CHAPTER 1

MOLECULAR BIOLOGY FOR THE PRACTICAL BIOINFORMATICIAN

See-Kiong Ng

Institute for Infocomm Research
skng@i2r.a-star.edu.sg

Bioinformatics is a marriage of computer science with molecular biology. A practical bioinformatician must learn to speak the language of biology to enable fruitful cross-fertilization. However, the complexity of biological knowledge is daunting and the technical vocabulary that describes it is ever-expanding. In order to sift out the core information necessary for formulating a solution, it can be difficult for a non-biologically trained bioinformatician not to be lost in the labyrinths of confounding details. The purpose of this introductory chapter is therefore to provide an overview of the major foundations of modern molecular biology, so that a non-biologically trained bioinformatician can begin to appreciate the various intriguing problems and solutions described in the subsequent chapters of this book.

ORGANIZATION.

Section 1. We begin with a brief history of the major events in modern molecular biology, motivating the marriage of biology and computer science into bioinformatics.

Section 2. Then we describe the various biological parts that make up our body.

Section 3. Next we describe the various biological processes that occurs in our body.

Section 4. Finally, we describe the various biotechnological tools that have been developed by scientists for further examination of our molecular selves and our biological machineries.

1. Introduction

In the 1930s—years before the first commercial computer was created—a brilliant mathematician named Alan Turing conceived a theoretical computing machine that essentially encapsulated the essence of digital computing. Turing's sublime creation was as simple as it was elegant. A "Turing machine" consists of three key components—a long re-writable tape divided into single-digit cells each inscribed with either a 0 or a 1, a read/write head scanning the tape cells, and a table of simple instructions directing it, such as "`if in State 1 and scanning a 0: print 1, move right, and go into State 3`". This deceptively simple concoction of Turing has since been proven to be able to compute anything that a modern digital computer can compute.

About a decade and a half later in the 1950s, James Watson and Francis Crick deduced the structure of the DNA.[875] Their revelation also unveiled the uncanny parallel between Turing's machines and Nature's own biological machinery of life. With few exceptions, each of our cells in our body is a biochemical Turing machine. Residing in the brain of each cell is a master table of genetic instructions encoded in three billion DNA letters of A's, C's, G's, and T's written on about six feet of tightly-packed DNA, just like Turing's tape. This master table of genetic instructions—also known as the "genome"—contains all the instructions for everything our cells do from conception until death. These genetic instructions on the DNA are scanned by the ribosome molecules in the cells. Just like Turing's read/write heads, the ribosome molecules methodically decipher the encoded instructions on the tape to create the various proteins necessary to sustain life.

While the computer's alphabet contains only 1 and 0, the genetic instructions on the DNA are also encoded with a very simple set of alphabet containing only four letters—A, C, G, and T. Reading the genetic sequence on the DNA is like reading the binary code of a computer program, where all we see are seemingly random sequences of 1's and 0's. Yet when put together, these seemingly meaningless sequences encode the instructions to perform such complex feats as compute complicated arithmetic on the computer, or perform sophisticated functions of life in the cells. The DNA is indeed our book of life; the secrets of how life functions are inscribed amongst its many pages. This book of life was previously only readable by the ribosome molecules in our cells. Thanks to the tireless efforts of international scientists in the Human Genome Project, the book of life is now completely transcribed from the DNA into digitally-readable files.[467, 859] Each and everyone of us can now browse our book of life just like the ribosome molecules in the cells. Figure 1 outlines the various significant events leading up to the so-called "genome era" of mapping and sequencing the human genome.

While the book of life reads like gibberish to most of us, therein lies the many secrets of life. Unraveling these secrets requires the decryption of the 3 billion letter-long sequence of A's, C's, G's, T's. This has led to the necessary marriage of biology and computer science into a new discipline known as "bioinformatics". However, the complexity of biological knowledge is daunting and the technical vocabulary that describes it is ever-increasing. It can be difficult for a non-biology trained bioinformatician to learn to speak the language of biology. The purpose of

Year	Event
1865	Gregor Mendel discovered genes
1869	DNA was discovered
1944	Avery and McCarty demonstrated that DNA is the major carrier of genetic information
1953	James Watson and Francis Crick deduced the three-dimensional structure of DNA[875]
1960	Elucidation of the genetic code, mapping DNA to peptides (proteins)
1970's	Development of DNA sequencing techniques
1985	Development of polymerase chain reaction (PCR) for amplifying DNA[585, 586]
1980-1990	Complete sequencing of the genomes of various organisms
1989	Launch of the Human Genome Project
2001	The first working draft of the human genome was published[467, 859]
2003	The reference sequence of the human genome was completed

Fig. 1. Time line for major events in modern molecular biology leading to the genome era.

this chapter is to give an introduction to the major foundations of modern molecular biology. We hope to provide sufficient biological information so that a practical bioinformatician can appreciate various problems and solutions described in subsequent chapters of this book and—ultimately—the chapters in the book of life.

The rest of this chapter is divided into three main sections. In Section 2: *Our Molecular Selves*, we describe various biological parts that make up our body. In Section 3: *Our Biological Machineries*, we describe various biological processes that occurs in our body. Finally, in Section 4: *Tools of the Trade*, we describe various biotechnological tools that have been developed by scientists for examination of our molecular selves and our biological machineries.

2. Our Molecular Selves

No matter how complex we are, life begins with a single cell—the "zygote"—a cell resulting from the fusion of a sperm from a male and an ovum from a female. This cell must contain all the programmatic instructions for it to develop into a complex multi-cellular organism in due course. In this section, we describe how nature organizes such genetic information at the molecular level. We also study how this complex information is permuted for diversity and passed on from one generation to another.

2.1. *Cells, DNAs, and Chromosomes*

Despite the outward diversity that we observe, all living organisms are strikingly similar at the cellular and molecular levels. We are all built from basic units called "cells". Each cell is a complex automaton capable of generating new cells which are self-sustaining and self-replicating. Figure 2 shows the structure of the cell together with its genetic components.

The "brain" of each cell is a set of DNA molecules that encode the requisite genetic information for life, usually found in the nucleus of the cell. The name DNA is an acronym for "deoxyribonucleic acid". The DNA molecules of all organisms—animals and plants alike—are chemically and physically the same. A DNA molecule is made up of four types of base molecules or "nucleotides". Each nucleotide comprises a phosphate group, a sugar (deoxyribose), and a base—either an adenine (A), a cytosine (C), a guanine (G), or a thymine (T). As such, we refer to the nucleotides by their bases, and we represent the content of a DNA molecule as a sequence of A's, C's, G's, and T's.

Each DNA in the cell is a highly compressed macromolecule. It comprises two intertwining chains of millions of nucleotides in a regular structure commonly known as the "double helix", first described by Watson and Crick in 1953.[875] The two strands of the double helix are held together by hydrogen bonds between specific pairings of the nucleotides: an A on one strand is always paired to a T on the other strand, and a G paired to a C. This is why the term "base pairs" is often also used to refer to the nucleotides on the double helix.

As a result of the specific pairings of the nucleotides, knowing the sequence of one DNA strand implies the sequence of the other. In other words, each strand contains all the information necessary to reconstruct the complementary strand. In fact, this is the basis for DNA replication in the cell as well as in the laboratories, where single strands of complementary DNA are repeatedly used as templates for DNA replication; see polymerase chain reaction or PCR in Section 4.1.2 for example. In practice, we therefore refer to the genetic recipe on the DNA as a single string containing letters from A, C, G, and T. This means that not only is the sequence of the DNA important, but the direction is also important. With a strand of DNA, one end is called the 5'-end and the other end the 3'-end. A DNA sequence—for example, 5'-GATCATTGGC-3'—is always written in a left-to-right fashion with the "upstream" or 5'-end to the left and the 3'-end to the right.

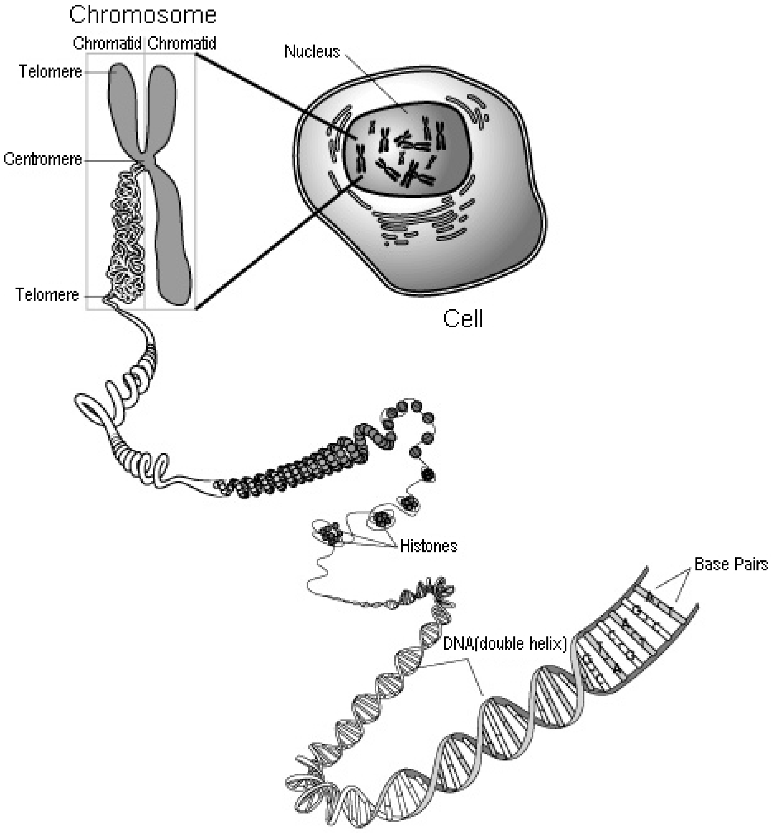

Fig. 2. The cell, chromosome, and DNA. (*Image credit: National Human Genome Research Institute.*)

This is also how it is scanned by the ribosomes in our cells; see Section 3.1 for more details. The corresponding complementary or “antisense” strand—namely, 3’-CTAGTAACCG-5’ for the example DNA sequence—can then serve as the template for replicating the original DNA sequence using the nucleotide pairing rules.

Typically, the DNA in a cell is arranged in not one but several physically separate molecules called the “chromosomes”. This particular arrangement forms—in effect—a distributed DNA database for the genetic information in the cells. While different species may have different number of chromosomes, the specific arrangement amongst all members in the same species is always consistent. Any

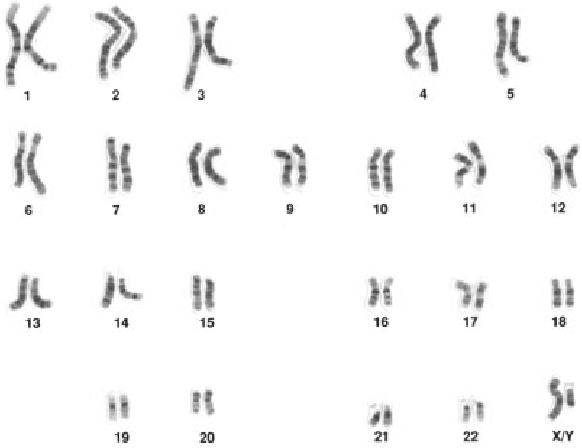

Fig. 3. 23 pairs of human chromosomes. (*Image credit: National Human Genome Research Institute.*)

aberration from the default chromosomal arrangement is often lethal or lead to serious genetic disorders. A well-known chromosomal disease in humans is the Down's Syndrome, in which an extra copy of one of the chromosomes causes mental retardation and physical deformation.

The chromosomes are usually organized in homologous (matching) pairs—each chromosome pair containing one chromosome from each parent. In humans, there are 23 pairs of homologous chromosomes ranging in length from about 50 million to 250 million base pairs. The human chromosomes are numbered from 1 to 22, with X/Y being the sex chromosomes. Figure 3 shows a photograph of how the 23 pairs of human chromosomes look like under a microscope. Collectively, the genetic information in the chromosomes are called the "genome". As each cell divides, the entire genome in the DNA is copied exactly into the new cells. This mechanism of copying the entire genetic blueprint in each and every cell is rather remarkable considering that the human body contains approximately 100 trillion cells. In theory, any of these 100 trillion cell on our body possesses the full complement of the genetic instructions for building a complex living organism like ourselves from it—if we can decipher the many secrets within the pages of our book of life in the DNA.

2.2. *Genes and Genetic Variations*

In 1865, an Augustinian monk in Moravia named Gregor Mendel published a then-obscure paper, “Versuche über Pflanzen-Hybriden” (Experiments in Plant Hybridization), describing his experiments on peas in the monastery garden. His paper was rediscovered *post mortem* 35 years later, and Mendel is now accredited as the Father of Genetics. In his legendary experiments, Mendel mated pea plants with different pairs of traits—round vs. wrinkled seeds, tall or dwarf plants, white or purple flowers, *etc*.—and observed the characteristics of the resulting offsprings. His results defied the then popular thinking which theorized that a tall plant and a short plant would have medium offspring. Based on his experiments, Mendel was able to theorize that the offspring must have received two particles—now called genes—one from each parent. One gene was dominant, the other recessive. His novel concept of genes explained why instead of medium plants, a tall and a short plant would produce some tall plants and some short plants.

Today, we know that Mendel’s genes are indeed the basic physical and functional units of heredity. The human genome, for example, is estimated to comprise more than 30,000 genes. Biochemically, genes are specific sequences of bases that encode the recipes on how to make different proteins, which in turn determine the expression of physical traits such as hair color or increased susceptibility to heart diseases. We describe how this works in Section 3.

Genes are linearly ordered along each chromosome like beads on a necklace. However, there are large amounts of non-gene DNA sequences interspersed between the genes. In fact, it is estimated that genes comprise only about 2% of the human genome. The remainder 98% consists of non-coding regions, whose specific biological functions are still unknown. As mentioned earlier, genes are like recipes; cells follow these recipes to make the right proteins for each specific trait or “phenotype”—a protein for red hair cells, for example. The actual versions—or “alleles”—of the genetic recipes that each of us have may differ from one another. *E.g.*, some of us may have inherited the recipe for making red hair cells, while others may have the blonde hair cell recipe at the gene that determines the hair color phenotype. The actual versions of genetic recipes that each of us have is called the “genotype”. With some exceptions of genes on the sex chromosomes, each person’s genotype for a gene comprises two alleles inherited from each of the parents. Cells can follow the recipe of either allele—a dominant allele always overpowers a recessive gene to express its trait; a recessive gene remains unseen unless in the presence of another recessive gene for that trait. Many of us may actually carry a disease gene allele although we are perfectly healthy—the disease gene allele is probably a recessive one. In this way, even fatal disease genes can “survive” in

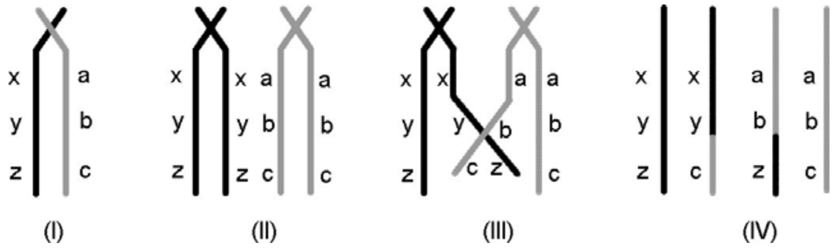

Fig. 4. Crossing-over during meiosis.

the gene pool unchecked as it is passed on from generations to generations until it is eventually paired up with another copy of the disease allele. Furthermore, common genetic diseases as well as many other traits are often "polygenic", determined by multiple genes acting together. Together with the dominant-recessive control mechanism, many important gene alleles may remain dormant until all the key factors come together through serendipitous genetic shufflings.

A key mechanism of genetic shufflings occur during reproduction, when genetic information are transmitted from parents to their offsprings. Instead of simply passing on one of the chromosomes in the chromosome pairs in each parent to the offspring intact, the parental chromosome pairs "cross over" during a special cell division process called "meiosis". Meiosis is the cell division process for generating the sexual cells which contains only half the genetic material of the normal cells in the parents. Before the cells split into halves to form egg or sperm cells, crossing over occurs to allow interchanging of homologous DNA segments in the chromosomal pairs in each parent.

Figure 4 shows a simplified example on how such DNA recombination during meiosis can give rise to further genetic variations in the offsprings. For illustration, we focus on only three gene loci on a chromosome pair in one of the parents here—say, the biological father. Let us supposed that the father has the genotypes `(x, a)`, `(y, b)`, and `(z, c)` for the three genes, and the alleles are arranged on the two chromosomes as shown in Part I of the figure. The order of the alleles on a chromosome is also known as the "haplotypes"—the two haplotypes on the father here are `(x, y, z)` and `(a, b, c)`. As we shall see shortly, meiotic cross-over events ensures diversity in the haplotypes (and hence genotypes) being passed on to the offsprings from each parent.

To prepare for DNA recombination by crossing over, the chromosome pairs double during the first phase of meiosis, resulting in two copies of each chromosome as shown in Part II of the figure. The exact copies are paired together, and the two pairs of homologous chromosomes line up side by side. Then, crossing-over takes place between the chromosomes, allowing recombination of the DNA. Part III of the figure depicts a cross-over between two of the four homologous chromosomes. As a result, four homologous chromosomes are generated, each with a different allelic set or haplotype for the three genes, as shown in Part IV. Each of these four chromosomes then goes into a different sperm cell, and an offspring may inherit any one of these four chromosome from the father. A similar process occurs in the mother in the formation of egg cells, so that the offspring inherits a randomly juxtaposed genotype from each of the parents. This explains why children tend to look like their parents, but not exactly and certainly not an absolutely equal 50/50 mix. They usually inherit some distinct characteristics from one parent, some from the other and some from their grandparents and great grandparents. In this way, different alleles are continuously being shuffled as they are transmitted from one generation to another, thereby ensuring further haplotypic and genotypic diversity in the population.

2.2.1. *Mutations and Genetic Diseases*

While meiotic recombination ensures that we do not always inherit the same genotypes from our parents, the types of alleles in the population's gene pool still remain the same despite the genetic shufflings. New genetic diversity is introduced into the gene pool via such genetic mutations as edit changes in the DNA "letters" of a gene or an alteration in the chromosomes. We call a particular DNA sequence variation a "polymorphism" if it is common, occurring in more than 1% of a population. By nature of their common occurrence, DNA polymorphisms are typically neutral—that is, they are neither harmful nor beneficial and therefore do not affect the balance of the population too much. The most common type of genetic variation is called a Single Nucleotide Polymorphism (SNP). SNPs are the smallest possible change in DNA, involving only a single base change in a DNA sequence. They are found throughout the human genome with a very high frequency of about 1 in 1,000 bases. This means that there could be millions of SNPs in each human genome. The abundance, stability, and relatively even distribution of SNPs and the ease with which they can be measured make them particularly useful as "genetic markers" or genomic reference points among people to track the flow of genetic information in families or population, and even for predicting an individual's genetic risk of developing a certain disease or predicting how an individual responds to a medicine, if a genetic marker is directly linked to a phenotype of interest.

If a particular genetic variation occurs in less than 1% of the population, we call it a "mutation" instead of a polymorphism. Mutations can arise spontaneously during normal cell functions, such as when a cell divides, or in response to environmental factors such as toxins, radiation, hormones, and even diet. On average, our DNA undergoes about 30 mutations during a human lifetime. Fortunately, as a large part of our DNA are non-coding, most mutations tend to happen there and therefore do not cause problems. In addition, nature has also provided us with a system of finely tuned enzymes that find and fix most DNA errors. However, those unrepaired mutations that change a gene's coding recipe may cause disease, and genetic mutations in the sperm or egg cells can cause diseases that pass on to the next generation.

The Human Genome Project has provided us with a readable copy of our book of life written in genetic code. The next step is to begin to understand what is written in it and what it means in terms of human health and disease. Scientists have already used genomic data to pinpoint many genes that are associated with human genetic diseases. For example, the disease genes for cystic fibrosis,[429] breast cancer[315, 899] and Parkinson disease[672] have been identified. It is no longer unrealistic to hope that in the not-too-distant future, such advances in molecular biology can transform the practice of medicine to one that is more effective, personalized, and even preventive.

3. Our Biological Machineries

Although genes get a lot of attention, it is actually the proteins that perform most life functions; see Figure 5 for some examples. The DNA only provides the blueprint for making the proteins—the actual workhorses of our biological machineries. It is through the proteins that our genes influence almost everything about us, including how tall we are, how we process foods, and how we respond to infections and medicines.

Proteins are large, complex molecules made up of long chains of smaller subunits called "amino acids". There are twenty different kinds of amino acids found in proteins. While each amino acid has different chemical properties, their basic structure is fairly similar, as shown in Figure 6. All amino acids have an amino group at one end and a carboxyl group at the other end—where each amino acid differs is in the so-called "R" group which gives an amino acid its specific properties. R can be as simple as a single hydrogen atom—as in the amino acid Glycine—or as complex side chains such as CH_3-S-$(CH_2)_2$ in Methionine.

Like the nucleotides on the DNA, the amino acids are also arranged side-by-side in a protein molecules like beads on long necklaces. A protein can contain from 50 to 5,000 amino acids hooked by peptide bonds from end-to-end. However, these amino acid "necklaces" do not remain straight and orderly in the cell—they

Protein	Function
Hemoglobin	Carry oxygen in our blood to every part of our body.
Enzymes in saliva, stomach, and small intestines	Help digest food in our body.
Muscle proteins like actin and myosin	Enable all muscular movements from blinking to running.
Ion channel proteins	Control signaling in the brain by allowing small molecules into and out of nerve cells.
Receptor proteins	Hang around on the cell surface or vicinity to help transmit signal to proteins on the inside of the cells.
Antibodies	Defend our body against foreign invaders such as bacteria and viruses.

Fig. 5. Some interesting proteins in our body.

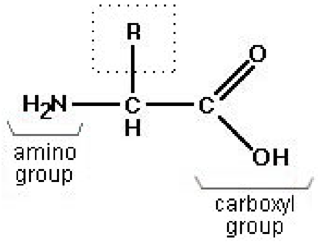

Fig. 6. Structure of an amino acid.

twist and buckle, fold in upon themselves, form knobs of amino acids, and so on. Unlike the DNA where the linear sequence of the nucleotides pretty much determines the function—with proteins, it is mostly their three-dimensional structures that dictate how they function in the cell. Interaction between proteins is a three-dimensional affair—a protein interacts with other proteins via "lock-and-key" arrangements. Misfolding of a protein can thus lead to diseases. *E.g.*, the disease cystic fibrosis is known to be caused by the misfolding of a protein called CFTR (cystic fibrosis transmembrane conductance regulator). The misfolding—in this case, due to the deletion of a single amino acid in CFTR—disrupts the molecule's function in allowing chloride ions to pass through the outer membranes of cells. This functional disruption causes thick mucus to build up in the lungs and digestive organs, and it often results in the death of patients at an early age.

The shapes of proteins are therefore a critical determinant for the proper functioning of biological faculties in our body. Without knowing the three-dimensional structures of the proteins, we cannot fully understand or make any predictions about the phenotype of the organism. The "Protein Folding Problem" remains one of the most fundamental unsolved problems in bioinformatics. While some sections on the proteins fold into regular recognizable shapes such as spiral-shaped "alpha helices" or pleated "beta sheets", scientists are still unable to reliably predict the final 3-D structure from the amino acid sequence of a protein.

Clearly, solving the folding problem has many rewarding implications. For example, if we are able to predict how a protein folds in the cell, we can theoretically design exact drugs on a computer to, say, inhibit its function without a great deal of costly experimentation. Much concerted efforts by the bioinformatics community has been invested in solving this important problem. Over the years, the structural bioinformaticians have made significant progress with this problem, as demonstrated by the results reported in the regular community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP).[580]

3.1. *From Gene to Protein: The Central Dogma*

A healthy body is a complex dynamic biological system that depends on the continuous interplay of thousands of proteins, acting together in just the right amounts and in just the right places. We have described how the DNA contains the genetic recipes for making proteins. We have also pointed out that the proteins are the actual workhorses that perform most life functions. Now, how does a sequence of DNA bases turn into a chain of amino acids and form a protein in the cell?

For a cell to make a protein, the information from a gene recipe is first copied—base by base—from a strand of DNA in the cell's nucleus into a strand of messenger RNA (mRNA). Chemically, the RNA—or ribonucleic acid—and the DNA are very similar. RNA molecules are also made up of four different nucleotides $\{$A, C, G, U$\}$—the nucleotide U (uracil) in RNA replaces the T (thymine) in DNA. Like thymine, the uracil also base-pairs to adenine. After copying the genetic recipes on the DNA in the nucleus, the mRNA molecules then travel out into the cytoplasm, and becomes accessible to cell organelles there called ribosomes. Here, each ribosome molecule reads the specific genetic code on an mRNA, and translates the genetic code into the corresponding amino acid sequence based on a genetic coding scheme; see Figure 9 in Section 3.2. With the help of transfer RNA (tRNA) molecules that transport different amino acids in the cell to the ribosome molecule as needed, the prescribed protein molecule is

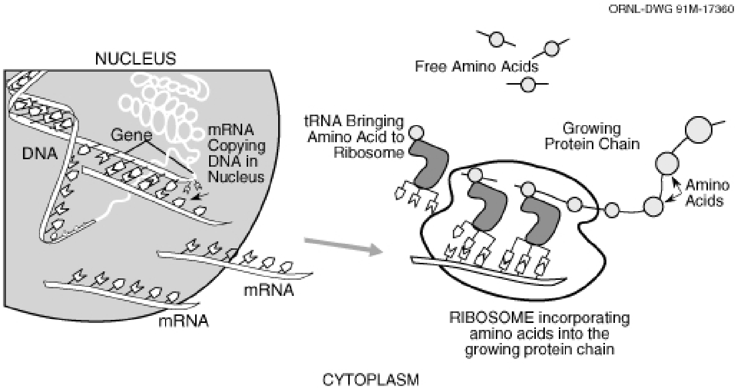
ORNL-DWG 91M-17360
NUCLEUS
DNA
Gene
mRNA Copying DNA in Nucleus
mRNA
mRNA
Free Amino Acids
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Amino Acids
RIBOSOME incorporating amino acids into the growing protein chain
CYTOPLASM

Fig. 7. From gene to protein. (*Image credit: U.S. Department of Energy Human Genome Program.*)

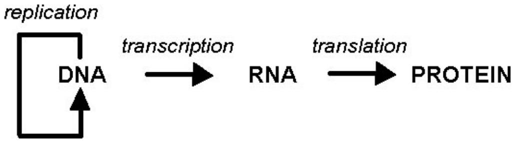

Fig. 8. The Central Dogma of Molecular Biology.

assembled—amino acid by amino acid—as instructed by the genetic recipe. Figure 7 illustrates how information stored in DNA is ultimately transferred to protein in the cell.

Figure 8 provides a schematic view of the relationship between the DNA, RNA, and protein in terms of three major processes:

(1) Replication—the process by which the information in the DNA molecule in one cell is passed on to new cells as the cell divides and the organism grows. The double-stranded complementary structure of the DNA molecule, together with the nucleotide pairing rules described earlier in Section 2.1 provide the framework for single DNA chains to serve as templates for creating complementary DNA molecules. Entire genetic blueprint can thus be passed on from cell to cell through DNA replication. In this way, virtually all the cells in our body have the full set of recipes for making all the proteins necessary to sustain life's many different functions.
(2) Transcription—the process by which the relevant information encoded in DNA is transferred into the copies of messenger RNA molecules during

the synthesis of the messenger RNA molecules. Just as in the DNA replication process, the DNA chains also serve as templates for synthesizing complementary RNA molecules. Transcription allows the amount of the corresponding proteins synthesized by the protein factories—the ribosomes in the cytoplasm—to be regulated by the rate at which the respective mRNAs are synthesized in the nucleus. Microarray experiments—see Section 4.2.2—measure gene expression with respect to the amount of corresponding mRNAs present in the cell, and indirectly infer the amount of the corresponding proteins—the gene products—present in the cell.

(3) Translation—the process by which genetic information on the mRNA is transferred into actual proteins. Protein synthesis is carried out by the ribosomes, and it involves translating the genetic code transcribed on the mRNA into a corresponding amino-acid string which can then fold into the functional protein. The genetic code for the translation process has been "cracked" by scientists in the 1960s—we describe it in more details in the next section.

This multi-step process of transferring genetic information from DNA to RNA to protein is known as the "Central Dogma of Molecular Biology"—often also considered as the backbone of molecular biology. This multi-step scheme devised by nature serves several important purposes. The installation of the transcription process protects the central "brain" of the entire system—the DNA—from the caustic chemistry of the cytoplasm where protein synthesis occurs. At the same time, the transcription scheme also allows easy amplification of gene information—many copies of an RNA can be made from one copy of DNA. The amount of proteins can then be controlled by the rate at which the respective mRNAs are synthesized in the nucleus, as mentioned earlier. In fact, the multi-step DNA-protein pathway in the dogma allows multiple opportunities for controlling in different circumstances. For example, in eukaryotic cells, after copying the genetic recipe from the DNA, the mRNA is spliced before it is processed by a ribosome. Alternative splicing sites on the mRNA allows more than one type of proteins to be synthesized from a single gene.

3.2. *The Genetic Code*

But how does a mere four letter alphabet in the DNA or RNA code for all possible combinations of 20 amino acids to make large number of different protein molecules? Clearly, to code for twenty different amino acids using only letters from $\{$A, C, G, U$\}$, a scheme of constructing multi-letter "words" from the 4-letter alphabet to represent the amino acids is necessary.

It turns out that the genetic code is indeed a triplet coding scheme—a run of three nucleotides, called a "codon", encodes one amino acid. A two-letter scheme is clearly inadequate as it can only encode $4^2 = 16$ different amino acids. Since $4^3 = 64 > 20$, a triplet coding scheme is sufficient to encode the 20 different amino acids. Figure 9 shows all possible mRNA triples and the amino acids that they specify—as revealed by industrious scientists such as Marshall Nirenberg and others who "cracked" the code in the 1960s.[177, 480, 616] The triplet scheme can theoretically code for up to 64 distinct amino acids. As there are only 20 distinct amino acids to be coded, the extraneous coding power of the triplet coding scheme is exploited by nature to encode many of the 20 amino acids with more than one mRNA triplet, thus introducing additional robustness to the coding scheme.

Since the genetic code is triplet-based, for each genetic sequence there are three possible ways that it can be read in a single direction. Each of these possibility is called a "reading frame", and a proper reading frame between a start codon and an in-frame stop codon is called an "open reading frame" or an ORF. Typically, the genetic messages work in such a way that there is one reading frame that makes sense, and two reading frames that are nonsense. We can illustrate this point using an analogy in English. Consider the following set of letters:

```
shesawthefatmanatethehotdog
```

The correct reading frame yields a sentence that makes sense:

```
she saw the fat man ate the hot dog
```

The other two alternative reading frames produce nonsensical sentences:

```
s hes awt hef atm ana tet heh otd og
sh esa wth efa tma nat eth eho tdo g
```

It is thus possible to computationally identify all the ORFs of a given genetic sequence given the genetic translation code shown in Figure 9.

3.3. *Gene Regulation and Expression*

Each of us originates from a single cell—the "zygote"—a single fertilized cell that contains the complete programmatic instructions for it to develop into a complex multi-cellular organism. We now know that these genetic instructions are encoded in the DNA found in nucleus of the cell, and we know how cells execute the genetic instructions under the Central Dogma. But how do cells in our body which is made up of trillions of cells with many different varieties, shapes, and sizes, selectively executes the genetic instructions so that a bone cell makes bone proteins and a liver cell makes liver proteins, and not vice versa?

1st	2nd position				3rd
position	U	C	A	G	position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	*stop*	*stop*	A
	Leu	Ser	*stop*	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Fig. 9. The Genetic Code. AUG is also known as the “start codon” that signals the initiation of translation. However, translation does not always start at the first AUG in an mRNA—predicting the actual translation initiation start sites is an interesting bioinformatics problem.[658, 869, 928] The three combinations that do not specify any amino acids—UAA, UAG, and UGA—are “stop codons” that code for chain termination.

The distinct variety of cells in our body indicates that different subsets of genes must be “switched on” in different cells to selectively generate different kinds of proteins necessary for the cells’ particular biological functions. In order for a single zygote to develop into a complex multi-cellular organism comprising trillions of cells of different varieties, the workings of the genetic machinery in our cells must be a highly regulated process—the types of genes switched on at any particular time must be controlled precisely, and the amount of each protein expressed by each active gene regulated according to different cellular events.

There are regulatory proteins in the cells that recognize and bind to specific sequences in the DNA such as the "promoters"—sequences upstream of a coding gene that contains the information to turn the gene on or off—to influence the transcription of a gene. Some genes are normally inactive—their transcription is blocked by the action of repressor proteins. On the occasion when the production of the protein is needed, the corresponding gene is induced by the arrival of an inducer molecule that binds to the repressor protein, rendering it inactive and re-enabling transcription of the gene. Other genes are normally active and must be constantly transcribed. For these genes, their repressor proteins are produced in the inactive form by default. When the production of a protein needs to be reduced or stopped, transcription of these genes is blocked with the arrival and binding of a suitable corepressor molecule that forms a complex with the repressor molecule to act as a functional repressor that blocks the gene's transcription. More detail of transcription is discussed in greater details in Chapter 5.

Other than controlling expression of genes at transcription initiation, regulation can also occur at the other points in the gene-to-protein pathway as governed by the Central Dogma depicted in Figure 8. Regulation can occur post-transcription in the "alternative splicing" process mentioned earlier, allowing biologically different proteins to be generated from the same gene under different circumstances. Regulation can also take place at translational initiation by modulating the activities of translational initiation factors such as the phosphorylation of initiation factors and their regulated association with other proteins. Regulation can also be implemented post-translation—modification such as glycosylation and acetylation can be employed to switch certain proteins on or off as necessary. More detail of translation is discussed in Chapter 4.

In short, the biological mechanisms that regulate our gene expression constitute a highly parallel and finely tuned system with elaborate control structure that involves extensive multi-factorial feedback and other circumstantial signals. At best, the sequencing of the genomes by the Human Genome Project and other sequencing projects can only give us the recipes for making the different proteins in our cells. But to begin to answer the who, the what, the when, the how, and the why of the various biological processes in our body, we need much more information than just the genetic recipes. In the rest of this chapter, we describe several key "tools of the trade" that biologists have acquired thus far in helping them decipher our book of life. In addition to these conventional developments in experimental tools, the genome era has also seen the creation of "bioinformatics"—a new tool of the trade from the marriage of computer science with biology. The emergence of bioinformatics has brought about a novel *in silico* dimension to modern molecular biology research. We focus on describing the conventional experimental tools here, and leave it to the other authors of this book to describe the wide array of bioinformatics applications in the subsequent chapters.

4. Tools of the Trade

To manage the many challenges in genomics research, molecular biologists have developed an interesting arsenal of ingenious technological tools to facilitate their study of the DNA. We can categorize the basic operations for manipulating the DNA for scientific study as follows:

(1) Cutting—the DNA is a fairly long molecule; thus to perform experiments on DNA in the laboratory, scientists must be able to cut the DNA into smaller strands for easy manipulation. The biologically-derived "restriction enzymes" provide the experimental apparatus for cutting DNA at specific regions.
(2) Copying—the cut DNA must be replicated or "amplified" to a sufficient quantity so that experimental signals be detected. The "polymerase chain reaction"—or PCR—is an ingenious solution devised by biologists for rapidly generating millions of copies from a DNA fragment.
(3) Separating—oftentimes, a mixture of different DNA fragments is resulted; they must be separated for individual identification, and this can be done by a process called "electrophoresis".
(4) Matching—one way of rapidly identifying a strand of DNA is matching by "hybridization". Hybridization uses the complementary nature of DNA strands due to specific pairings of the nucleotides to match identical—or rather complementary—strands of DNA. This is also the basis for DNA microarrays which allows scientists to monitor whole-genome genetic expression in the cell in parallel with quantitative DNA-DNA matching.

In the rest of this chapter, we describe the various staple biotechnological tools of the trade for the above basic operations in further details. We then conclude the chapter by describing how these basic operations are put together to enable two key applications in the genome era—the sequencing of the genome and the concurrent monitoring of gene expression by microarray.

4.1. *Basic Operations*

4.1.1. *Cutting DNA*

To study the DNA in the laboratory, we must be able to cut the long DNA at specific pre-determined points to produce defined and identifiable DNA fragments for further manipulation such as the joining of DNA fragments from different origins in recombinant DNA research. A pair of “molecular scissors” that cuts discriminately only at specific sites in their sequence is needed.

Fortunately, nature has provided us with a class of such molecular scissors called “restriction enzymes”. Many bacteria make these enzymatic molecular scissors to protect themselves from foreign DNA brought into their cells by viruses. These restriction enzymes degrade foreign DNAs by cutting the area that contain specific sequences of nucleotides.

A restriction enzyme functions by scanning the length of a DNA molecule for a particular 4 to 6-nucleotide long pattern—the enzyme’s so-called “recognition site”. Once it encounters its specific recognition sequence, the enzyme binds to the DNA molecule and makes one cut at the designated cleavage site in each of the DNA strand of the double helix, breaking the DNA molecule into fragments. The enzyme is said to “digest” the DNA, and the process of cutting the DNA is therefore called a restriction digest or a digestion.

Since restriction enzymes are isolated from various strains of bacteria, they have such names as *Hin*dII and *Eco*RI, where the first part of the name refers to the strain of bacteria which is the source of the enzyme—*e.g.*, *Haemophilus influenzae* Rd and *Escherichia coli* RY 13—and the second part of the name is a Roman numeral typically indicating the order of discovery. The first sequence-specific restriction enzyme called “*Hin*dII” was isolated around 1970 by a group of researchers working with *Haemophilus influenzae* bacteria at Johns Hopkins University.[779] *Hin*dII always cuts DNA molecules at a specific point within the following 6-base pair sequence pattern:

5’-GT[T/C]$^{\downarrow}$[A/G]AC-3’
3’-CA[A/G]$_{\uparrow}$[T/C]TG-5’

Restriction enzymes are useful experimental tools for molecular biologists for manipulating DNA because they allow the molecular biologists to cut DNA at specific pre-determined locations. Different restriction enzymes recognize different sequence patterns. Ever since the discovery of “*Hin*dII”, molecular biologists have already isolated many restriction enzymes from different strains of bacteria in their arsenal of specialized molecular scissors, enabling them to cut DNA at hundreds of distinct DNA cleavage sites in the laboratory.

4.1.2. *Copying DNA*

Another essential operation in the laboratory is the replication of DNA. For experimental analysis, the DNA must often be replicated or cloned many times to provide sufficient material for experimental manipulation. DNA can be replicated either *in vivo* or *in vitro*—for clarity, we use the term "cloning" for replicating DNA in host cells, and "amplification" for the test-tube equivalent.

Cloning is done by a cut-and-paste process to incorporate a target DNA fragment into the DNA of a "vector" such as a bacteriophage—a virus that infects bacteria—for *in vivo* replication. Scientists carry out the cut-and-paste process on a DNA molecule in the following steps. The source and vector DNA are first isolated and then restriction enzymes are used to cut the two DNAs. This creates ends in both DNAs that allows the source DNA to connect with the vector. The source DNA are then bonded to the vector using a DNA ligase enzyme that repairs the cuts and creates a single length of DNA. The DNA is then transferred into a host cell—a bacterium or another organism where the recombinant DNA is replicated and expressed. As the cells and vectors are small and it is relatively easy to grow a lot of them, copies of a specific part of a DNA or RNA sequence can be selectively produced in an unlimited amount. Genome-wide DNA libraries can be constructed in this way for screening studies.

In comparison, amplification is done by the polymerase chain reaction or PCR—an *in vitro* technique that allows one to clone a specific stretch of DNA in the test tube, without the necessity of cloning and sub-cloning in bacteria. The basic ingredients for a PCR process are: a DNA template, two primers, and DNA polymerase—an enzyme that replicates DNA in the cell. The template is the DNA to be replicated; in theory, a single DNA molecule suffice for generating millions of copies. The primers are short chains of nucleotides that correspond to the nucleotide sequences on either side of the DNA strand of interest. These flanking sequences can be constructed in the laboratory or simply purchased from traditional suppliers of reagents for molecular biology.

The PCR process was invented by Kary Mullis[585] in the mid 1980's and it has since become a standard procedure in molecular biology laboratories today. In fact, the process has been fully-automated by PCR machines, often also known as thermal cyclers. Replication of DNA by PCR is an iterative process. By iterating the replication cycle, millions of copies of the specified DNA region can be rapidly generated at an exponential rate by a combinatorial compounding process. Each PCR cycle is a 3-step procedure, as depicted in Figure 10:

(1) Denaturation—the first step of a PCR cycle is to create single-stranded DNA templates for replication. In its normal state, the DNA is a two-stranded molecule held together by hydrogen bonds down its center. Boiling a solution of DNA adds energy and breaks these bonds, making the DNA single-stranded. This is known as "denaturing"—or metaphorically, "melting"—the

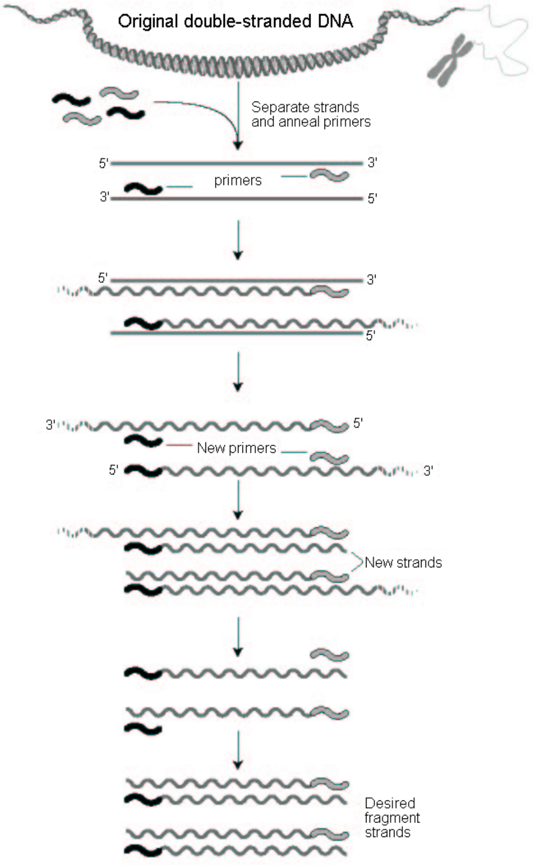

Fig. 10. Polymerase Chain Reaction. (*Image credit: National Human Genome Research Institute.*)

DNA, and it usually takes place around 94 °C. This is, however, a generally unstable state for DNA, and it will spontaneously re-form a double-helix if permitted to cool slowly.

(2) Primer annealing—the second step is to specify a region on the DNA to be replicated by creating a sort of "content-addressable pointer" to the start and the end of the target region. To do so, we label the starting point for DNA synthesis with a synthetic oligonucleotide primer that anneals—at a lower temperature, say about 30–65 °C to the single-stranded DNA templates at that

point. By supplying a pair of flanking primers, only the DNA region between the marked points is amplified by the DNA polymerase in due process.

(3) DNA synthesis—the third step is to replicate the DNA region specified by the primers. This occurs at a temperature of about 72 °C that facilitates elongation of DNA from the annealed primers by a thermostable polymerase such as the *Taq* polymerase. The *Taq* polymerase was first isolated from *Thermus aquaticus*—a bacterium found in the geysers in Yellowstone National Park. The elongation process results in new complementary strands on the templates, which can then be denatured, annealed with primers, and replicated again by cycling between the three different levels of temperature for DNA denaturing, annealing, and elongation. This explains why the term "thermal cyclers" is also used for machines that automate the PCR process.

By going through a PCR cycle, the concentration of the target sequence is doubled. After n cycles, there will be up to 2^n replicated DNA molecules in the resulting PCR mixture. Typically, the PCR cycle is repeated for about 30–60 cycles, generating millions of copies of DNA in a couple of hours.

4.1.3. *Separating DNA*

Experiments often produce mixtures of DNA fragments which must then be separated and identified. For example, when a DNA molecule is digested by restriction enzymes, fragments of DNA of different lengths are produced which must then be separated and identified. Even in PCR, although the resulting product is expected to contain only fragments of DNA of a defined length, a good experimentalist always verifies it by taking a sample from the reaction product and check for the presence of any other fragments by trying to separate the DNA produced.

The common way to separate macromolecules such as DNA and proteins in the laboratory is by size or their electric charge. When exposed to an electric field, a mixture of macromolecules travels through a medium—such as an agarose or acrylamide gel—at different rates depending on their physical properties. By labeling the macromolecules with a radioactive or fluorescent molecule, the separated macromolecules can then be seen as a series of bands spread from one end of the gel to the other. If we know what molecules are expected to be in the mixture, we may then deduce the identities of the molecules in individual bands from their relative size difference. We can run molecular weight standards or DNAs with known sizes on the gel together with the unknown DNA molecules if it is necessary to determine the actual sizes of these DNA molecules; see Figure 11. We can then use the calibration curve generated from the gel positions of the molecular size standards to interpolate the actual sizes of the unknown DNA molecules.

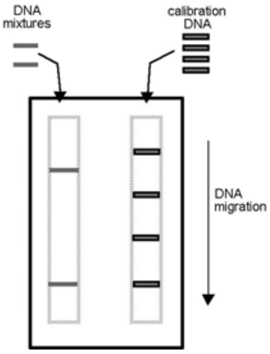

Fig. 11. Gel Electrophoresis.

This laboratory technique of separating and identifying biomolecules is called "gel electrophoresis". Gel electrophoresis can also be used for the isolation and purification of macromolecules. Once the molecules in the mixture are separated into discrete bands on the gels, they can be individually retrieved from the gel for further processing by a "blotting" procedure that transfers the separated macromolecules from the gel matrix onto an inert support membrane. If DNA is the type of macromolecules being separated, the process is called "Southern blotting". It is called a "Northern blot" and a "Western blot" respectively if mRNA and proteins are the macromolecules being separated.

4.1.4. *Matching DNA*

A direct way to identify DNA is by sequence similarity. This can be determined experimentally using a technique called "DNA hybridization". Hybridization is based on the property that complementary nucleic acid strands bind quite specifically to each other—the underlining principle being the base-pairing of nucleotides, *viz.* A-T and G-C for DNA, and A-U and G-C for RNA. The more similar two single-stranded DNA molecules are in terms of their sequences, the stronger they bind or hybridize to each other.

In DNA hybridization experiments, a small nucleic acid sequence—either a DNA or RNA molecule—can be used as a "probe" to detect complementary sequences within a mixture of different nucleic acid sequences. For example, we can chemically synthesize an oligonucleotide of a sequence complementary to the gene of interest. Just like in gel electrophoresis, we can label the probe with a dye or other marker for detection.

Typically, different DNA or RNA fragments—for example, those obtained by blotting from gel electrophoresis—are spotted and immobilized on a membrane and then allowed to react with a labeled probe. The membrane is then washed extensively to remove non-specifically bound probes, and the spots where the probe have remained bound are visualized. In this way, the sequence identity of the DNA or RNA in those spots on the membrane is revealed. It is important to note that some degree of hybridization can still occur between DNA with inexact match with the probe. For accurate detection, it is important to design the probe such that it is sufficiently specific for its target sequence so that it does not hybridize with any other DNA or RNA sequences that are also present. Fortunately, probe specificity can be achieved quite easily—even with a short probe length of say 20 nucleotides, the chance that two randomly occurring 20-nucleotide sequence are matching is very low, namely 1 in 4^{20} or approximately 1 in 10^{12}.

4.2. ***Putting It Together***

We conclude this section with two applications in the genome era—the sequencing of the genome, and the concurrent monitoring of gene expression. Genome sequencing allows us to read our book of life letter by letter. Gene expression profiling allows us to take snapshots of the genetic expression of our cells in different conditions and localization to reveal how they responds to different needs.

4.2.1. *Genome Sequencing*

The Human Genome Project was an ambitious project started in 1990 that aims to read—letter by letter—the 3 billion units of human DNA in 15 years. Thanks to the concerted efforts of international scientists and the timely emergence of many innovative technologies, the project was completed ahead of its schedule—a working draft was published in 2001, followed with a more accurate reference sequence in 2003. We describe here, how the various basic tools of the trade that we have described in the previous section are put together to achieve the Herculean task of reading the formidable human genome.

As there are no technology that allows us to effectively and reliably read the A's, C's, G's, and T's from one end of the genome to the other end in a single pass, we must first cut the DNA molecule into smaller fragments. For this, we employ the restriction enzymes to cut the DNA into small manageable fragments as described earlier. Then, in a process ingeniously modified from the DNA replication process in PCR, we read out sequence on a DNA fragment letter by letter.

To do so, the four major ingredients for a PCR reaction are put together into a sequencing mixture: the source DNA template, its corresponding primers, the DNA polymerase, and sufficient quantity of free nucleotides for replication. The only difference from standard PCR is that for sequencing reactions, two different classes of the free nucleotides are added. In addition to the normal free nucleotides {A, C, G, T}, we also add a few modified ones attached with an extra fluorescent dye {A*, C*, G*, T*}. These colored nucleotides have the special property that when they are attached to a growing strand during the elongation process, it stops the new DNA strand from growing any further. Even though we have used the same "*" symbol to denote the attachment of a dye to a nucleotide, in a sequencing reaction, a different colored dye is attached to each of the four kinds of bases.

The reaction is then iterated over many cycles to replicate many complementary copies of the source DNA template just like in the PCR process. However, unlike regular PCR, the presence of the colored nucleotides introduces many occurrences of incomplete replication of the source DNA at each cycle. This results in many new DNA templates in the mixture—each starting at the same primer site but ending at various points downstream with a colored nucleotide. In due course, the completed sequencing reaction ends up containing an array of colored DNA fragments, each a one-sided complementary substring of the original sequence, differing by a base at a time. For example, a sequencing reaction for a DNA template with the source sequence ATGAGCC is going to end up with the following products in its sequencing mixture:

```
[primer]-T*
[primer]-T-A*
[primer]-T-A-C*
[primer]-T-A-C-T*
[primer]-T-A-C-T-C*
[primer]-T-A-C-T-C-G*
[primer]-T-A-C-T-C-G-G*
```

Note that the sequence of the longest colored DNA fragment generated, TACTCGG, is the complementary sequence of the source sequence.

The DNA molecules produced in the sequencing reaction must then be separated and identified using an electrophoretic process. The products of sequencing reactions are fed into an automated sequencing machine. There, a laser excites the fluorescent dyes attached to the end of each colored DNA fragment, and a camera detects the color of the light emitted by the excited dyes. In this way, as the DNA molecules passes down the gel in increasing number of nucleotides, the sequencing machine detects the last nucleotide of each fragment, and the sequence can be read out letter by letter by a computer. Figure 12 shows an example of how the differently colored bands for each DNA fragments can be read out into a sequence.

Each single sequencing reaction allows us to read out the sequence of a few

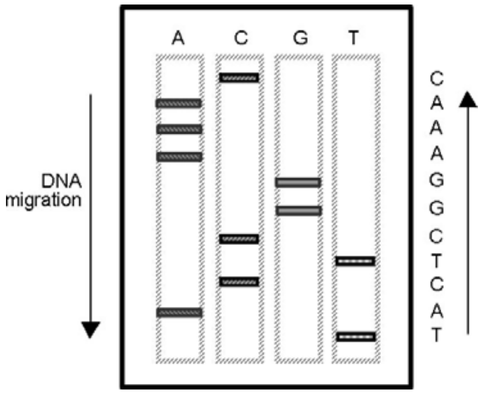

Fig. 12. Reading out the sequence.

hundred letters of DNA at a time. This amounts to a small fraction of a page in our Book of Life. By exploiting the overlaps in the sequences, the various fragments can then stitched together by bioinformaticians into an intact whole on the computer using sequence assembly programs. By the cross-disciplinary efforts of the biologists and the bioinformaticians, scientists have managed to assemble our Book of Life ahead of the planned schedule. We can now begin to read it in its entirety, letter by letter, to uncover the secrets of how life works.

4.2.2. *Gene Expression Profiling*

The sequencing of genomes by the Human Genome Project and other sequencing projects can only give us the recipes for making the different proteins in our cells. Our body is a complex and dynamic biological system that depends on the interplay of thousands of proteins in the cells, acting together in just the right amounts and in just the right places. To begin to uncover the underlying networks and intricacies of the various biomolecules, we need to be able to peep into the cell to discover what genes are turned on—or expressed—at different times and places. This can be done by gene expression profiling, using the so-called microarray which is based on the laboratory technique of DNA matching by hybridization described previously.

There are two key recent technological advancements that make gene expression profiling a possibility for scientists. First, the advent of high-throughput DNA sequencing—as described in the previous section—has resulted in the availability of many genome sequences. As a result, the DNA sequences of the genes on the genome are now known to scientists for experimentation by hybridization studies. Second, the emergence of DNA microarray technology has made it possible for scientists to systematically and efficiently profile the types and amounts of mRNA produced by a cell to reveal which genes are expressed in different conditions.

A microarray is a solid support such as a glass microscope slide, a nylon membrane, or a silicon chip, onto which DNA molecules are immobilized at fixed spots. The DNA samples may be applied to the array surface by pins, nibs or ink-jet technology,[632] or they may be directly synthesized onto the array by *in situ* photolithographic synthesis of oligonucleotides.[142, 518] With the help of robotics, tens of thousands of spots—each containing a large number of identical DNA molecules—can be placed on a small array of dimension, say, less than an inch. In a whole-genome scan by gene expression, for instance, each of these spots would contain a unique DNA fragment that identify a gene in the genome.

Microarrays make use of our ability to match DNA by hybridization as described in Section 4.1.4. Through the preferential binding of complementary single-stranded nucleic acid sequences, we can rapidly determine the identity of an mRNA and the degree it is expressed in the cell by hybridization probing. Let us suppose that we are interested in comparing the gene expression levels of four different genes, A, B, C, and D, in a particular cell type in two different states, namely: a healthy and diseased state. First, a microarray slide is prepared with unique DNA fragments from genes A, B, C, and D immobilized onto known locations on the array as probing "targets". Then, the mRNAs from both the healthy and diseased cells are extracted. The mRNAs are then used as templates to create corresponding cDNAs that are labeled with different fluorescent tags to mark their cellular origins—for example, a green dye for the healthy cells and a red dye for the diseased cells.

The two color-labeled samples from the healthy and diseased cells are then mixed and incubated with the microarray. The labeled molecules—or "probes"—then bind to the spots on the array corresponding to the genes expressed in each cell. The excess samples are washed away after the hybridization reaction. Note that in the terminology of Affymetrix, a "probe" refers to the unlabeled oligonucleotides synthesized on their microarray—known as GeneChip—rather than the solution of labeled DNA to be hybridized with a microarray as described here.

The microarray is then placed in a "scanner" to read out the array. Just as in the case of sequencing machines, the fluorescent tags on the molecules are excited by laser. A camera, together with a microscope, captures the array with a digital image which can then be analyzed by image quantitation software. The software identifies each spot on the array, measures its intensity and compares to the background. For each target gene, if the RNA from the healthy sample is in abundance, the spot is green. If the RNA from the diseased sample is in relative abundance, it is red. If there are no differential expression between the two cellular states, the spot is yellow. And if the gene is never expressed in the cell regardless of whether it is healthy or not, then its corresponding spot does fluoresce and it appears black. In this way, we estimate the relative expression levels of the genes in both samples using the fluorescence intensities and colors for each spot.

Microarray technology is a significant advance because it enables scientists to progress from studying the expression of one gene in several days to hundreds of thousands of gene expressions in a single day, or even just several hours. To cope with the huge amount of data generated by microarray studies, increasingly sophisticated bioinformatics methods and tools are required to derive new knowledge from the expression data. We discuss the analysis of microarray data with bioinformatics methods in Chapter 14. In general, it can be said that the new tools and technologies for biological research in the genome era and post-genome era have led to huge amounts of data being generated at a rate faster than we can interpret it. Thus far, informatics has played a critical role in the representation, organization, distribution and maintenance of these data in the digital form, thereby helping to make the unprecedented growth of biological data generation a manageable process. The new challenge, however, is for informatics to evolve into an integral element in the science of biology—helping to discover useful knowledge from the data generated so that ultimately, the immense potential of such ingenious inventions as PCR and microarrays can be unleashed for understanding, treating, and preventing diseases that affect the quality of human life each day.

5. Conclusion

As bioinformatics becomes an indispensable element for modern genomics, a practical bioinformatician must learn to speak the language of biology and be appreciative of how the data that we analyze are derived. This introductory chapter provides a brief overview of the major foundations of modern molecular biology and its various tools and applications. Its objective is to prime a non-biologically trained bioinformatician with sufficient basic knowledge to embark—together with the biologists—on the exciting and rewarding quest to uncover the many intriguing secrets in our book of life.

As practical bioinformaticians, we must keep in mind that biology—just like information technology—is now a rapidly evolving field. Our self-education in biology should be a firm commitment of a continuous process—it should not stop at this chapter. The future discovery of noteworthy methodologies and pivotal findings in molecular biology definitely requires the involved participation of the bioinformaticians together with the biologists—ultimately, it is those who are truly trained in the cross-disciplinary spirit who are able to make the most remarkable contributions.