Chapter One

Biological Foundations

No single agreed-upon definition seems to exist for the term *bioinformatics*, which has been used to mean a variety of things ranging in scope and focus. To cite but a few examples from textbooks, Lodish et al. (2000) state that "bioinformatics is the rapidly developing area of computer science devoted to collecting, organizing, and analyzing DNA and protein sequences." A more general and encompassing definition, given by Brown (2002), is that bioinformatics is "the use of computer methods in studies of genomes." More general still: "bioinformatics is the science of refining biological information into biological knowledge using computers" (Draghici, 2003). Kohane et al. (2003) observe that the "breadth of this commonly used definition of bioinformatics risks relegating it to the dustbin of labels too general to be useful" and advocate being more specific about the particular bioinformatics techniques employed.

While it is true that the field of bioinformatics has traditionally dealt primarily with biological data encoded in digital symbol sequences, such as nucleotide and amino acid sequences, in this book we will be mainly concerned with extracting information from gene expression measurements and genomic signals. By the latter we mean any measurable events, principally the production of messenger ribonucleic acid (RNA) and protein, that are carried out by the genome. The analysis, processing, and use of genomic signals for gaining biological knowledge and translating this knowedge into systems-based applications is called *genomic signal processing*.

In this chapter, our aim is to place this material into a proper biological context by providing the necessary background for some of the key concepts that we shall use. We cannot hope to comprehensively cover the topics of modern genetics, genomics, cell biology, and others, so we will confine ourselves to brief overviews of some of these topics. We particularly recommend the book by Alberts et al. (2002) for a more comprehensive coverage of these topics.

1.1 GENETICS

Broadly speaking, genetics is the study of genes. The latter can be studied from different perspectives and on a molecular, cellular, population, or evolutionary level. A gene is composed of deoxyribonucleic acid (DNA), which is a double helix consisting of two intertwined and complementary nucleotide chains. The entire set of DNA is the *genome* of the organism. The DNA molecules in the genome are assembled into *chromosomes*, and genes are the functional regions of DNA.

Each gene encodes information about the structure and functionality of some protein produced in the cell. Proteins in turn are the machinery of the cell and the major determinants of its properties. Proteins can carry out a number of tasks, such as catalyzing reactions, transporting oxygen, regulating the production of other proteins, and many others. The way proteins are encoded by genes involves two major steps: transcription and translation. *Transcription* refers to the process of copying the information encoded in the DNA into a molecule called messenger RNA (mRNA). Many copies of the same RNA can be produced from only a single copy of DNA, which ultimately allows the cell to make large amounts of proteins. This occurs by means of the process referred to as *translation*, which converts mRNA into chains of linked amino acids called *polypeptides*. Polypeptides can combine with other polypeptides or act on their own to form the actual proteins. The flow of information from DNA to RNA to protein is known as the *central dogma* of molecular biology. Although it is mostly correct, there are a number of modifications that need to be made. These include the processes of reverse transcription, RNA editing, and RNA replication.

Briefly, *reverse transcription* refers to the conversion of a single-stranded RNA molecule to a double-stranded DNA molecule with the help of an enzyme aptly called *reverse transcriptase*. For example, HIV virus consists of an RNA genome that is converted to DNA and inserted into the genome of the host. *RNA editing* refers to the alteration of RNA after it has been transcribed from DNA. Therefore, the ultimate protein product that results from the edited RNA molecule does not correspond to what was originally encoded in the DNA. Finally, *RNA replication* is a process whereby RNA can be copied into RNA without the use of DNA. Several viruses, such as hepatitis C virus, employ this mechanism. We will now discuss some preliminary concepts in more detail.

1.1.1 Nucleic Acid Structure

Almost every cell in an organism contains the same DNA content. Every time a cell divides, this material is faithfully replicated. The information stored in the DNA is used to code for the expressed proteins by means of transcription and translation. The DNA molecule is a polymer that is strung together from monomers called deoxyribonucleotides, or simply *nucleotides*, each of which consists of three chemical components: a sugar (deoxyribose), a phosphate group, and a nitrogenous base. There are four possible bases: adenine, guanine, cytosine, and thymine, often abbreviated as A, G, C, and T, respectively. Adenine and guanine are *purines* and have bicyclic structures (two fused rings), whereas cytosine and thymine are *pyrimidines*, and have monocyclic structures. The sugar has five carbon atoms that are typically numbered from $1'$ to $5'$. The phosphate group is attached to the $5'$-carbon atom, whereas the base is attached to the $1'$ carbon. The $3'$ carbon also has a hydroxyl group (OH) attached to it.

Figure 1.1 illustrates the structure of a nucleotide with a thymine base. Although this figure shows one phosphate group, up to three phosphates can be attached. For example, adenosine $5'$-triphosphate (ATP), which has three phosphates, is the molecule responsible for supplying energy for many biochemical cellular processes.

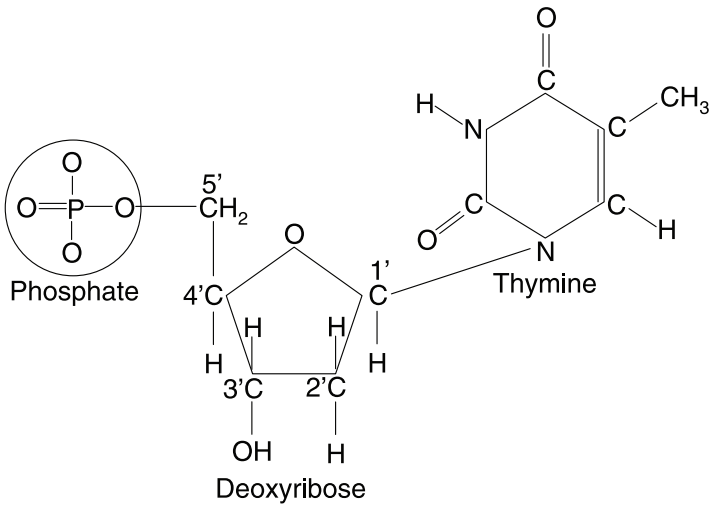

Figure 1.1. The chemical structure of a nucleotide with a thymine base.

Ribonucleic acid is a polymer that is quite close in structure to DNA. One of the differences is that in RNA the sugar is ribose rather than deoxyribose. While the latter has a hydrogen at the $2'$ position (figure 1.1), ribose has a hydroxyl group at this position. Another difference is that the thymine base is replaced by the structurally similar uracil (U) base in a ribonucleotide.

The deoxyribonucleotides in DNA and the ribonucleotides in RNA are joined by the covalent linkage of a phosphate group where one bond is between the phosphate and the $5'$ carbon of deoxyribose and the other bond is between the phosphate and the $3'$ carbon of deoxyribose. This type of linkage is called a *phosphodiester bond*. The arrangement just described gives the molecule a $5' \rightarrow 3'$ polarity or directionality. Because of this, it is a convention to write the sequences of nucleotides starting with the $5'$ end at the left, for example, $5'$-ATCGGCTC-$3'$. Figure 1.2 is a simplified diagram of the phosphodiester bonds and the covalent structure of a DNA strand.

DNA commonly occurs in nature as two strands of nucleotides twisted around in a double helix, with the repeating phosphate–deoxyribose sugar polymer serving as the backbone. This backbone is on the outside of the helix, and the bases are located in the center. The opposite strands are joined by hydrogen bonding between the bases, forming *base pairs*. The two backbones are in opposite or antiparallel orientations. Thus, one strand is oriented $5' \rightarrow 3'$ and the other is $3' \rightarrow 5'$. Each base can interact with only one other type of base. Specifically, A always pairs with T (an $A \cdot T$ base pair), and G always pairs with C (a $G \cdot C$ base pair). The bases in the base pairs are said to be *complementary*. The $A \cdot T$ base pair has two hydrogen bonds, whereas the $G \cdot C$ base pair has three hydrogen bonds. These bonds are responsible for holding the two opposite strands together. Thus, if a DNA molecule contains many $G \cdot C$ base pairs, it is more stable than one containing many $A \cdot T$ base pairs. This also implies that DNA that is high in $G \cdot C$ content requires a higher temperatue

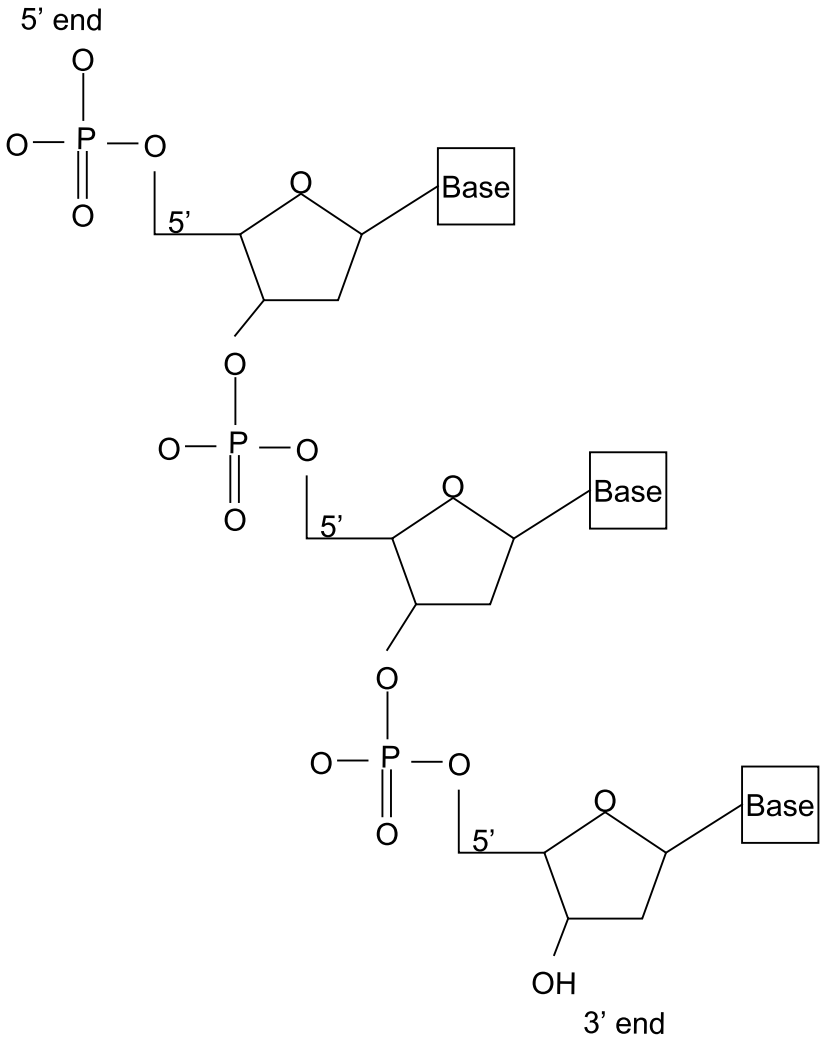

Figure 1.2. A diagram of one DNA strand showing the sugar-phosphate backbone with the phosphodiester bonds.

to separate, or *denature*, the two strands. Although the individual hydrogen bonds are rather weak, because the overall number of these bonds is quite high, the two strands are held together quite well.

Although in this book we will focus on gene expression, which involves transcription, it is important to say a few words about how the DNA molecule duplicates. Because the two strands in the DNA double helix are complementary, they carry the same information. During replication, the strands separate and each one acts as a *template* for directing the synthesis of a new complementary strand. The two new double-stranded molecules are passed on to daughter cells during

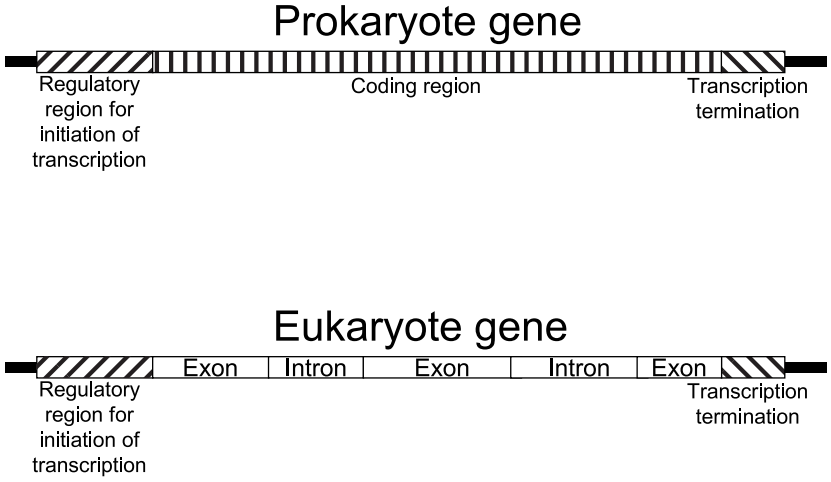

Figure 1.3. The gene structure in prokaryotes and eukaryotes.

cell division. The DNA replication phase in the cell cycle is called the *S* (synthesis) *phase*. After the strands separate, the single bases (on each side) become exposed. Thus, they are free to form base pairs with other free (complementary) nucleotides. The enzyme that is responsible for building new strands is called *DNA polymerase*.

1.1.2 Genes

Genes represent the functional regions of DNA in that they can be transcribed to produce RNA. A gene contains a *regulatory region* at its upstream (5′) end to which various proteins can bind and cause the gene to initiate transcription in the adjacent RNA-encoding region. This essentially allows the gene to receive and then respond to other signals from within or outside the genome. At the other (3′) end of the gene, there is another region that signals termination of transcription.

In eukaryotes (cells that have a nucleus), many genes contain *introns*, which are segments of DNA that have no information for, or do not code for, any gene products (proteins). Introns are transcribed along with the coding regions, which are called *exons*, but are then cut out from the transcript. The exons are then spliced together to form the functional messenger RNA that leaves the nucleus to direct protein synthesis in the cytoplasm. Prokaryotes (cells without a nucleus) do not have an exon/intron structure, and their coding region is contiguous. These concepts are illustrated in figure 1.3.

The parts of DNA that do not correspond to genes are of mostly unknown function. The amount of this type of *intergenic* DNA present depends on the organism. For example, mammals can contain enormous regions of intergenic DNA.

1.1.3 RNA

Before we go on to discuss the process of transcription, which is the synthesis of RNA, let us say a few words about RNA and its roles in the cell. As discussed above, most RNAs are used as an intermediary in producing proteins via the process of translation. However, some RNAs are also useful in their own right in that they can carry out a number of functions. As mentioned earlier, the RNA that is used to make proteins is called messenger RNA. The other RNAs that perform various functions are never translated; however, these RNAs are still encoded by some genes.

One such type of RNA is transfer RNA (tRNA), which transports amino acids to mRNA during translation. tRNAs are quite general in that they can transport amino acids to the mRNA corresponding to any gene. Another type of RNA is ribosomal RNA (rRNA), which along with different proteins comprises *ribosomes*. Ribosomes coordinate assembly of the amino acid chain in a protein. rRNAs are also general-purpose molecules and can be used to translate the mRNA of any gene. There are also a number of other types of RNA involved in splicing (snRNAs), protein trafficking (scRNAs), and other functions. We now turn to the topic of transcription.

1.1.4 Transcription

Transcription, which is the synthesis of RNA on a DNA template, is the first step in the process of gene expression, leading to the synthesis of a protein. Similarly to DNA replication, transcription relies on complementary base pairing. Transcription is catalyzed by an *RNA polymerase* and RNA synthesis always occurs from the $5'$ to the $3'$ end of an RNA molecule. First, the two DNA strands separate, with one of the strands acting as a template for synthesizing RNA. Which of the two strands is used as a template depends on the gene. After separation of the DNA strands, available ribonucleotides are attached to their complementary bases on the DNA template. Recall that in RNA uracil is used in place of thymine in complementary base pairing. The RNA strand is thus a direct copy (with U instead of T) of one of the DNA strands and is referred to as the *sense* strand. The other DNA strand, the one that is used as a template, is called the *antisense* strand. This is illustrated in figure 1.4.

Transcription is initiated when RNA polymerase binds to the double-stranded

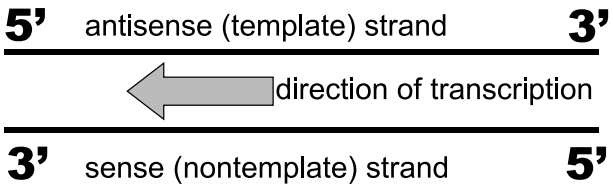

Figure 1.4. The direction of transcription. The antisense strand, oriented $3' \to 5'$, is used as a template to which ribonucleotides base-pair for synthesizing RNA, which always grows in the $5' \to 3'$ direction.

DNA. The actual site at which RNA polymerase binds is called a *promoter*, which is a sequence of DNA at the start of a gene. Since RNA is synthesized in the $5' \rightarrow 3'$ direction (figure 1.5), genes are also viewed in the same orientation by convention. Therefore, the promoter is always located upstream ($5'$side) of the coding region. Certain sequence elements of promoters are conserved between genes. RNA polymerase binds to these common parts of the sequences in order to initiate transcription of the gene. In most prokaryotes, the same RNA polymerase is able to transcribe all types of RNAs, whereas in eukaryotes, several different types of RNA polymerases are used depending on what kind of RNA is produced (mRNA, rRNA, tRNA).

In order to allow the DNA antisense strand to be used for base pairing, the DNA helix must first be locally unwound. This unwinding process starts at the promoter site to which RNA polymerase binds. The location at which the RNA strand begins to be synthesized is called the *initiation site* and is defined as position +1 in the gene. As shown in figure 1.5, RNA polymerase adds ribonucleotides to the $3'$ end of the RNA. After this, the helix is re-formed once again.

Since the transcript must eventually terminate, how does the RNA polymerase know when to stop synthesizing RNA? This is accomplished by the recognition of certain specific DNA sequences, called *terminators*, that signal the termination of transcription, causing the RNA polymerase to be released from the template and ending the RNA synthesis. Although there are several mechanisms for termination, a common direct mechanism in prokaryotes is a terminator sequence arranged in such a way that it contains self-complementary regions that can form *stem-loop* or *hairpin* structures in the RNA product. Such a structure, shown in figure 1.6, can cause the polymerase to pause, thereby terminating transcription. It is interesting to note that the hairpin structure is often GC-rich, making the self-complementary base pairing stronger because of the higher stability of $G \cdot C$ base pairs relative to $A \cdot U$ base pairs. Moreover, there are usually several U bases at the end of the hairpin structure, which, because of the relatively weaker $A \cdot U$ base pairs, facilitates dissociation of the RNA.

In eukaryotes, the transcriptional machinery is somewhat more complicated than in prokaryotes since the primary RNA transcript (pre-mRNA) must first

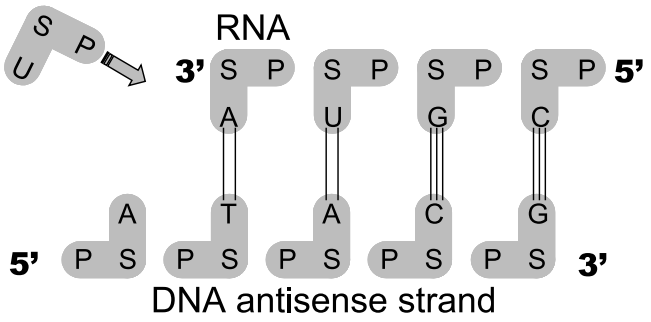

Figure 1.5. The addition of a uracil to the $3'$ end of the synthesized RNA. P stands for phosphate and S stands for sugar.

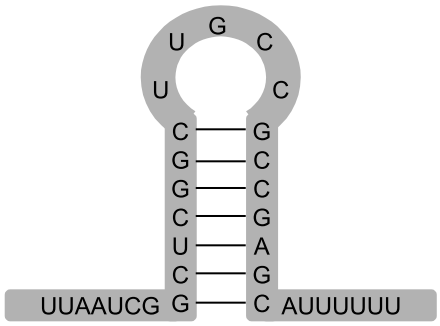

Figure 1.6. RNA hairpin structure used as a termination site for RNA polymerase in prokaryotes.

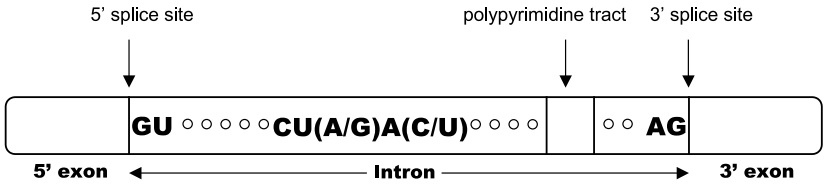

Figure 1.7. The splice sites, the branchpoint sequence, and the polypyrimidine tract.

be processed before being transported out of the nucleus (recall that prokaryotes have no nucleus). This processing first involves *capping*—the addition of a 7-methylguanosine molecule to the 5′ end of the transcript, linked by a triphosphate bond. This typically occurs before the RNA chain is 30 nucleotides long. This cap structure serves to stabilize the transcript but is also important for splicing and translation. At the 3′ end, a specific sequence (5′-AAUAAA-3′) is recognized by an enzyme which then cuts off the RNA at approximately 20 bases further down, and a *poly(A) tail* is added at the 3′ end. The poly(A) tail consists of a run of up to 250 adenine nucleotides and is believed to help in the translation of mRNA in the cytoplasm. This process is called 3′ *cleavage* and *polyadenylation*.

The final step in converting pre-mRNA into mature mRNA involves *splicing*, or removal of the introns and joining of the exons. In order for splicing to occur, certain nucleotide sequences must also be present. The 5′ end of an intron almost always contains a 5′-GU-3′ sequence, and the 3′ end contains a 5′-AG-3′ sequence. The AG sequence is preceded by a *polypyrimidine tract*—a pyrimidine-rich sequence. Further upstream there is a sequence called a *branchpoint sequence*, which is 5′-CU(A/G)A(C/U)-3′ in vertebrates. The splice sites as well as the conserved sequences related to intron splicing are shown in figure 1.7.

The actual splicing reaction consists of two steps. In the first step, the G at the 5′ splice site is attacked by the 2′-hydroxyl group in the adenine (fourth base) in the branchpoint sequence, which creates a circular molecule called a *lariat*, thereby freeing the exon upstream of the intron. In the second step, the 3′ splice site is cleaved, which releases the lariat and joins the two exons. The released intron

is then degraded. We should briefly mention at this point that in eukaryotes it is possible for a particular pre-mRNA to produce several types of mature mRNA by means of *alternative splicing*. In other words, certain exons can be removed by splicing and therefore are not retained in the mature mRNA product. These alternative mRNA forms ultimately give rise to different proteins. A good example of this is the fibronectin gene, which by means of alternative splicing can produce two different protein isoforms. One isoform of this protein is produced by fibroblasts (a connective tissue cell), whereas the other is secreted by hepatocytes (an epithelial liver cell). Two exons that are responsible for fibronectin adhering to cell surfaces are found in mRNA produced in fibroblasts, while in hepatocytes, they are spliced out. Consequently, the fibronectin produced by hepatocytes cannot adhere to cell surfaces and can easily be transported in the serum.

1.1.5 Proteins

A protein is a chain of linked *amino acids*. A total of 20 amino acids can occur, each having unique properties. In a protein, they are linked by covalent bonds called *peptide bonds*, which are formed by the removal of water molecules between the amino acids. Although the picture of amino acids linked in a chain suggests a simple linear arrangement, proteins are structurally quite complex. The linear sequence of amino acids constitutes the protein's *primary structure*. However, various forces acting between atoms at different locations within the protein cause it to take on a specific shape. The *secondary structure* of a protein refers to the regular arrangement of the amino acids within localized regions of the protein. The most common types of secondary structure are the α helix (a spiral structure) and the β sheet (a planar structure).

The term *tertiary structure* describes the three-dimensional arrangement of all amino acids. In most proteins, various combinations of α helices and β sheets can fold into compact globular structures called *domains*, which are the basic units of tertiary structure. It is often the case that amino acids that are far apart in the linear sequence are in close proximity in the tertiary structure. Different domains in large proteins are often associated with specific functions. For example, one domain might be responsible for catalytic activity, while another may modulate DNA-binding ability. Often, several identical or different types of folded (tertiary) structures bind together to form a *quaternary structure*.

What is important to remember is that the function of a protein is determined by its three-dimensional structure, which in turn is determined by the sequence of amino acids. Also important is the fact that most proteins are chemically modified after they are produced. This may have the effect of altering their life span, activity, or cellular location. Some modifications, such as those involving linkage of various chemical groups to the protein, are often reversible, whereas others, such as removal of entire peptide segments, are not.

Table 1.1 The 20 Amino Acids and Their Corresponding Codons

Amino Acid	**Codons**
Alanine	GCA, GCC, GCG, GCU
Arginine	AGA, AGG, CGA, CGC, CGG, CGU
Asparagine	AAC, AAU
Aspartic acid	GAC, GAU
Cysteine	UGC, UGU
Glutamic acid	GAA, GAG
Glutamine	CAA, CAG
Glycine	GGA, GGC, GGG, GGU
Histidine	CAC, CAU
Isoleucine	AUA, AUC, AUU
Leucine	UUA, UUG, CUA, CUC, CUG, CUU
Lysine	AAA, AAG
Methionine	AUG
Phenylalanine	UUC, UUU
Proline	CCA, CCC, CCG, CCU
Serine	AGC, AGU, UCA, UCC, UCG, UCU
Threonine	ACA, ACC, ACG, ACU
Tryptophan	UGG
Tyrosine	UAC, UAU
Valine	GUA, GUC, GUG, GUU
STOP	UAA, UAG, UGA

The last row shows the three STOP codons. As can be seen, 18 of the amino acids are specified by more than one codon. Codons that specify the same amino acid are said to be synonymous.

1.1.6 Translation

The sequence of amino acids in a protein is determined from the sequence of nucleotides in the gene that encodes the protein. Recall that the sequence of nucleotides in DNA is transcribed into an mRNA sequence. After this, ribosomes, which are large macromolecular assemblies, move along the mRNA in $5' \rightarrow 3'$ direction and read the mRNA sequence in nonoverlapping chunks of three nucleotides. Each nucleotide triplet, called a *triplet codon*, codes for a particular amino acid. As there are 4 different possible nucleotides, there are $4^3 = 64$ possible codons. Since there are only 20 possible amino acids and 64 possible codons, most amino acids can be specified by more than one triplet. The *genetic code* specifies the correspondence between the codons and the 20 amino acids, as shown in table 1.1.

Each codon is recognized by an *anticodon*—a complementary triplet located on the end of a transfer RNA molecule—having a cloverleaf shape. The anticodon binds with a codon by the specific RNA-RNA base complementarity. Note that since codons are read in the $5' \rightarrow 3'$ direction, anticodons are positioned on tRNAs in the $3' \rightarrow 5'$ direction. This is illustrated in figure 1.8.

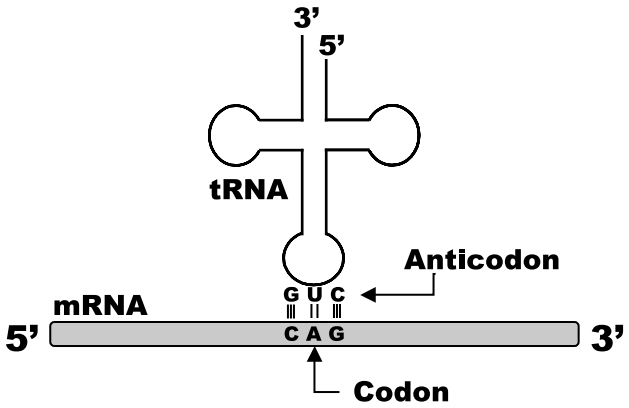

Figure 1.8. The anticodon at one end of a tRNA molecule binds to its complementary codon in mRNA.

Each tRNA molecule carries a specific amino acid attached to its free 3′ end. Moreover, each tRNA is designed to carry only one of the 20 possible amino acids. However, some amino acids can be carried by more than one type of tRNA molecule. Also, certain tRNA molecules require accurate base-pair matching at only the first two codon positions and can tolerate a mismatch, also called a *wobble*, at the third position. tRNAs are essentially "adaptor" molecules in the sense that they can convert a particular sequence of three nucleotides into a specific amino acid, as specified in table 1.1.

The tRNAs and mRNA meet at the ribosome, which contains specific sites at which it binds to mRNA, tRNAs, and other factors needed during protein synthesis. It is important that the *reading frame* progresses accurately, three nucleotides at a time, without skipping any nucleotides, for if this were to occur, the triplets would code for the wrong amino acids and in turn would synthesize the wrong protein, which might have highly adverse phenotypic effects. The ribosome ensures this precise movement.

The ribosome has an aminoacyl-tRNA-binding site (*A site*), which holds the incoming tRNA molecule that carries an amino acid, and a peptidyl-tRNA-binding site (*P site*), which holds the tRNA molecule linked to the growing polypeptide chain. These two sites are located quite close together so that the two tRNA molecules bind to two adjacent codons on the mRNA. After an aminoacyl-tRNA molecule binds to the A site by forming base pairs with the codon, the amino acid linked to the tRNA molecule positioned at the P site becomes uncoupled from its host tRNA molecule and forms a peptide bond to the new amino acid attached to the tRNA that just arrived at the A site. Then, the ribosome moves exactly three nucleotides along the mRNA, the tRNA molecule at the A site is moved to the P site, the old tRNA that used to be at the P site (without its amino acid) is ejected from the ribosome and reenters the tRNA pool in the cytoplasm. This leaves the A site free for the newly arriving aminoacyl-tRNA molecule. This three-step process, referred to as the *elongation* phase of protein synthesis, is illustrated in figure 1.9.

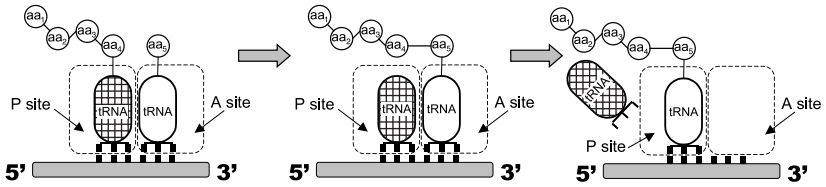

Figure 1.9. The three-step process of the elongation phase in protein synthesis.

The last row in table 1.1 does not contain an amino acid; rather, it shows the three possible codons (UAA, UAG, and UGA), called *stop codons*, that are used to terminate translation. If the ribosome is positioned in such a way that the stop codon is at the A site, certain proteins called *release factors* recognize and bind directly to this stop codon. By a series of steps, this terminates transcription and releases the polypeptide chain. The ribosome, having finished its job, also dissociates and is ready to reassemble on a new mRNA molecule to begin protein synthesis again.

We have now discussed how the ribosome accurately moves one codon at a time along the mRNA and how it terminates translation when it reaches a stop codon. But we have not yet discussed how the ribosome initiates the process of protein synthesis or, more specifically, how it knows where to begin. Indeed, depending on where the process starts, the subsequent nonoverlapping nucleotide triplets code for completely different polypeptide chains. Thus, there are three possible reading frames. The initiation process of protein synthesis consists of several rather complex steps catalyzed by proteins called *initiation factors* and involves the ribosome, the mRNA, initiator tRNA (which carries methionine), and guanosine 5′-triphosphate (GTP). In eukaryotes, the initiator tRNA binds to the triplet AUG, which codes for methionine, as table 1.1 shows. In *Escherichia coli*, for example, AUG or GUG can serve as the initiation codon. Since there could in principle be many AUGs, how is the correct initiation codon selected? In prokaryotes, there is a conserved purine-rich sequence 8–13 nucleotides in length, called the *Shine-Dalgarno sequence*, which is located upstream of the initiation codon. This sequence is believed to position the ribosome correctly by base-pairing with one of the subunits of the ribosome.

1.1.7 Transcriptional Regulation

All cells contain the same DNA content. What, then, differentiates a liver cell from a white blood cell? The properties of a cell, including its architecture and ability to participate in various activities while interacting with its environment, are determined by gene activities and, in particular, the expressed proteins. This implies that there must be a control mechanism, internal, external, or both, that regulates expression of the proteins characterizing the cell type or the functional state in which a particular cell is found. The process of gene regulation can be extremely complex, especially in higher eukaryotes.

As an example of why it is important for a cell to produce certain proteins in response to various environmental conditions, let us consider a bacterium. Sugar metabolism involves a number of enzymes. There are various sugars, such as lactose, glucose, and galactose, that can be used as an energy source. However, depending on the type of sugar, a different set of enzymes is needed to enable the sugar to enter the cell and then break down the sugar. Of course, one possibility would be to have all these enzymes available so that when a particular sugar is presented, the necessary enzyme would be readily available for it. Such a strategy, however, would be highly wasteful and inefficient, requiring too much energy for producing all the enzymes, many of which might never be needed. Thus, the strategy a bacterium uses is to synthesize only the needed enzymes by activating the genes that encode these enzymes and inactivating or repressing the genes that encode unnecessary enzymes. These genes can therefore be activated and inactivated in response to various environmental conditions.

We have already discussed one of the mechanisms necessary for transcription to occur: RNA polymerase must bind to the promoter of a gene. However, various other DNA-binding proteins can determine whether or not transcription can take place. In prokaryotes, for example, certain sites in the vicinity of the promoter can be bound by regulatory proteins called *activators* or *repressors*. An activator protein, as the name suggests, must bind to its target site in order for transcription to occur. A repressor protein, on the other hand, must not bind to its target site in order to enable transcription—if it binds, transcription is blocked. One of the mechanisms by which activators and repressors are able to alter the transcription of a gene involves physical interaction with RNA polymerase (bound to the nearby promoter); a bound repressor can physically interfere with RNA polymerase binding, whereas an activator can assist the RNA polymerase in attaching to the DNA.

If the binding of a regulatory protein can determine whether a gene transcript is produced, what determines this protein's ability to bind to its target site? In many cases, this ability is modulated by the interplay between the protein's DNA-binding domain (a region on the protein that can directly bind to specific DNA sequences) and the protein's *allosteric site* (a site on a protein where small molecules, called *allosteric effectors*, can bind and cause conformational changes). When an allosteric effector binds to the allosteric site on a regulatory protein, the structure of the DNA-binding domain can change, thereby enabling or disabling the protein's ability to bind to its target site and ultimately determining whether transcription can occur. For example, in an activator, the presence of an allosteric effector might enable the protein's ability to bind DNA. On the other hand, a repressor might be able to block transcription (by binding to its target site) only in the absense of the effector molecule. For instance, in lactose metabolism in *E. coli*, the *lac* repressor loses its affinity for its target site and falls off the DNA when lactose (its allosteric effector) binds to it. This in turn allows the RNA polymerase situated nearby to do its job and transcribe the nearby genes encoding certain enzymes needed for the metabolism of lactose (Figure 1.10).

The *lac* repressor itself is encoded by a regulatory gene (*lac*I). Thus, this is a good example of how the product of one gene is able to control the transcription

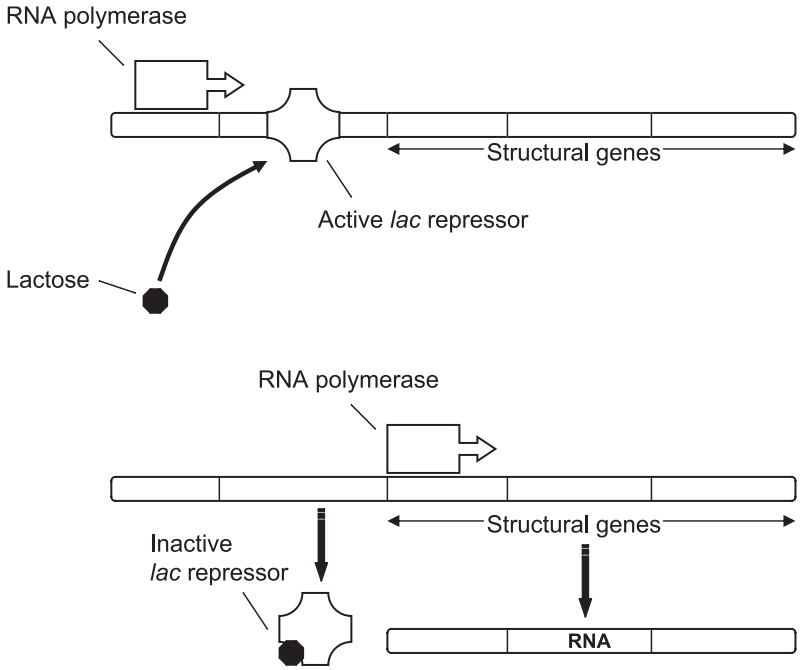

Figure 1.10. Lactose serving as an allosteric effector of the *lac* repressor protein. The repressor binds to the DNA, blocking RNA polymerase and inhibiting transcription of the nearby structural genes encoding enzymes. When lactose is introduced, it binds to the *lac* repressor and causes a change in its conformation, causing it to fall off the DNA and thereby freeing the RNA polymerase to quickly initiate transcription.

of other genes. However, such control mechanisms can be much more complicated even in prokaryotes, involving a number of other genes and factors in a multivariate fashion—a topic that we will deal with later in the book when we study models of genetic networks. As an example of an additional control mechanism, let us again consider the lactose system in *E. coli.* It is known that the cell is able to capture more energy from the breakdown of glucose than of lactose. Thus, if both lactose and glucose are present, the cell will favor glucose. Consequently, the cell will not require the production of *lac* enzymes to metabolize lactose until it runs out of glucose. This implies that there is an additional level of control due to the concentration of glucose.

Indeed, when glucose concentration is high, a substance called cyclic adenosine monophosphate (cAMP) is present at low concentrations. However, when glucose begins to run out, the level of cAMP increases, serving as an alert signal for the cell. cAMP then forms a complex with the cAMP receptor protein (also called catabolite activator protein or CAP), and this complex binds to its target site just upstream of the RNA polymerase site. Without cAMP, CAP cannot bind to its

target site. Thus, cAMP is an allosteric effector, and CAP is an activator protein. After the cAMP-CAP complex binds, it induces a bend in the DNA, and this is believed to enhance the affinity of RNA polymerase for the *lac* promoter. Putting all this together, we see that low levels of glucose induce the transcription of genes encoding the enzymes necessary for the metabolism of lactose. We can conclude that the presence of lactose can activate these genes, but this is not a sufficient condition; glucose must also not be present.

The regulation of transcription in eukaryotes is fundamentally similar to that in prokaryotes. The regulatory proteins, often called *trans-acting elements*, bind to specific sequences in DNA, which are often called *cis-acting elements*. Unlike the situation in prokaryotes, however, cis-acting elements can be located at great distances from the transcription start sites. There are also several classes of cis-acting elements.

The RNA polymerase–binding region constitutes the core promoter. This region usually contains a *TATA box*—a highly conserved sequence approximately 30 base pairs (bp) upstream of the transcription initiation site. In some eukaryotes, instead of a TATA box, there is a promoter element called an initiator. Nearby are the promoter-proximal cis-acting sequences; as in prokaryotes, the proteins that bind to these sites can physically interact with RNA polymerase and help it bind to the promoter. Promoter-proximal elements are typically within 100–200 bp of the transcription initiation site. These elements also contain conserved sequences, such as CCAAT. Finally, there can be cis-acting elements located very far away from the promoter. These fall into two classes: enhancers and silencers. *Enhancers* increase transcription rates, whereas *silencers* inhibit transcription. Enhancers and silencers can act as far as thousands of base pairs away from the start site. They can also be located upstream or downstream of the promoter. It is also interesting that many enhancers are cell-type-specific, as are some promoter-proximal elements. This means that they are able to exert their effect only in particular differentiated cell types. The mechanism by which these distant cis-acting elements can execute control over transcription is believed to involve DNA loops, whereby the trans-acting regulatory protein bound at a distant enhancer site is brought into close proximity with other proteins in the promoter-proximal cis-acting elements. This can also explain why enhancers and silencers can be either upstream or downstream of the promoter. Figure 1.11 illustrates this mechanism.

Finally, we should mention that regulation by feedback is common and often necessary for proper cell functioning. For example, positive feedback refers to the situation where a regulatory protein activates its own transcription. More complicated feedback mechanisms are possible, such as when protein A regulates the transcription of gene B whose gene product regulates the transcription of gene A. Negative feedback is also common. Simple positive self-feedback can allow a gene to be switched on continuously. For example, the *myoD* gene, which begins to be expressed when a cell differentiates into a muscle cell, is kept continually expressed by means of its own protein binding to its own cis-acting elements. Thus, the cell continues to synthesize this muscle-specific protein and remains differentiated, passing on the protein to its daughter cells.

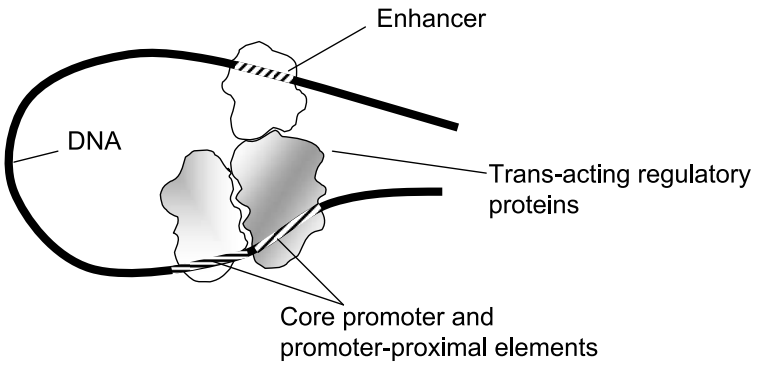

Figure 1.11. How a DNA loop can bring a trans-acting regulatory protein bound at a distant enhancer site into close proximity with other regulatory proteins bound at the core promoter and promoter-proximal elements.

1.2 GENOMICS

The field of genomics is concerned with investigating the properties and behavior of large numbers of genes, typically in a high-throughput manner. The overarching goal of this enterprise is to understand how the genome functions as a whole. That is, how do genes and proteins work together, regulating the functions of cells and the development of organisms, and what goes wrong when they do not work together as they should in diseases such as cancer?

Genomic studies can be conducted on different levels roughly corresponding to DNA, RNA, and protein. On the DNA level, most genomic studies have addressed the problem of gene mapping and sequencing for entire genomes, such as in the Human Genome Project. A popular method, called comparative genomic hybridization (CGH), can be used to produce a map of DNA sequence copy numbers as a function of chromosomal location throughout the entire genome. Thus, it can be used to perform genomewide scanning of differences in DNA sequence copy numbers with many applications in cancer research. High-throughput CGH methods using array technology are now available.

Several recently developed technologies now allow us to study the *transcriptome*—the aggregation of mRNAs present in a cell at the time of measurement—in a high-throughput manner. This involves identifying the mRNAs and quantifying their abundances or, in other words, measuring gene expressions on the transcript level. First, a few words about cDNA libraries.

A direct way to determine what transcripts are produced would be to collect the mRNA produced in cells and create from it, a *complementary DNA* (cDNA) library that can be subsequently sequenced. Recall that reverse transcriptase is the enzyme that uses RNA as a template to produce cDNA. The reason for constructing such a cDNA library is that most of the DNA in the genome (introns, promoters, intergenic regions) is not needed for determining gene expressions. Thus, the cDNAs contain no intron sequences and cannot be used to express the protein encoded by

the corresponding mRNA.

After reverse transcription of the mRNA, the synthesized cDNA is introduced into bacteria. This is accomplished by using *cloning vectors*: a DNA fragment is joined to, say, a chromosome of a bacterial virus (phage), and the whole recombinant DNA molecule is introduced into bacteria for replication (recall that a virus also uses the host cell for its own replication). In addition to phage vectors, plasmid vectors—small, circular molecules of double-stranded DNA derived from naturally occurring bacterial cells—are often used. In either case, the bacterial host cells are then *transfected* with the vectors, which means that the recombinant DNA molecules are inserted into the cells, which then keep dividing and produce more and more of the foreign DNA. Cloning vectors typically incorporate a gene that allows those cells that contain the vector to be selected from those that do not contain it. This selectivity gene usually provides resistance to some toxin, such as an antibiotic. Thus, only the cells with inserted vectors ultimately survive. Each bacterial colony in a culture dish houses a unique vector that contains the cDNA counterpart of a particular mRNA. Today, large sequenced cDNA clone libraries are commercially available.

Although in principle every clone in a cDNA library can be sequenced to study the transcriptome of a given cell system, this approach requires a large amount of work and is not practical. One high-throughput approach for studying the transcriptome is called serial analysis of gene expression or SAGE. This method is based on the fact that a short (e.g., 12-bp) nucleotide sequence, called a *SAGE tag*, can uniquely identify a gene as long as the sequence is always from a certain position in the transcript. Indeed, 12 bp can, in theory, distinguish up to $4^{12} = 16\,777\,216$ different transcripts. The abundance of each SAGE tag is then quantified and used as a measurement of the level of expression of the corresponding gene. Many SAGE tags can be concatenated (covalently linked together within a single clone) head to tail to produce a single *concatamer*, which can then be sequenced. This step reduces the number of sequencing reactions that need to be carried out; information on 30 or more genes can be obtained with one sequencing reaction.

1.2.1 Microarray Technology

Another extremely popular high-throughput technique for studying transcriptomes is the DNA *microarray* (DNA *chip*). For completely sequenced genomes, DNA microarrays allow the screening of whole-genome expression, meaning that every gene in the organism can be simultaneously monitored. This is already possible for the human genome. Although there are several kinds of microarrays reflecting the method of their construction, such as in situ synthesized oligonucleotide arrays and spotted cDNA arrays, on a basic level a microarray is a gridlike arrangement of thousands of unique DNA molecules, called *probes*, attached to some support surface such as glass or a nylon membrane. We will not attempt to cover in detail different microarray technologies and all aspects of microarray production. Instead, we refer the reader to several books, such as Schena (2003), Blalock (2003), Hardiman (2003), and Zhang et al. (2004). In this section, we will outline the generic principles of microarrays and use the popular spotted cDNA or long oligo

arrays as our example.

The DNA fragments are arrayed (spotted) on a simple glass microscope slide by a robot. The robot typically dips its pins into solutions, typically placed in 96-well plates, that contain the DNA. Then the tiny amounts of solution that adhere to the pins are transferred to a support surface, most commonly glass. Each such transfer by a pin results in a tiny printed spot containing DNA molecules, and the printed array consists of a gridlike arrangement of such spots. The pins can be either solid, similar to a sharp needle, or quill-type (also called split-pin), which have narrow slits as in a quill pen. In the former case, the robot must redip the pins into the solution before touching down on the support surface, whereas in the latter case, the robot can perform multiple printings without redipping. As a result, solid-pin robots are slower, but quill-pin robots are more prone to pin damage and clogging, resulting in printing dropouts. Yet another robotic printing approach is the inkjet method, which is the same method used in standard printers, whereby a precise amount of sample is expelled from a miniature nozzle equipped with a piezoelectric fitting by applying an electric current. Robots are now capable of printing upward of 50,000 spots on a standard microscope glass slide. The slides are often precoated with polylysine or polyamine in order to immobilize the DNA molecules on the surface. The DNA is denatured so that each spot contains deposits of single-stranded DNA.

In spotted microarrays, each spot contains a different DNA sequence (probe) corresponding to the spliced mRNA (exons) of a specific gene (in the case of eukaryotic genes). Thus, different spots can serve as detectors of the presence of mRNAs that were transcribed from different genes in the sample of interest. This is accomplished as follows. The RNA is first extracted from the cells whose transcriptome we wish to study. Then, the RNA is converted to cDNA and amplified by the reverse transcriptase–polymerase chain reaction (rtPCR). During this process, fluorescent molecules (tags) are chemically attached to the DNA. If a specific mRNA molecule was produced by the cells in the sample, then the corresponding fluorescently labeled cDNA molecule will eventually stick to its complementary single-stranded probe via base pairing. The rest of the cDNA molecules that are unable to find their complementary "partners" on the array will be washed away during a washing step. Since the fluorescent tags will still be attached to the molecules that base-pair with the probes, the corresponding spots will fluoresce when measured by an instrument that provides fluorescence excitation energy and detects the level of emitted light, yielding a digital image of the microarray in which the level of measured fluorescence is converted to pixel intensities. After an image processing step intended to segment the spots and summarize the overall pixel intensity in each spot, a data file containing measurements of gene expressions (and a host of other measurements, such as background intensity, spot area, signal-to-noise ratio, and so on) is produced. These data are then subjected to statistical analysis and modeling.

In the case of cDNA microarrays, two distinct fluorescent dyes are commonly used to compare relative gene expressions between two samples. In this scenario, one dye is incorporated into the cDNA of one sample (say, tumor) and the cDNA from the other sample (say, normal) is tagged with the other dye. Commercial scanners use multiple laser sources to excite each of the dyes at their specific

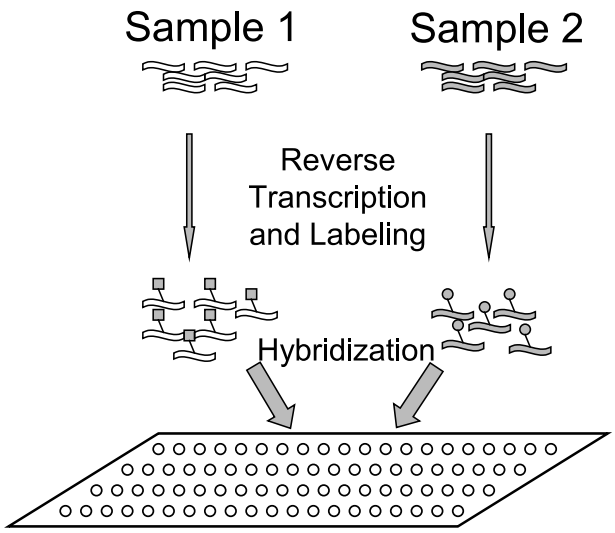

Figure 1.12. Two-color hybridization. The RNA extracted from each sample is reverse-transcribed and labeled to make cDNA so that each sample is labeled with a different fluorescent dye. The labeled cDNAs are then hybridized to the probes on the microarray so that the fluorescent signal intensities corresponding to the abundances of mRNA transcripts produced in each of the samples can be quantified by a laser scanner and image processing.

wavelengths so that their fluorescent emissions can be detected. This "two-color" microarray approach allows one to directly compare relative transcript abundance between two samples and provides a type of self-normalization in that possible local variations in the array affecting only some of the spots affect both samples equally. The two-color approach is illustrated in figure 1.12.

Many of the current approaches in computational and systems biology that we will discuss in this book make use of microarray data. Although this technology is undergoing rapid change, an understanding of the fundamental principles behind microarray experiments is essential for investigators wishing to apply computational, mathematical, and statistical methods to microarray data. Each of the multitude of steps, including biological sample handling, slide preparation and printing, labeling, hybridization, scanning, and image analysis, presents its own challenges and can greatly influence the results of the experiment. Without an appreciation of the ways each of these steps can affect the data produced in a microarray experiment, computational and modeling efforts run the risk of yielding unsound conclusions. We strongly recommend that the interested reader, especially one who intends to work with microarray data, consult the available literature on microarray experiments, quality control, and basic data analysis (e.g., Schena, 2003; Baldi and Hatfield, 2002; Zhang et al., 2004).

1.3 PROTEOMICS

As we have discussed, genomics is concerned with studying large numbers of genes and, in particular, how they function collectively. However, genes execute their functions via the transcriptome and the *proteome*, the latter pertaining to the collection of functioning proteins synthesized in the cell. Thus, a comprehensive understanding of how the genome controls the development and functioning of living cells and how they fail in disease requires the study of transcriptomics and proteomics. While these disciplines specifically refer to the study of RNA and proteins, respectively, we take the broader view that genomics must encompass all these levels precisely because the genetic information stored in DNA is manifested on the RNA and protein levels. Indeed, transcriptional profiling using a high-throughput technology such as microarrays is already commonly considered part of genomics. Thus, many of the methods and models that we describe in this book are applicable not only to transcriptional but also to protein expression measurements.

Transcriptional measurements can show us which genes are expressed in a cell at a given time or under a given condition, but they cannot give us accurate information about the synthesized proteins despite the fact that RNA is translated into proteins. One reason for this is that different mRNAs can be translated at different times and a transcriptional measurement at a given time may not reflect the protein content. Furthermore, while some proteins are being synthesized, others are being degraded in a dynamical process. Thus, direct measurements of protein activity are also necessary.

Two common techniques are two-dimensional gel electrophoresis and mass spectrometry. The first method is used for separating the individual proteins in the proteome. It is based on standard polyacrylamide gel electrophoresis, which is used to separate proteins according to their molecular weights but also involves a second step where the gel is rotated 90° and the proteins are then separated according to their charges. This procedure results in a two-dimensional arrangement of spots, each one corresponding to a different protein. This method can be used to detect differences between two proteomes by comparing the presence or intensity of spots in corresponding locations. Another application of this procedure in proteomics is the isolation of proteins for further characterization by mass spectroscopy.

The protein in a given spot can be further analyzed and identified by mass spectrometry, which is a technique used to separate ions according to their charge-to-mass ratios. A common technique used in proteomics is matrix-assisted laser desorption ionization time of flight (MALDI-TOF). In this technique, protein ions are accelerated through an electric field. The smallest ones arrive at the detector first as they travel through the flight tube. This time of flight in the electric field is used as a measure of the charge-to-mass ratio. Mass spectrometry has greatly enhanced the usefulness of two-dimensional gels.

Another important goal in proteomics is the identification of interactions between two or more proteins. This information can be highly useful in characterizing the function of a new or uncharacterized protein. For example, if the unknown protein interacts with another protein known to be located on the cell surface, this

may indicate that the unknown protein is involved in cell-cell signaling. There are several popular methods, such as phage display and the yeast two-hybrid system, that can be used in a high-throughput fashion to detect protein-protein interactions. These techniques can generate protein interaction maps, which show all the interactions between the proteins. Let us give a brief outline of the yeast two-hybrid method.

This technique is based on the ability of two physically bound proteins to activate transcription of a reporter gene. The reporter gene, integrated into the yeast genome, is expressed only when a suitable transcription factor is present. Key to this method is the fact that transcription factors require two separate domains, a DNA-binding domain and an activation domain. Although these two domains are usually part of the same molecule, an artificial transcription factor can be constructed from two proteins, each carrying one of the two domains. First, the gene encoding the target protein is cloned in a vector that contains the cDNA sequence for the binding domain. Then, a library of uncharacterized cDNA sequences, each one cloned in a vector containing the cDNA sequence for the activation domain, is created. Thus, the target protein is fused to the binding domain, and each of the uncharacterized proteins is fused to the activation domain. These are all mixed together and cotransfected into a yeast strain containing an integrated reporter gene. If a protein-protein interaction occurs between the target protein and one of the uncharacterized proteins, the corresponding fused binding domain and activation domain are brought together, forming a transcription factor that can activate the reporter gene. The cells where the reporter is expressed can then be selected in order to identify the specific protein-protein interaction that took place.

Bibliography

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. (2002) *Molecular Biology of the Cell*, 4th ed., Garland Science, New York.

Baldi P, Hatfield GW. (2002) *DNA Microarrays and Gene Expression*, Cambridge University Press, Cambridge, UK.

Blalock EM, ed. (2003) *A Beginner's Guide to Microarrays*, Kluwer Academic Publishers, Norwell, MA.

Brown TA. (2002) *Genomes*, 2nd ed., John Wiley & Sons, New York.

Draghici S. (2003) *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, Boca Raton, FL.

Hardiman G, ed. (2003) *Microarrays Methods and Applications: Nuts & Bolts*, DNA Press, Skippack, PA.

Kohane IS, Kho A, Butte AJ. (2003) *Microarrays for an Integrative Genomics*, MIT Press, Cambridge, MA.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell JE. (2000) *Molecular Cell Biology*, 4th ed., W. H. Freeman & Co., New York.

Schena M. (2003) *Microarray Analysis*, John Wiley & Sons, Hoboken, NJ.

Zhang W, Shmulevich I, Astola J. (2004) *Microarray Quality Control*, John Wiley & Sons, Hoboken, NJ.