crick's general transfers. The third one is *replication* (DNA to DNA). Special transfers are represented by *reverse transcription* (RNA to DNA). There is no evidence for the *unknown transfers*.

In humans (and other eukaryotic[3] organisms), transcription takes place in the cell nucleus, and translation in the cytoplasm—outside the nucleus. Most human genes contain noncoding sequences (*introns*), which have to be removed from mRNA before it is translated into protein. The process that eliminates introns is called *splicing*. During translation, which takes place at ribosomes, the mRNA sequential information is translated into a string of amino acids that are used to synthesize the protein. First, the mRNA sequence is divided into three-letter codons representing amino acids. Subsequently, the amino acids are linked together to create the protein. Translation ends when one of the stop codons is encountered, and the newly created protein leaves the ribosome.

1.1.2 Genome

The term *genome* can be understood either as the complete set of genetic information or as the entire set of genetic material contained in an organism's cell. When applied to the human genome, this definition includes both the *nuclear genome* and the *mitochondrial genome.*[4] Nevertheless, in the area of gene expression analysis, we often use the term genome as referring to the nuclear genome only, that is, understood as the complete DNA sequence of one set of the organism's chromosomes.

[2] Crick referred to RNA since mRNA had yet to be discovered when he formulated the dogma.

[3] Eukaryotic cells are cells that have a nucleus. Eukaryotes, that is, organisms with eukaryotic cells, may be unicellular or multicellular (e.g., fungi, plants, and animals). Prokaryotes are unicellular organisms (such as bacteria) that have no nuclei. *Pro* in the term means "prior to" and *karyot* means "nucleus". The prefix *eu* means "true" (Garrett and Grisham 2007).

[4] The mitochondrial genome represents organellar genomes carried by cells of most eukaryotic organisms. Another example of organellar genomes is the *chloroplast genome* in plants.

In humans, every nucleated cell (circulating mature red blood cells have no nucleus) contains the nuclear genome organized within the nucleus into the DNA molecules called chromosomes—22 pairs of autosomes (nonsex chromosomes) and two heterosomes (sex chromosomes). The length of the human genome sequence is about 3×10^9 nucleotides (base pairs).[5] For two randomly selected humans, the order of nucleotides in their genomes is about 99.9% identical (it is more than that if the two are related). However, 0.1 percent of the three billion bases amounts to three million places where two such genomes differ. The space for different DNA sequences is huge, beyond huge actually (up to $4^{3000000}$).[6] We are different; no two humans (with the exception of identical twins) have identical DNA.

The Human Genome Project was an international research program aimed at obtaining a high-quality sequence of the euchromatic (i.e., gene-rich) portion of the human genome. The project was initiated in 1990 and was officially completed in 2003. In February 2001, two draft versions of the human genome sequence were simultaneously announced and published, one from the International Human Genome Sequencing Consortium (International Human Genome Sequencing Consortium 2001) and the other from Celera Genomics (Venter et al. 2001). Each of the drafts contained over 100,000 gaps and the draft sequences were missing about 10% of the euchromatic portion of the genome (Stein 2004). In 2003, the 'finished' human genome sequence was announced. It covered 99% of the euchromatic portion of the human genome (and about 94% of the total genome), had only 341 gaps,[7] and contained only about one error per 100,000 bases (International Human Genome Sequencing Consortium 2004).

1.1.3 Proteome

A simplified definition could state that the *proteome* is the complete set of protein products expressed by the genome (see the central dogma). However, unlike the genome that can be considered a rather stable entity, the proteome constantly changes (mainly due to protein–protein interactions and changes in a cell's environmental conditions). Furthermore, the set of proteins expressed in a cell depends on the type of cell. Thus, a proteome can also be interpreted as a snapshot of all the proteins expressed in a cell or a tissue at a particular point in time. This means that depending on the context, we may refer to the single proteome of an organism (i.e., the *complete proteome* understood as

[5] This is the length of the *haploid* human genome, that is, the length of the DNA sequence of the 24 distinct chromosomes, one from each pair of the 22 autosomes and the 2 heterosomes. The *diploid* human genome, including the DNA sequence from all 46 chromosomes, would have about 6×10^9 nucleotides, or six gigabases (6 Gb). The latest technological advances allowed for sequencing the diploid genome of James Watson in two months (Wheeler et al. 2008).

[6] This number is the upper limit of potentially significant differences since not all changes in the DNA sequence are necessarily associated with differences in functions.

[7] Many of these gaps were associated with segmental duplications that could not be sequenced with available technologies. This was also one of the reasons why The Human Genome Project did not target the heterochromatic (gene-poor) regions of the human genome, which contain highly repetitive sequences. The estimated size of the gaps was: about 2.8×10^7 bases in the euchromatic portion and about 2.0×10^8 bases in the heterochromatic portion of the genome.

the set of all protein products that can be expressed in the organism) or to many *cellular proteomes* that are qualified by location and time.

The number of proteins in the human proteome is much larger than the number of the underlying protein-coding genes. Currently, the number of protein-coding genes in the human genome is estimated to be under 21,000 (Clamp et al. 2007). The current estimate for the size of the complete human proteome is about one million proteins. Why are there more proteins than genes if each protein is synthesized by reading the sequence of a gene? This is due to such events as *alternative splicing* of genes and *post-translational modifications* of proteins.[8]

The Human Proteome Organisation (HUPO) plans to identify and characterize all proteins in the complete human proteome. However, due to the scale and complexity of this task, the goal of the first phase of the Human Proteome Project (HPP) is limited to the identification of one representative protein for each protein-coding human gene. After this first stage, the catalogue of human proteins will be extended to eventually include all protein isoforms (Uhlen 2007; Pearson 2008; Service 2008).

1.1.4 DNA (Deoxyribonucleic Acid)

DNA is a nucleic acid that encodes genetic information. DNA is capable of self-replication and synthesis of RNA. In 1944, Oswald Avery and colleagues identified DNA as the genetic material of inheritance (Avery et al. 1944). In 1953, James Watson and Francis Crick proposed and described a model of DNA as the double helical structure[9] (Watson and Crick 1953a, 1953b).

DNA consists of two long chains[10] (strands) of nucleotides twisted into a *double helix* (Fig. 1.2) and joined by hydrogen bonds between the complementary nucleotide bases. Each of the strands has a sugar phosphate backbone, to which the bases are covalently attached.

There are four types of nucleotides in DNA and each of them contains a different base: *adenine* (A), *guanine* (G), *cytosine* (C), or *thymine* (T).[11] The four bases are letters of the alphabet in which genetic information is encoded. Strings of these letters are read only in one direction. Each of the two DNA strands has its polarity—the $5'$ end and the $3'$ end[12]—and the two complementary strands are bound with the opposite

[8] Alternative splicing is described in Section 1.1.8. Post-translational modifications are various chemical alterations that proteins may undergo after their synthesis. Examples of these modifications include the enzymatic cleavage (cutting a protein into smaller functional proteins), covalent attachment of additional biochemical groups (e.g., phosphorylation, methylation, oxidation, acetylation), and forming covalent bonds between proteins.

[9] Watson and Crick admitted that their work was stimulated by unpublished experimental results (X-ray crystallographic data) and ideas of Rosalind Franklin and Maurice Wilkins; a part of these results was published in the same issue of *Nature* as the Watson and Crick main paper (Franklin and Gosling 1953; Wilkins et al. 1953).

[10] The entire DNA in one human cell (in all 46 chromosomes) is about two meters long, yet fits into a cell nucleus, which is 2–3 micrometers wide.

[11] Due to their chemical structure, A and G nucleotides are purines, whereas C and T are pyrimidines.

[12] These ends correspond to $5'$ and $3'$ carbon atoms in a ribose residue, which are not linked to another nucleotide (Singleton 2008).

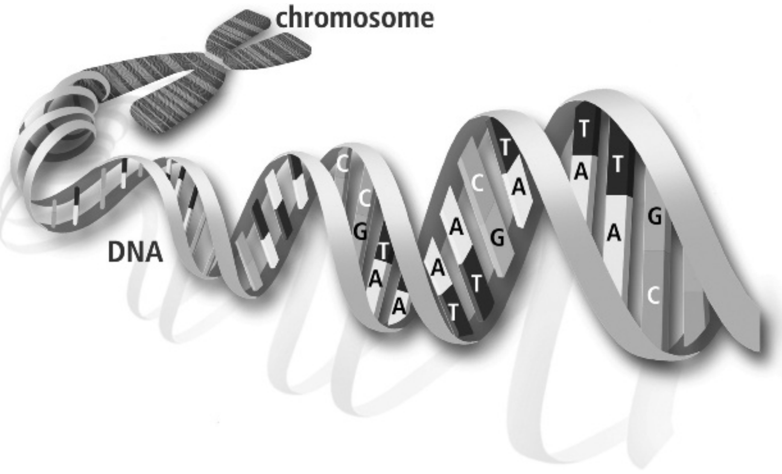

Figure 1.2: The DNA double helix (courtesy: The U.S. Department of Energy Genome Programs, http://genomics.energy.gov). The DNA structure includes two antiparallel deoxyribose-phosphate helical chains, which are connected via hydrogen bonds between complementary bases on the chains (base pairs). The base pairs "rungs of the ladder" are spaced 0.34 nm apart. This double helix structure repeats itself every 10 base pairs, or 3.4 nm (Watson and Crick 1953a; Garrett and Grisham 2007). (See color insert.)

directionality (i.e., they are *antiparallel*). The sequence of nucleotides determines the genetic information. In nature, base pairs form only between A and T and between G and C. Therefore, the sequence of bases in one strand of DNA can be deduced from the sequence of bases in its complementary strand (which enables DNA replication). During transcription, the sequential information contained in DNA is transferred into the single-stranded RNA.[13]

The latest discoveries indicate that the majority of human DNA is transcribed into various RNA transcripts (The ENCODE Project Consortium 2007). The messenger RNA (mRNA) is the only type of these RNA transcripts that are subsequently translated into proteins. Consequently, these discoveries indicate that there are few unused sequences in the human genome.

1.1.5 RNA (Ribonucleic Acid)

RNA is a nucleic acid molecule similar to DNA but containing the sugar component *ribose* rather than deoxyribose[14] and the *uracil* (U) base instead of thymine (T) in DNA. RNA is formed upon a DNA template, but is almost always single-stranded

[13] The DNA strand that has the same sequence as the synthesized transcript RNA (except for the replacement of Ts by Us) is called the *sense* strand. The other DNA strand is called the *antisense* strand (Campbell and Farrell 2006).

[14] The prefix *deoxy* means that an oxygen atom is missing (H instead of OH) from one of the ribose carbon atoms.

and has a much shorter sequence of nucleotides. There are many kinds of RNA molecules, with the three main classes of cellular RNA being: mRNA—messenger RNA; tRNA—transfer RNA; and rRNA—ribosomal RNA. Other classes of RNAs include microRNA (miRNA) and small interfering RNA (siRNA), both of which can regulate gene expression during and after transcription. Since only mRNAs are translated into proteins, some genes do not encode proteins but various RNA functional products.

Uracil is very similar to thymine and they both carry the same information. One may ask why there is thymine in DNA, but uracil in RNA. One of the common DNA mutations is chemical degradation of cytosine that forms uracil. If uracil was a valid base in DNA, cellular repair mechanisms could not efficiently correct such mutations. Thymine is then more appropriate for the stability of long-lived DNA. For short-lived RNA molecules, for which quantity is more important than long-term stability, uracil is more appropriate since it is energetically less expensive to produce than thymine.

1.1.6 mRNA (messenger RNA)

Messenger RNA (mRNA) is the type of RNA that contains coding information for protein synthesis. Messenger RNA is transcribed from a DNA template (a sequence of bases on one strand of DNA), and subsequently undergoes maturation that includes splicing. The mature mRNA is then transported from the cell nucleus to the cytoplasm where proteins are made during the process called translation. During translation, which takes place at the cytoplasm's ribosomes, the mRNA information contained in the sequence of its nucleotides is translated into amino acids. Amino acids are the building blocks of proteins. First, the mRNA sequence is divided into three-letter codons representing amino acids. Subsequently, the amino acids are linked together to produce a polypeptide. Translation ends when one of the stop codons is encountered. A functional protein is the result of properly folded polypeptide or polypeptides.

1.1.7 Genetic Code

The four bases of mRNA—adenine (A), guanine (G), cytosine (C), and uracil (U)—are the "letters" of the *genetic code*. Amino acids, which are building blocks of proteins, are coded by three-letter words of this code, called *codons*. The genetic code (Table 1.1) defines the relation between mRNA codons and amino acids of proteins.

With the four letters of the genetic code, there are $4^3 = 64$ possible triplets (codons). All 64 codons are meaningful: 61 of them code for amino acids and three (UAA, UAG, and UGA) serve as the *stop codons* signaling the end of the protein. The stop codons are called *nonsense codons* as they do not correspond to any amino acid. AUG codes for methionine, but is also the *start codon*—a part of the initiation signal to start protein synthesis.

Since there are 61 *sense* codons and only 20 amino acids in proteins (see Table 1.2), the genetic code is *degenerate*, which means that almost all amino acids

TABLE 1.1: Genetic Code

	Second Base in Codon				
First Base in Codon	U	C	G	A	Third Base in Codon
U	Phe	Ser	Cys	Tyr	U
	Phe	Ser	Cys	Tyr	C
	Leu	Ser	Trp	*Stop*	G
	Leu	Ser	*Stop*	*Stop*	A
C	Leu	Pro	Arg	His	U
	Leu	Pro	Arg	His	C
	Leu	Pro	Arg	Gln	G
	Leu	Pro	Arg	Gln	A
G	Val	Ala	Gly	Asp	U
	Val	Ala	Gly	Asp	C
	Val	Ala	Gly	Glu	G
	Val	Ala	Gly	Glu	A
A	Ile	Thr	Ser	Asn	U
	Ile	Thr	Ser	Asn	C
	Met	Thr	Arg	Lys	G
	Ile	Thr	Arg	Lys	A

The gray background denotes *four-fold degeneracies*—codons with irrelevant third base. Sixty one out of 64 codons code for 20 amino acids. Three codons are the stop codons (UAA, UAG, and UGA). The AUG is the start codon, but it also codes for methionine. Codons that code for the same amino acid (e.g., the six codons coding for Leu, or leucine) are called *synonymous codons*.

(18 out of 20) are associated with more than one codon. Methionine (Met) and tryptophan (Trp) are exceptions—each of them is coded by only one codon.

When the third base of a codon is irrelevant, that is, four codons with the same first and second base code for the same amino acid, we have *four-fold degeneracy* (sometimes called *third-base degeneracy*). The four-fold degenerate families of codons (such as UC* or CU*, where the symbol * denotes an irrelevant third base) are marked in Table 1.1 with the gray background. Another main family of degeneracy is *two-fold degeneracy*, when two codons with a different base in the third position are associated with the same amino acid (for instance, UG[U|C], where the [U|C] notation means the third base is either U or C).

TABLE 1.2: Amino Acids

Amino acid	3-letter code	1-letter code	Number of codons	Codons
Alanine	Ala	A	4	GC*
Arginine	Arg	R	6	CG*, AG[A|G]
Asparagine	Asn	N	2	AA[U|C]
Aspartic acid	Asp	D	2	GA[U|C]
Cysteine	Cys	C	2	UG[U|C]
Glutamic acid	Glu	E	2	GA[A|G]
Glutamine	Gln	Q	2	CA[A|G]
Glycine	Gly	G	4	GG*
Histidine	His	H	2	CA[U|C]
Isoleucine	Ile	I	3	AU[U|C|A]
Leucine	Leu	L	6	CU*, UU[A|G]
Lysine	Lys	K	2	AA[A|G]
Methionine	Met	M	1	AUG
Phenylalanine	Phe	F	2	UU[U|C]
Proline	Pro	P	4	CC*
Serine	Ser	S	6	UC*, AG[U|C]
Threonine	Thr	T	4	AC*
Tryptophan	Trp	W	1	UGG
Tyrosine	Tyr	Y	2	UA[U|C]
Valine	Val	V	4	GU*

Twenty amino acids that are building blocks of proteins. Amino acids are coded by one to six codons. For example, arginine is coded by six codons, CG* and AG[A|G]. The CG* notation means four codons starting with CG, that is, [CGU, CGC, CGG, CGA], and AG[A|G] means that the first and second bases are AG, and the third base is either A or G, [AGA, AGG].

1.1.8 Gene

For the purpose of mining gene expression data, we could use the following description of a gene:

> *A gene is the segment of DNA, which is transcribed into RNA, which may then be translated into a protein. Human (and other eukaryotic) genes often include non-coding sequences (introns) between coding regions (exons). Such genes encode one or more functional products, which are proteins or RNA transcripts.*[15] *Each gene is associated with a promoter sequence, which initiates gene transcription and regulates its expression.*

This description is satisfactory for the analysis of gene expression microarray data. Nevertheless, as a definition of the gene this description may be incomplete. The same is true for other definitions of the gene that were proposed during the last hundred years. Although it may be true that none of those definitions has been

[15] A transcript is the RNA product of the gene transcription process.

generally accepted (Falk 1986; Rheinberger and Müller-Wille 2008), some of them were popular enough to prevail in their time. Here are a few examples of the concept of a gene as it was evolving in time (Gerstein et al. 2007; Rheinberger and Müller-Wille 2008).

- *A gene as a discrete unit of heredity that determines a characteristic of an organism as well as the heritability of this characteristic.*
- *A gene as a stretch of DNA that codes for one protein.*
- *A gene as "a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology"* (Wain et al. 2002; HUGO Gene Nomenclature Committee 2008).

New discoveries have been forcing changes to the concept of a gene. The following are just a few among such discoveries:

- *Splicing*—the process of removing introns from the primary RNA transcript and using only exons to form the mature mRNA (see Fig. 1.3). Splicing redefined the gene as including a series of exons (coding sequences) separated by introns (noncoding sequences).
- *Alternative splicing*—generally, combining different subsets of gene exons to form (code for) different mRNA transcripts (see Fig. 1.4). Alternative splicing invalidated the "one gene—one protein" paradigm.
- Nonprotein-coding genes—genes that "*encode functional RNA molecules*[16] *that are not translated into proteins*" (HUGO Gene Nomenclature Committee 2008).
- Overlapping protein-coding genes—genes sharing the same DNA sequence. New exons have been discovered far away from the previously known gene locations, some of them within the sequence of another gene (Pennisi 2007). It is possible for a gene to be completely contained within an intron of another gene; it is also possible for two genes to share the same stretch of DNA without sharing any exons (Gerstein et al. 2007).
- Nonprotein-coding genes sharing the same DNA sequence with protein-coding genes (Gingeras 2007).
- Some RNA transcripts are composed from exons belonging to two genes.
- Transcribed *pseudogenes*. A pseudogene is a sequence of DNA that is almost identical to a sequence of an existing and functional gene but is or appears to be inactive, for instance, due to mutations that make it nonfunctional (Ganten and Ruckpaul 2006; Singleton 2008). However, the Encyclopedia of DNA Elements project (The ENCODE Project Consortium 2007) revealed recently that some pseudogenes—although by definition lacking protein coding potential—can produce RNA transcripts that could potentially have some regulatory functions (Gerstein et al. 2007; Gingeras 2007; Zheng et al. 2007).

[16] Some of these nonprotein-coding RNA transcripts are well known functional RNA molecules, for example ribosomal RNAs (rRNA) or transfer RNAs (tRNA). Others include recently discovered microRNAs (miRNA) and small interfering RNAs (siRNA); both of them can regulate expression of specific genes (Gingeras 2007).

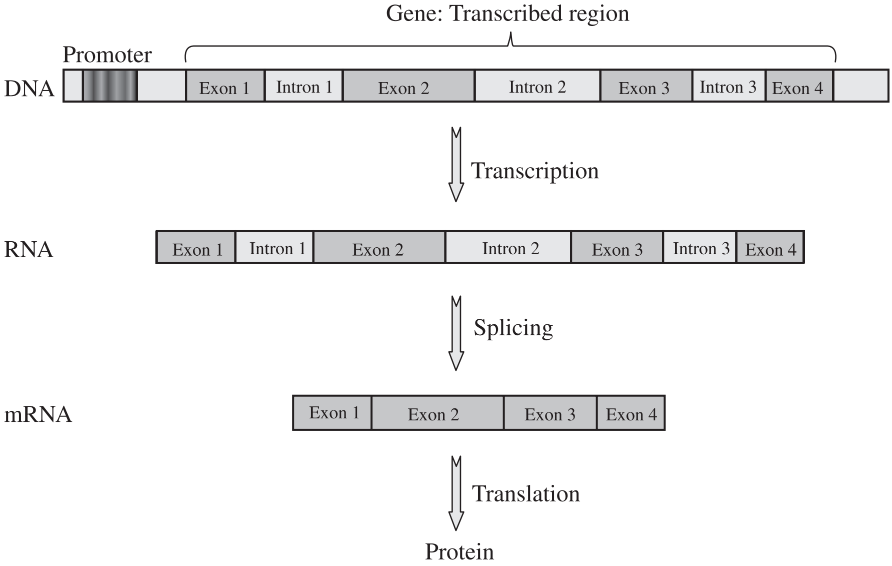

Figure 1.3: A human protein-coding gene: the structure, transcription, splicing, and translation. The top row shows a schematic structure of the gene. The transcribed region consists of exons that are interrupted by introns (on average, introns are about 20 times longer than exons). The promoter associated with the gene is a regulatory sequence that facilitates initiation of gene transcription and controls gene expression. Enhancers and silencers are other gene-associated regulatory sequences that may activate or repress transcription of the gene (they are not shown here as they are often located distantly from the transcribed region). The gene's exons and introns are first transcribed into a complementary RNA (called nuclear RNA, primary RNA transcript or pre-mRNA). Then, the splicing process removes introns, joins exons, and creates mRNA (called also mature mRNA). The mRNA travels from the nucleus to the cytoplasm where, in ribosomes, it is translated into a protein. (See color insert.)

Since the concept of a gene has been changed over and over again, we may consider one of two options: (i) stop using this concept at all, or (ii) formulate a gene definition in a way that is either quite general (and perhaps vague) or deliberately verbose in an attempt to cover all possibilities. Here are two recently proposed gene definitions:

- "*A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.*" (Gerstein et al. 2007)
- "*A gene is a discrete genomic region whose transcription is regulated by one or more promoters and distal regulatory elements and which contains the information for the synthesis of functional proteins or non-coding RNAs, related by the sharing of a portion of genetic information at the level of the ultimate products (proteins or RNAs).*" (Pesole 2008)

These two definitions provide different answers to the question "*What is a gene?*" Consequently, if we were able to enumerate all the genes in the human

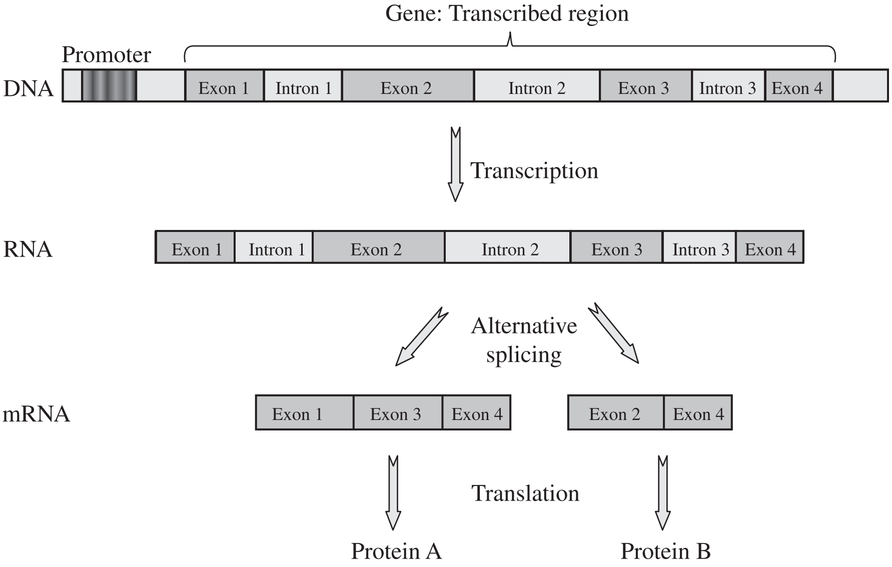

Figure 1.4: A human protein-coding gene: an example of alternative splicing. During alternative splicing of the primary RNA transcript, different subsets of exons are joined to create two (or more) different mRNA isoforms. These mRNA isoforms are then translated into usually distinct proteins. The majority of human protein-coding genes undergo alternative splicing (Stamm 2006). (See color insert.)

genome, these definitions would also lead to different answers to the question "*How many genes are there in the human genome?*" To simplify enumeration of human genes, we may start with counting only protein-coding genes. A recently performed analysis of the three most popular human gene databases[17] suggests that there may be fewer than 21,000 protein-coding genes in the human genome (Clamp et al. 2007).

Although cells of different tissues have the same genome, only a limited and tissue-specific subset of genes is expressed in each tissue. While tissues can be identified by general patterns of their gene expression, these patterns are not constant. Gene expression levels in a cell change in time and in response to changes in a cell's environmental conditions. Changes in the expression profile (turning on or off some genes or changing expression levels of some of the expressed genes) alter protein production and may result in significant changes in tissue functioning; for instance, in transforming a normal tissue into a cancerous one.

1.1.9 Gene Expression and the Gene Expression Level

Gene expression is the process that converts information encoded in a gene into functional products of the cell (see the central dogma). In the context of genomic projects

[17] The three databases are: Ensembl (Flicek et al. 2008), RefSeq (Pruitt et al. 2007), and Vega (Wilming et al. 2008).

involving gene expression microarray data, we are interested in the *gene expression level* that refers to the number of copies of RNA transcripts created by transcription of a particular gene (for the protein-coding genes, this is the number of mRNA transcripts generated from the gene) at a given time. The genes that are expressed include the protein-coding genes as well as the genes coding for RNA functional products. Most often, the gene expression analysis focuses, however, on the expression level of the protein-coding genes.

Although the same DNA is contained in every cell of an organism, cells of different tissues are different. The differences result from different levels of gene expressions, that is, different patterns of gene activations. If a gene is active (expressed), the protein (or proteins) encoded by the gene is (are) synthesized in the cell. The higher the gene expression level, the more of the protein is produced. There may be various reasons for a gene to be over-expressed, under-expressed or not expressed at all—hereditary factors, environmental factors, or their combinations. The hereditary factors may mean a mutation in a single gene, simultaneous mutations in a set of genes, chromosomal abnormalities, etc. If this prevents or significantly influences the production of one or more important proteins, we have hereditary diseases (such as cystic fibrosis for a single gene mutation, or Down syndrome when there are three copies of chromosome 21). On the other hand, many diseases are not necessarily related to genetic abnormalities, but are caused by the environmental factors that are changing the expression level of otherwise “normal” genes. This may be more complicated when such changes in the gene expression level are still moderated by some gene mutations, which by themselves are not causing any diseases. And here is the promise of genomics—identification of gene expression patterns associated with a disease and linking the patterns to underlying biological and environmental factors may lead to cure or prevention.

1.1.10 Protein

Protein is a biological macromolecule composed of one or more chains of amino acids synthesized in the order determined by the DNA sequence of the gene coding for the protein. Short chains of amino acids are called *peptides* and longer ones *polypeptides*. Since there is no precise distinction between them, we can say that the term *polypeptide* is used for chains longer than several dozen amino acids. We can define protein as a molecule composed of one or more polypeptide chains (Garrett and Grisham 2007).

A protein folds into a three-dimensional structure, which determines the function of the protein. The way the protein folds into its three-dimensional form is determined by its sequence of amino acids. Proteins are essential components of all living cells and each of them has a unique function. Examples of proteins include enzymes, hormones, and antibodies.

The number of proteins in the human proteome is much larger than the number of protein-coding genes in the human genome. This is mostly due to *alternative splicing* (when more than one protein is produced by different combinations of exons of the same gene) and *post-translational modifications* of proteins.

1.2 OVERLAPPING AREAS OF RESEARCH

Biomedical research based on the analysis of gene or protein expression data is an interdisciplinary field including molecular biology, computational biology, bioinformatics, data mining, computer science, statistics, mathematics, . . . and the list of overlapping disciplines is far from being exhaustive. New terms are constantly coined. Furthermore, the same term may have different meanings in some of the overlapping areas. Here are a few terms that are important for the subject of this book.

1.2.1 Genomics

Genomics can be understood as the systematic analysis of genome-scale data in order to expand biomedical knowledge. Genomic studies investigate structure and function of genes and do this simultaneously for all the genes in a genome. Investigating patterns of gene expression—one of the main themes of this book—is an important part of *functional genomics* that focuses on analysis of genes and their products at the functional level.

Although there are opinions that genomics and genetics should be combined and considered together, it is not currently the case and genetics is still understood as the study of genes in the context of inheritance.

1.2.2 Proteomics

Proteomics is the study of functions and structures of proteins. Proteomics is often seen as the next stage of research after genomics, and as the area that should give us more direct insight into biological processes (since proteins are direct players in the cell physiology whereas genes are mostly intermediate entities). Proteomics seems to be much more complicated than genomics, the main reason being that proteomes are constantly changing and that different cells of the same organism may have different proteomes.

1.2.3 Bioinformatics

One may find many definitions of *bioinformatics* and it is not unusual for them to limit this field to the areas researched by the definition authors. Defining it more generally, we may say that bioinformatics is the science of managing, analyzing, mining, and interpreting biological data. Bioinformatics, computational biology, and data mining are overlapping in their use of mathematical, statistical, and computer science tools and methods to analyze large sets of biological data.

1.2.4 Transcriptomics and Other -omics . . .

Though the term *genomics* is already well established in biomedical sciences, the term itself is an "*unusual scientific term because its definition varies from person to person*" (Campbell and Heyer 2007). Once genomics became popular, many new related -*omics* terms were created. Whether their areas belong to genomics or not is pretty much an open question. For example, *transcriptomics* is most often defined[18] as the area covering analysis of gene expression data (expressed genes are called *transcripts*). One could then say that the focus of this book is transcriptomics. Nevertheless, we will use the widely recognized term *genomics*, which has been used for this kind of investigations since the beginning of high-throughput gene expression data analysis.

1.2.5 Data Mining

We define *data mining* as efficient ways of extracting new information or new knowledge from large data sets or databases. Some more classical definitions limit data mining to extraction of 'useful information.' However, in biomedical research any new information reflecting underlying biological processes can be potentially useful. Furthermore, translating the extracted new information into new biomedical knowledge is often a nontrivial task and the data mining definition (and methods) should be extended as well to include the information-to-knowledge stage of investigations.

[18] There are other definitions of transcriptomics, one—for example—defines it as the study of transcription process itself.