3 *Molecular Biology Primer*

To understand bioinformatics in any meaningful way, it is necessary for a computer scientist to understand some basic biology, just as it is necessary for a biologist to understand some basic computer science. This chapter provides a short and informal introduction to those biological fundamentals. We scanned existing bioinformatics books to find out how much biological material was "relevant" to those books and we were surprised how little biological knowledge was actually presented. It would be safe to say that the minimum biological background one needs in order to digest a typical bioinformatics book could fit into ten pages.[1] In this chapter we give a brief introduction to biology that covers most of the computational concepts discussed in bioinformatics books. Some of the sections in this chapter are not directly related to the rest of the book, but we present them to convey the fascinating story of molecular biology in the twentieth century.

3.1 What Is Life Made Of?

Biology at the microscopic level began in 1665 when a maverick and virtuoso performer of public animal dissections, Robert Hooke, discovered that organisms are composed of individual compartments called *cells*. Cell theory, further advanced by Matthias Schleiden and Theodor Schwann in the 1830s, marked an important milestone: it turned biology into a science beyond the reach of the naked eye. In many ways, the study of life became the study of cells.

[1] This is not to say that computer scientists should limit themselves to these ten pages. More detailed discussions can be found in introductory biology textbooks like Brown (17), Lewin (66), or Alberts (3).

A great diversity of cells exist in nature, but they all have some common features. All cells have a life cycle: they are born, eat, replicate, and die. During the life cycle, a cell has to make many important decisions. For example, if a cell were to attempt to replicate before it had collected all of the necessary nutrients to do so, the result would be a disaster. However, cells do not have brains. Instead, these decisions are manifested in complex networks of chemical reactions, called *pathways*, that synthesize new materials, break other materials down for spare parts, or signal that the time has come to eat or die. The amazingly reliable and complex algorithm that controls the life of the cell is still beyond our comprehension.

One can envision a cell as a complex mechanical system with many moving parts. Not only does it store all of the *information* necessary to make a complete replica of itself, it also contains all the *machinery* required to collect and manufacture its components, carry out the copying process, and kickstart its new offspring. In macroscopic terms, a cell would be roughly analogous to a car factory that could mine for ore, fabricate girders and concrete pillars, and assemble an exact working copy of itself, all the while building family sedans with no human intervention.

Despite the complexity of a cell, there seems to be a few organizing principles that are conserved across all organisms. All life on this planet depends on three types of molecule: DNA, RNA, and proteins.[2] Roughly speaking, a cell's DNA holds a vast library describing how the cell works. RNA acts to transfer certain short pieces of this library to different places in the cell, at which point those smaller volumes of information are used as templates to synthesize proteins. Proteins form enzymes that perform biochemical reactions, send signals to other cells, form the body's major components (like the keratin in our skin), and otherwise perform the actual work of the cell. DNA, RNA, and proteins are examples of *strings* written in either the four-letter alphabet of DNA and RNA or the twenty-letter alphabet of proteins. This meshes well with Schrödinger's visionary idea about an "instruction book" of life scribbled in a secret code. It took a long time to figure out that DNA, RNA, and proteins are the main players in the cells. Below we give a brief summary of how this was discovered.

[2] To be sure, other types of molecules, like lipids, play a critical role in maintaining the cell's structure, but DNA, RNA, and proteins are the three primary types of molecules that biologists study.

3.2 What Is the Genetic Material?

Schleiden's and Schwann's studies of cells were further advanced by the discovery of threadlike chromosomes in the cell nucleii. Different organisms have different numbers of chromosomes, suggesting that they might carry information specific for each species. This fit well with the work of the Augustinian monk Gregor Mendel in the 1860s, whose experiments with garden peas suggested the existence of *genes* that were responsible for inheritance. Evidence that traits (more precisely, genes) are located on chromosomes came in the 1920s through the work of Thomas Morgan. Unlike Mendel, Morgan worked in New York City and lacked the garden space to cultivate peas, so he instead used fruit flies for his experiments: they have a short life span and produce numerous offspring. One of these offspring turned out to have white eyes, whereas wild flies had red eyes. This one white-eyed male fly born in Morgan's "fly room" in New York City became the cornerstone of modern genetics.

The white-eyed male fly was mated with its red-eyed sisters and the offspring were followed closely for a few generations. The analysis of offspring revealed that white eyes appeared predominantly in males, suggesting that a gene for eye color resides on the X chromosome (which partly determines the gender of a fruit fly). Thus, Morgan suspected that genes were located on chromosomes. Of course, Morgan had no idea what chromosomes were themselves made of.

Morgan and his students proceeded to identify other mutations in flies and used ever more sophisticated techniques to assign these mutations to certain locations on chromosomes. Morgan postulated that the genes somehow responsible for these mutations were also positioned at these locations. His group showed that certain genes are inherited together, as if they were a single unit. For example, Morgan identified mutants with a black body color (normal flies are gray) and mutants with vestigial wings. He proceeded to cross black flies with vestigial wings with gray flies with normal wings, expecting to see a number of gray flies with vestigial wings, gray flies with normal wings, black flies with vestigial wings, and black flies with normal wings. However, the experiment produced a surprisingly large number of normal flies (gray body, normal wings) and a surprisingly large number of double mutants (black body, vestigial wings). Morgan immediately proposed a hypothesis that such *linked* genes reside close together on a chromosome. Moreover, he theorized, the more tightly two genes are linked (i.e., the more often they are inherited together), the closer they are on a chromosome.

Morgan's student Alfred Sturtevant pursued Morgan's chromosome theory and constructed the first genetic map of a chromosome that showed the order of genes. Sturtevant studied three genes: cn, which determines eye color; b, which determines body color; and vg, which determines wing size. Sturtevant crossed double-mutant b and vg flies with normal flies and saw that about 17% of the offspring had only a single mutation. However, when Sturtevant crossed double-mutant b and cn flies he found that 9% of the offspring had only a single mutation. This implied that b and cn reside closer together than b and vg; a further experiment with cn and vg mutants demonstrated an 8% single mutation rate. Combined together, these three observations showed that b lies on one side of cn and vg on the other. By studying many genes in this way, it is possible to determine the ordering of genes. However, the nature of genes remained an elusive and abstract concept for many years, since it was not clear how genes encoded information and how they passed that information to the organism's progeny.

3.3 What Do Genes Do?

By the early 1940s, biologists understood that a cell's traits were inherent in its genetic information, that the genetic information was passed to its offspring, and that the genetic information was organized into genes that resided on chromosomes. They did not know what the chromosomes were made of or what the genes actually did to give rise to a cell's traits. George Beadle and Edward Tatum were the first to identify the job of the gene, without actually revealing the true nature of genetic information. They worked with the bread mold *Neurospora*, which can survive by consuming very simple nutrients like sucrose and salt. To be able to live on such a limited diet, *Neurospora* must have some proteins (enzymes) that are able to convert these simple nutrients into "real food" like amino acids and the other molecules necessary for life. It was known that proteins performed this type of chemical "work" in the cell.

In 1941 Beadle and Tatum irradiated *Neurospora* with x-rays and examined its growth on the usual "spartan" medium. Not surprisingly, some irradiated *Neurospora* spores failed to grow on this diet. Beadle and Tatum conjectured that x-rays introduced some mutations that possibly "destroyed" one of the genes responsible for processing *Neurospora*'s diet into real food. Which particular gene was destroyed remained unclear, but one of the experiments revealed that the irradiated *Neurospora* survived and even flourished when

Beadle and Tatum supplemented its spartan diet with vitamin B_6. An immediate conclusion was that x-rays damaged a gene that produces a protein (enzyme) responsible for the synthesis of B_6. The simplest explanation for this observation was that the role of a gene was to produce proteins. The rule of "one gene, one protein" remained the dominant thinking for the next half-century until biologists learned that one gene may produce a multitude of proteins.

3.4 What Molecule Codes for Genes?

DNA was discovered in 1869 by Johann Friedrich Miescher when he isolated a substance he called "nuclein" from the nuclei of white blood cells. By the early 1900s it was known that DNA (nuclein) was a long molecule consisting of four types of bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Originally, biologists discovered five types of bases, the fifth being uracil (U), which is chemically similar to thymine. By the 1920s, nucleic acids were grouped into two classes called DNA and RNA, that differ slightly in their base composition: DNA uses T while RNA uses U.

DNA, or *deoxyribonucleic acid*, is a simple molecule consisting of a sugar (a common type of organic compound), a phosphate group (containing the element phosphorus), and one of four nitrogenous bases (A, T, G, or C). The chemical bonds linking together nucleotides in DNA are always the same such that the backbone of a DNA molecule is very regular. It is the A, T, C, and G bases that give "individuality" to each DNA molecule.

Ironically, for a long time biologists paid little attention to DNA since it was thought to be a repetitive molecule incapable of encoding genetic information. They thought that each nucleotide in DNA followed another in an unchanging long pattern like ATGCATGCATGCATGCATGC, like synthetic polymers. Such a simple sequence could not serve as Schrödinger's codescript, so biologists remained largely uninterested in DNA. This changed in 1944 when Oswald Avery and colleagues proved that genes indeed reside on DNA.

3.5 What Is the Structure of DNA?

The modern DNA era began in 1953 when James Watson and Francis Crick (fig. 3.1) determined the double helical structure of a DNA molecule. Just 3 years earlier, Erwin Chargaff discovered a surprising one-to-one ratio of

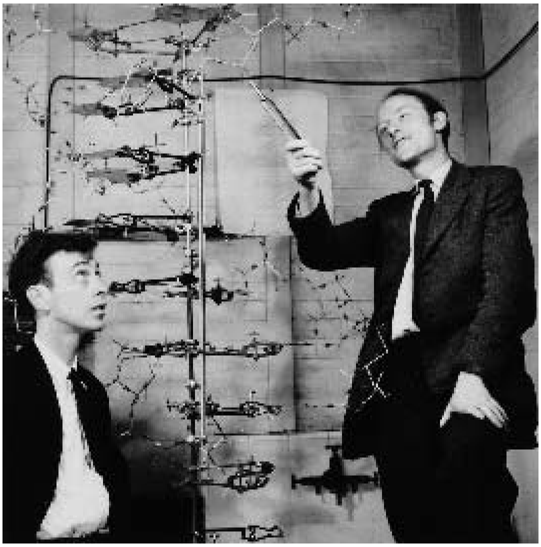

Figure 3.1 Watson and Crick puzzling about the structure of DNA. (Photo courtesy of Photo Researchers, Inc.)

the adenine-to-thymine and guanine-to-cytosine content in DNA (known as the *Chargaff rule*). In 1951, Maurice Wilkins and Rosalind Franklin obtained sharp x-ray images of DNA that suggested that DNA is a helical molecule.

Watson and Crick were facing a three-dimensional jigsaw puzzle: find a helical structure made out of DNA subunits that explains the Chargaff rule. When they learned of the Chargaff rule, Watson and Crick wondered whether A might be chemically attracted to T (and G to C) during DNA replication. If this was the case, then the "parental" strand of DNA would be complementary to the "child" strand, in the sense that ATGACC is complementary to TACTGG. After manipulating paper and metal Tinkertoy representations of bases[3] Watson and Crick arrived at the very simple and elegant double-stranded helical structure of DNA. The two strands were held together by hydrogen bonds between specific base pairings: A-T and C-G. The key ingredient in their discovery was the chemical logic begind the complementary relationship between nucleotides in each strand—it explained the

[3] Computers were not common at that time, so they built a six-foot tall metal model of DNA. Amusingly, they ran out of the metal pieces and ended up cutting out cardboard ones to take their place.

Chargaff rule, since A was predicted to pair with T, and C with G. Thus, the nucleotide string of one strand completely defined the nucleotide string of the other. This is, in fact, the key to DNA replication, and the missing link between the DNA molecule and heredity. As Watson and Crick gently put it in their one-page paper on April 25, 1953: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

3.6 What Carries Information between DNA and Proteins?

The double helix provided the key to DNA replication, but the question remained as to how DNA (a long but simple molecule) generates an enormous variety of different proteins. The DNA content of a cell does not change over time, but the concentrations of different proteins do. DNA is written in a four-letter alphabet while proteins are written in a twenty-letter alphabet. The key insight was that different pieces of a long DNA molecule coded for different proteins. But what was the code that translated texts written in a four-letter alphabet into texts written in a twenty-letter alphabet? How was this code read and executed?

First, we must realize that there are two types of cells: those that encapsulate their DNA in a *nucleus* and those that do not. The former are referred to as *eukaryotic* cells and the latter are *prokaryotic* cells. All multicellular organisms (like flies or humans) are eukaryotic, while most unicellular organisms (like bacteria) are prokaryotic. For our purposes, the major difference between prokaryotes and eukaryotes is that prokaryotic genes are continuous strings, while they are broken into pieces (called *exons*) in eukaryotes. Human genes may be broken into as many as 50 exons, separated by seemingly meaningless pieces called *introns*, whose function researchers are still trying to determine.

Understanding the connection between DNA and proteins began with the realization that proteins could not be made directly from DNA, since in eukaryotes DNA resides within the nucleus, whereas protein synthesis had been observed to happen outside the nucleus, in the *cytoplasm*. Therefore, some unknown agent had to somehow transport the genetic information from the DNA in the nucleus to the cytoplasm. In the mid 1950s Paul Zamecnik discovered that protein synthesis in the cytoplasm happens with the help of certain large molecules called *ribosomes* that contain RNA. This led to the suspicion that RNA could be the intermediary agent between DNA and proteins. Finally, in 1960 Benjamin Hall and and Sol Spiegelman demonstrated that RNA forms duplexes with single-stranded DNA, proving that the RNA (responsible for the synthesis of a particular protein) is complementary to the DNA segment (i.e., the gene) that codes for the protein. Thus, DNA served as a template used to copy a particular gene into *messenger RNA (mRNA)* that carries the gene's genetic information to the ribosome to make a particular protein.[4]

Chemically speaking, RNA, or *ribonucleic acid*, is almost the same as DNA. There are two main differences between RNA and DNA: there is no T base in RNA—the similar base U takes its place—and an oxygen atom is added to the sugar component. These two seemingly minor differences have a major impact on the biological roles of the two molecules. DNA is mostly inert and almost always double-stranded, helping it to serve as a static repository for information. RNA, on the other hand, is more chemically active and it usually lives in a single-stranded form. The effect is that RNA can carry short messages from the DNA to the cellular machinery that builds protein, and it can actively participate in important chemical reactions.

In 1960 Jerard Hurwitz and Samuel Weiss identified a molecular machine (composed of many proteins) that uses DNA as a template and adds ribonucleotide by ribonucleotide to make RNA. This process is called *transcription* and the molecular machine responsible for this process got the name *RNA polymerase*. Despite the advances in our understanding of the copying of DNA into RNA, how RNA polymerase knows where to start and stop transcribing DNA remains one of the many unsolved bioinformatics problems. Furthermore, the transcription of a gene into mRNA is tightly controlled, so that not all genes produce proteins at all times. Though some basic mechanisms of how gene transcription is controlled are known, a comprehensive understanding for all genes is still beyond our grasp.

In eukaryotes, a gene is typically broken into many pieces but it still produces a coherent protein. To do so, these cells have to cut the introns out of the RNA transcript and concatenate all the exons together prior to the mRNA entering the ribosome. This process of cutting and pasting the "raw" RNA version of the gene into the mRNA version that enters the ribosome is called *splicing* and is quite complicated at a molecular level.

[4] Later biologists discovered that not all RNAs are destined to serve as templates for building proteins. Some RNAs (like transfer RNA described below) play a different role.

DNA:	TAC	CGC	GGC	TAT	TAC	TGC	CAG	GAA	GGA	ACT
RNA:	AUG	GCG	CCG	AUA	AUG	ACG	GUC	CUU	CCU	UGA
Protein:	Met	Ala	Pro	Ile	Met	Thr	Val	Leu	Pro	Stop

Figure 3.2 The transcription of DNA into RNA, and the translation of RNA into a protein. Every amino acid is denoted with three letters, for example Met stands for the amino acid Methionine.

3.7 How Are Proteins Made?

In 1820 Henry Braconnot identified the first amino acid, glycine. By the early 1900s all twenty amino acids had been discovered and their chemical structure identified. Since the early 1900s when Emil Hermann Fischer showed that amino acids were linked together into linear chains to form proteins, proteins became the focus of biochemistry and molecular biology. It was postulated that the properties of proteins were defined by the composition and arrangement of their amino acids, which we now accept as true.

To uncover the code responsible for the transformation of DNA into protein, biologists conjectured that triplets of consecutive letters in DNA (called *codons*) were responsible for the amino acid sequence in a protein. Thus, a particular 30-base pair gene in DNA will make a protein of a specific 10 amino acids in a specific order, as in figure 3.2. There are $4^3 = 64$ different codons, which is more than three times as large as the number of amino acids. To explain this redundancy biologists conjectured that the *genetic code* responsible for transforming DNA into protein is degenerate: different triplets of nucleotides may code for the same amino acid. Biologists raced to find out which triplets code for which amino acids and by the late 1960s discovered the *genetic code* (table 3.1).[5] The *triplet rule* was therefore confirmed and is now accepted as fact.

Unlike the regular double-helical structure of DNA, the three-dimensional structure of proteins is highly variable. Researchers invest a large amount of effort into finding the structure of each protein; it is this structure that determines what role a protein plays in the cell—does it participate in the DNA replication process, or does it take part in some pathway that helps the cell metabolize sugar faster? Proteins perform most of the chemical work

[5] The exact genetic code and the set of start and stop codons may vary by species from the standard genetic code presented in table 3.1. For example, mitochondrial DNA or single-cell protozoan ciliates use a slightly different table.

Table 3.1 The genetic code, from the perspective of mRNA. The codon for methionine, or AUG, also acts as a “start” codon that initiates transcription. This code is translated as in figure 3.2.

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

in the cell, including copying DNA, moving materials inside the cell, and communicating with nearby cells. Biologists used to believe that one gene coded for one protein, but a more complex picture emerged recently with the discovery of *alternative splicing*, allowing one gene to code for many proteins.

Many chemical systems in the cell require *protein complexes*, which are groups of proteins that clump together into a large structure. A protein complex, known as RNA polymerase, begins *transcribing* a gene by copying its DNA base sequence into a short RNA base sequence (pairing a DNA T with an RNA A, a DNA A with an RNA U, and so on) called *messenger RNA*,[6] or mRNA. This short molecule is then attacked by large molecular complexes known as *ribosomes*, which read consecutive codons and locate the corresponding amino acid for inclusion in the growing polypeptide chain. Ribosomes are, in effect, molecular factories where proteins are assembled.

To help with the location of the proper amino acid for a given codon, a spe-

[6] More precisely, this is the case in prokaryotes. In eukaryotes, this RNA template undergoes the splicing process above to form mRNA.

cial type of RNA, called *transfer RNA* (tRNA), performs a specific and elegant function. There are twenty types of tRNAs, and twenty types of amino acids. Each type of amino acid binds to a different tRNA, and the tRNA molecules have a three-base segment (called an *anticodon*) that is complementary to the codon in the mRNA. As in DNA base-pairing, the anticodon on the tRNA sticks to the codon on the RNA, which makes the amino acid available to the ribosome to add to the polypeptide chain. When one amino acid has been added, the ribosome shifts one codon to the right, and the process repeats. The process of turning an mRNA into a protein is called *translation*, since it translates information from the RNA (written in a four-letter alphabet) into the protein (written in 20-letter alphabet). All proteins, including the ones necessary for this process, are produced *by* this process.

This flow of information,

$$\text{DNA} \rightarrow \textit{transcription} \rightarrow \text{RNA} \rightarrow \textit{translation} \rightarrow \text{protein},$$

is emphatically referred to as *the central dogma in molecular biology*.

3.8 How Can We Analyze DNA?

Over the years, biologists have learned how to analyze DNA. Below we describe some important techniques for copying, cutting, pasting, measuring, and probing DNA.

3.8.1 Copying DNA

Why does one need to copy DNA, that is, to obtain a large number of identical DNA fragments? From a computer science perspective, having the same string in 10^9 copies does not mean much since it does not increase the total amount of information. However, most experimental techniques (like gel electrophoresis, used for measuring DNA length) require many copies of the same DNA fragment. Since it is difficult to detect a single molecule or even a hundred molecules with modern instrumentation, amplifying DNA to yield millions or billions of identical copies is often a prerequisite of further analysis.

One method, *polymerase chain reaction* or *PCR*, is the Gutenberg printing press for DNA and is illustrated in figure 3.3. PCR amplifies a short (100- to 500-nucleotide) DNA fragment and produces a large number of identical DNA strings. To use PCR, one must know a pair of short (20- to 30-letter) strings in the DNA flanking the area of interest and design two *PCR primers*, synthetic DNA fragments identical to these strings.

Suppose we want to generate a billion copies of a DNA fragment of 500 nucleotides, that we know happens to be flanked by the 20-mer nucleotide sequence X on the left and the 20-mer nucleotide sequence Y on the right. PCR repeats a cycle of three operations: *denaturation, priming*, and *extension* to double the number of DNA fragments in every iteration. Therefore, after thirty iterations of PCR we will have on the order of 2^{30} DNA fragments, which is more than a billion copies. To start PCR, we only need a single copy of the target DNA, some artificially synthesized 20-nucleotide long DNA fragment $\overline{X}$ (many copies), some 20-nucleotide long DNA fragment Y (many copies), and billions of "spare" nucleotides (A,T,G,C).[7] We also need a molecular machine that will copy an existing DNA strand to produce a new DNA strand, and for this purpose we hijack DNA polymerase. DNA polymerase has an ability to add a complementary copy to a single-stranded DNA as long as there is a primer (i.e., $\overline{X}$ and Y) attached to the DNA strand and a sufficient supply of spare nucleotides

The denaturation step simply amounts to heating double-stranded DNA to separate it into two single strands (fig. 3.3 (top)). Priming is cooling down the solution to allow primers $\overline{X}$ and Y to hybridize to their complementary positions in DNA (fig. 3.3 (middle)). In the extension step, DNA polymerase extends the primer to produce two double-stranded DNA copies from single-stranded DNA (fig. 3.3 (bottom)). By repeatedly performing these three steps, one achieves an exponential increase in the amount of DNA, as shown in figure 3.4.

Another way to copy DNA is to *clone* it. In contrast to PCR, cloning does not require any prior information about flanking primers. However, biologists usually have no control over *which* fragment of DNA gets amplified. The process usually starts with breaking DNA into small pieces; to study an individual piece, biologists obtain many identical copies of each piece by cloning the pieces, and then try to select the individual piece of interest. Cloning incorporates a fragment of DNA into a *cloning vector*, which is a DNA molecule originating from a virus or bacterium. In this operation, the cloning vector does not lose its ability for self-replication, but carries the additional incorporated *insert* that the biologist plans to study. Vectors introduce foreign DNA into host cells (such as bacteria) which reproduce in large quantities. The self-replication process creates a large number of copies of the

[7] $\overline{X}$ stands for the Watson-Crick complement of the 20-mer X.

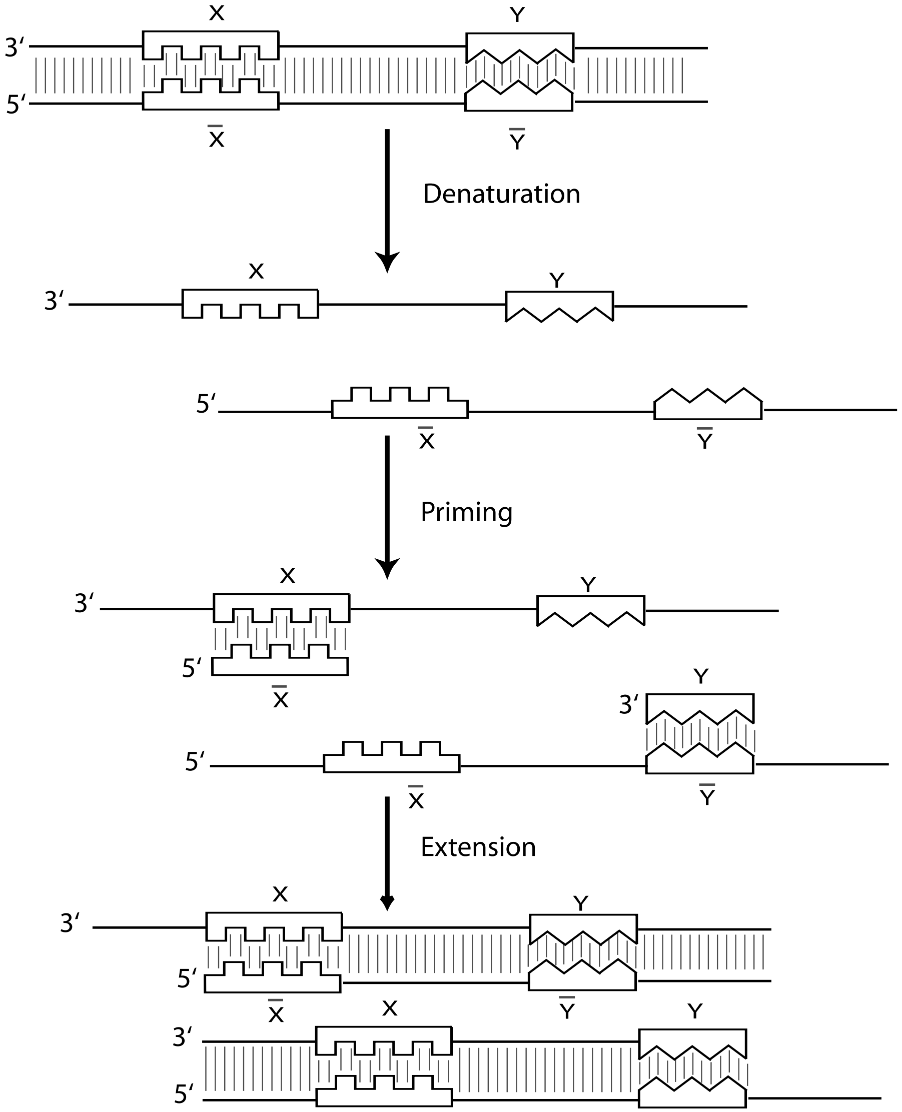

Figure 3.3 The three main operations in the polymerase chain reaction. Denaturation (top) is performed by heating the solution of DNA until the strands separate (which happens around 70 C). Priming (middle) occurs when an excess amount of primers $\overline{X}$ and Y are added to the denatured solution and the whole soup is allowed to cool. Finally, extension (bottom) occurs when DNA polymerase and excess free nucleotides (more precisely, nucleotide triphosphates) are added.

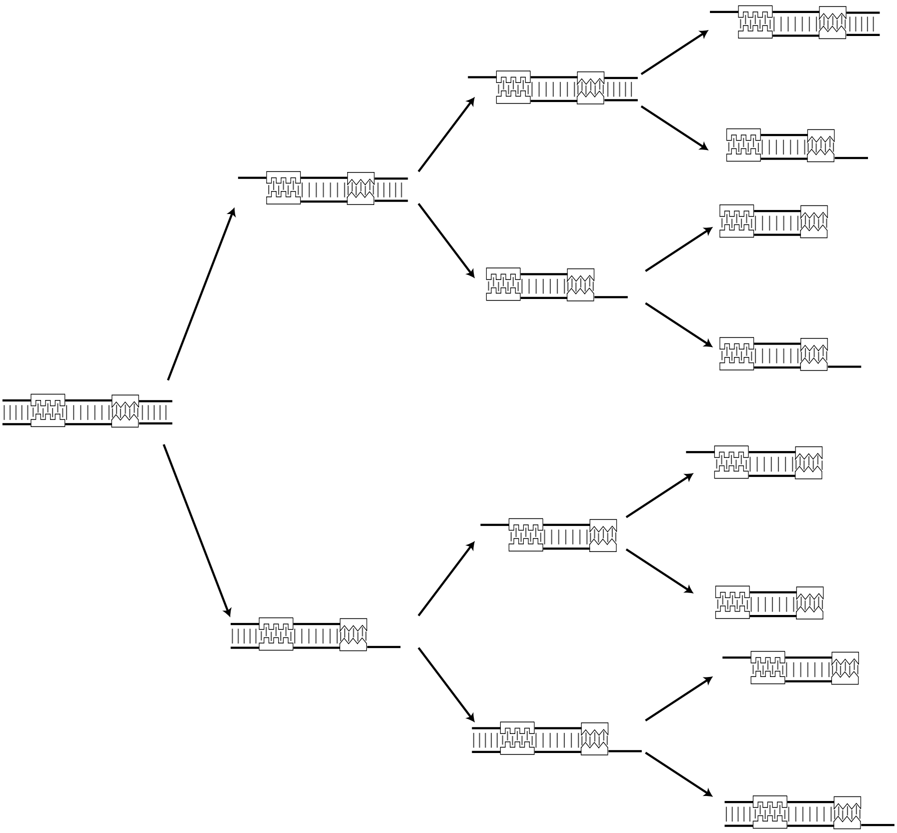

Figure 3.4 The first few iterations of PCR. Within three iterations we can go from one copy of the target DNA to eight copies.

fragment, thus enabling its properties to be studied. A fragment reproduced in this way is called a *clone*. Biologists can make *clone libraries* consisting of thousands of clones (each representing a short, randomly chosen DNA fragment) from the same DNA molecule. For example, the entire human genome can be represented as a library of 30,000 clones, each clone carrying a 100- to 200-kilobase (1000 base pairs) insert.

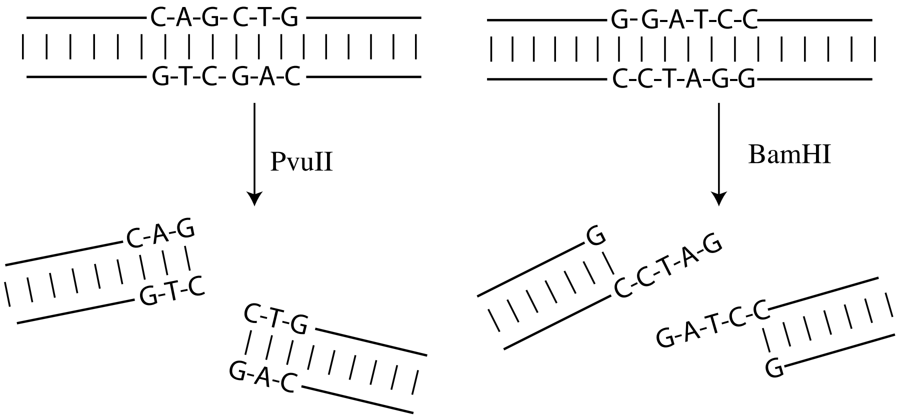

Figure 3.5 Sticky and blunt ends after cutting DNA with restriction enzymes. *Bam*HI and *Pvu*II cut at GGATCC and CAGCTG, respectively, both of which are palindromes. However, the result of *Bam*HI leaves four unmatched nucleotides on each of the strands that are cut (these unmatched nucleotides are called *sticky ends*); if a gene is cut out of one organism with *Bam*HI, it can be inserted into a different sequence that has also been cut with *Bam*HI because the sticky ends act as glue.

3.8.2 Cutting and Pasting DNA

In order to study a gene (more generally, a genomic region) of interest, it is sometimes necessary to cut it out of an organism's genome and reintroduce it into some host organism that is easy to grow, like a bacterium. Fortunately, there exist "scissors" that do just this task: certain proteins destroy the internal bonds in DNA molecules, effectively cutting it into pieces. *Restriction enzymes* are proteins that act as molecular scissors that cut DNA at every occurrence of a certain string (recognition site). For example, the *Bam*HI restriction enzyme cuts DNA into *restriction fragments* at every occurrence of the string GGATCC. Restriction enzymes first bind to the recognition site in the double-stranded DNA and then cut the DNA. The cut may produce blunt or sticky ends, as shown in figure 3.5.

Biologists have many ways to fuse two pieces of DNA together by adding the required chemical bonds. This is usually done by mimicking the processes that happen in the cell all the time: hybridization (based on complementary base-pairing) and ligation (fixing bonds within single strands), shown in figure 3.6.

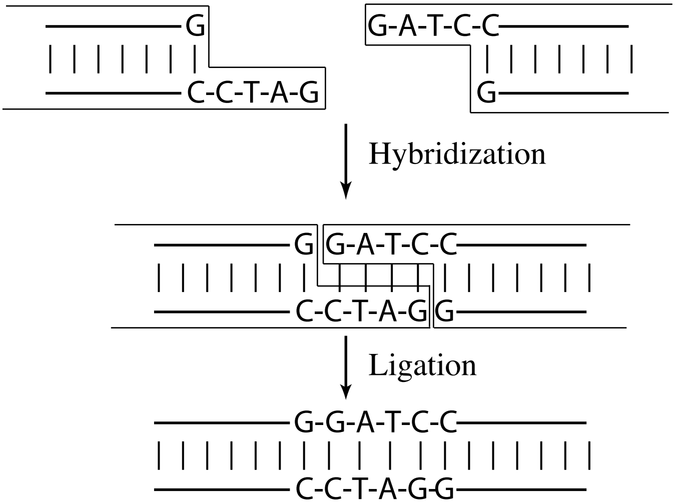

Figure 3.6 Cutting and pasting two fragments that have sticky ends (created by the restriction enzyme *Bam*HI). After hybridization, the bonds in the same DNA strands remain unfixed. The ligation step patches these bonds.

3.8.3 Measuring DNA Length

Gel electrophoresis is a technique that allows a biologist to measure the size of a DNA fragment without actually finding its exact sequence. DNA is a negatively charged molecule that migrates toward the positive pole of an electric field. The gel acts as a molecular "brake" so that long molecules move slower than short ones. The speed of migration of a fragment is related to the fragment's size, so the measurement of the migration distance for a given amount of time allows one to estimate the size of a DNA fragment. But, of course, you cannot actually see DNA molecules, so "molecular light bulbs," which are fluorescent compounds, are attached by a chemical reaction to the ends of the DNA fragments. With these bulbs, biologists can see how far different DNA fragments in a mixture migrate in the gel and thus estimate their respective lengths.

3.8.4 Probing DNA

A common task in biology is to test whether a particular DNA fragment is present in a given DNA solution. This is often done using *hybridization*: the process of joining two complementary DNA strands into a single double-stranded molecule. Biologists often use *probes*, which are single-stranded DNA fragments 20 to 30 nucleotides long that have a known sequence and a fluorescent tag. Hybridization of the probe to some unknown DNA fragment of interest can show a biologist the presence of the probe's complementary sequence in the larger DNA fragment.[8]

We can also probe RNA using a *DNA array* to see if a gene is on or off. A DNA array is essentially composed of "spots" bound to a solid support, such as a glass slide. On each spot are many copies of the complement of one gene's mRNA transcript. If the mRNA content of a cell is poured onto this slide, the mRNA will bind to the single-stranded spots and can be detected with the light-bulb technique described eariler. As a result, biologists can find out which genes are producing mRNA in a particular tissue under fixed conditions.

3.9 How Do Individuals of a Species Differ?

The genetic makeup of an individual manifests itself in *traits*, such as hair color, eye color, or susceptibility to malaria. Traits are caused by variations in genes. A surprising observation is that, despite the near similarity of genomes among all humans, no two individuals are quite the same. In fact, the variations among the same gene across different individuals are limited to a handful of different base pairs (if any). Roughly only 0.1% of the 3 billion nucleotide human genome (or 3 million bases) are different between any two individuals. Still, this leaves room for roughly $4^{3,000,000}$ different genomes, and is for all intents and purposes an endless diversity.

In other words, when we speak of "the" genome of a species, we are referring to some sort of "master" genome that is fairly representative of all the possible genomes that an individual of that species could have. While specific individuals of the species may differ in some bases, the basic long DNA sequence is roughly the same in all members of the species. Of course, this handful of differences is critically important, and the large Human Diversity Project is underway to understand how various individuals differ. This will hopefully identify the mutations reponsible for a number of genetic diseases.

[8] This is, essentially, a constant-time search of a database performed by molecular machines, something computer scientists only fantasize about!

3.10 How Do Different Species Differ?

The genomes of different organisms may be vastly different and amazingly similar.[9] The human genome consists of about 3 billion bases, while the fly genome has a scant 140 million bases. However, an analysis of the genomic sequences for two vastly different organisms (fruit flies and humans) has revealed that many genes in humans and flies are similar. Moreover, as many as 99% of all human genes are conserved across all mammals! Some human genes show strong similarity across not only mammals and flies but also across worms, plants, and (worse yet) deadly bacteria. A species, then, is a collection of individuals whose genomes are "compatible," in the sense of mating.

The likelihood that all currently living species could spontaneously develop the same gene with the same function is quite low, so it seems reasonable to assume that some process must exist that generates new species from old ones. This process is called *evolution*. The theory that all living things have evolved through a process of incremental change over millions of years has been at the heart of biology since the publication in 1859 of Charles Darwin's *On the Origin of Species*. However, only with the discovery of genomic sequences were biologists able to see how these changes are reflected in the genetic texts of existing species.

There are a number of sources for genetic variation across individuals in a species. Errors in the replication of DNA, and bizarre biological processes such as reverse transcription all cause the genomes of any two individuals in a species to be subtly different. However, genetic differences are not entirely spurious; many organisms have inherent processes that enforce genetic variation, so that no two individuals could be the same.[10] Occasionally, a variation in an individual's genome can produce a new trait, perhaps slightly stronger teeth or longer fins. If the mutations in an individual are beneficial in that individual's environment, then that individual will be more likely to be reproductively successful, passing along the mutation to its progeny. If the mutations are harmful, then that individual will be less likely to reproduce and the mutation will die out. This filtering of mutations is called *natural selection*. Over many generations the more successful individuals will become an increasingly large part of the population, to the end that the ben-

[9] There are some genetic similarities between species that are rather surprising; we will examine some of these similarities in later chapters.

[10] For example, chromosomes randomly crossover before they can make offspring. This occurs in *meiosis*, a cell replication mechanism required for most multicellular organisms to reproduce.

eficial mutation gradually takes root in all the living members of a species. As the species, as a whole, takes on the new trait, we say that it *adapts* to its environment.

If a species is divided into two isolated groups and placed into different environments, then the groups will adapt differently.[11] After many more generations, the two groups become so different that their individuals can no longer reproduce with each other, and they have become different species. This process is called *speciation*. Adaptation and speciation together form the basis of the process of evolution by natural selection, and explains the apparent paradox that there is such a diversity of life on the planet, yet so many of the genes seem similar at a sequence level. The recent abundance of genomic sequence data has enabled bioinformaticians to carry out studies that try to unravel the evolutionary relationships among different species. Since evolution by natural selection is the direct effect of adjustments to a species' genomic sequence, it stands to reason that studying the genomic sequences in different species can yield insight into their evolutionary history.

3.11 Why Bioinformatics?

As James Watson and Francis Crick worked to decipher the DNA puzzle, the 30 year-old English architect Michael Ventris tried to decipher an ancient language known as *Linear B*. At the beginning of the twentieth century, archaeologists excavated the ancient city of Knossos located on the island of Crete and found what might have been the palace of King Minos, complete with labyrinth. The archaeologists also found clay tablets with an unfamiliar form of writing. These were letters of an unknown language and there was nothing to compare them to.

The script that the ancient Cretans used (nicknamed "Linear B") remained a mystery for the next fifty years. Linguists at that time thought that Linear B was used to write in some hypothetical Minoan language (i.e., after King Minos) and cut off any investigation into the possibility that the language on

[11] For example, a small group of birds might fly to an isolated part of a continent, or a few lizards might float to an island on a log.

the tablets was Greek.[12] In 1936, a fourteen-year-old boy, Michael Ventris, went on a school trip to the Minoan exhibit in London and was fascinated with the legend of the Minotaur and the unsolved puzzle of the Minoan language. After seventeen years of code-breaking, Ventris decoded the Minoan language at about the same time Watson and Crick deciphered the structure of DNA.

Some Linear B tablets had been discovered on the Greek mainland. Noting that certain strings of symbols appeared in the Cretan texts but did not appear in Greek texts, Ventris made the inspired guess that those strings applied to cities on the island. Armed with these new symbols that he could decipher, he soon unlocked much more text, and determined that the underlying language of Linear B was, in fact, just Greek written in a different alphabet. This showed that the Cretan civilization of the Linear B tablets had been part of Greek civilization.

There were two types of clay tablets found at Crete: some written in Linear B and others written in a different script named *Linear A*. Linear A appears to be older than Linear B and linguists think that Linear A is the oldest written language of Europe, a precursor of Greek. Linear A has resisted all attempts at decoding. Its underlying language is still unknown and probably will remain undecoded since it does not seem to relate to any other surviving language in the world. Linear A and Linear B texts are written in alphabets consisting of roughly ninety symbols.

Bioinformatics was born after biologists discovered how to sequence DNA and soon generated many texts in the four-letter alphabet of DNA. DNA is more like Linear A than Linear B when it comes to decoding—we still know very little about the language of DNA. Like Michael Ventris, who mobilized the mathematics of code-breaking to decipher Linear B, bioinformaticians use algorithms, statistics, and other mathematical techniques to decipher the language of DNA.

For example, suppose we have the genomic sequences of two insects that we suspect are somewhat related, evolutionarily speaking—perhaps a fruit fly (*Drosophila melanogaster*) and a malaria mosquito (*Anopheles gamibae*). Taking the Michael Ventris approach, we would like to know what parts of the fruit fly genomic sequence are dissimilar and what parts are similar to the mosquito genomic sequence. Though the means to find this out may not be immediately obvious at this point, the alignment algorithms described later

[12] For many years biologists thought that proteins rather than DNA represent the language of the cell, which was another mistaken assumption.

in this book allow one to compare any two genes and to detect similarities between them. Unfortunately, it will take an unbearably long time to do so if we want to compare the entire fruit fly genome with the entire mosquito genome. Rather than giving up on the question altogether, biologists combined their efforts with algorithmists and mathematicians to come up with an algorithm (BLAST) that solves the problem very quickly and evaluates the statistical significance of any similarities that it finds.

Comparing related DNA sequences is often a key to understanding each of them, which is why recent efforts to sequence many related genomes (e.g., human, chimpanzee, mouse, rat) provide the best hope for understanding the language of DNA. This approach is often referred to as *comparative genomics*. A similar approach was used by the nineteenth century French linguist Jean-François Champollion who decoded the ancient Egyptian language.

The ancient Egyptians used hieroglyphs, but when the Egyptian religion was banned in the fourth century as a pagan cult, knowledge of hieroglyphics was lost. Even worse, the spoken language of Egyptian and its script (known as *demotic*) was lost soon afterward and completely forgotten by the tenth century when Arabic became the language of Egypt. As a result, a script that had been in use since the beginning of the third millennium BC turned into a forgotten language that nobody remembered.

During Napoleon's Egyptian campaign, French soldiers near the city of Rosetta found a stone (now known as the *Rosetta stone*) that was inscribed in three different scripts. Many of Napoleon's officers happened to be classically educated and one of them, a Lieutenant Bouchard, identified the three bands of scripts as hieroglyphic, demotic, and ancient Greek. The last sentence of the Greek inscription read: "This decree shall be inscribed on stelae of hard rock, in sacred characters, both native and Greek." The Rosetta stone thus presented a comparative linguistics problem not unlike the comparative genomics problems bionformaticians face today.

In recent decades biology has raised fascinating mathematical problems and has enabled important biological discoveries. Biologists that reduce bioinformatics to simply "the application of computers in biology" sometimes fail to recognize the rich intellectual content of bioinformatics. Bioinformatics has become a part of modern biology and often dictates new fashions, enables new approaches, and drives further biological developments. Simply using bioinformatics as a tool kit without a reasonable understanding of the main computational ideas is not very different from using a PCR kit without knowing how PCR works.

Bioinformatics is a large branch of biology (or of computer science) and this book presents neither a complete cross section nor a detailed look at any one part of it. Our intent is to describe those algorithmic principles that underlie the solution to several important biological problems to make it possible to understand any other part of the field.

Russell F. Doolittle, born 1931 in Connecticut, is currently a research professor at the Center for Molecular Genetics, University of California, San Diego. His principal research interests center around the evolution of protein structure and function. He has a PhD in biochemistry from Harvard (1962) and did postdoctoral work in Sweden. He was an early advocate of using computers as an aid to characterizing proteins.

For some it may be difficult to envision a time when the World Wide Web did not exist and every academician did not have a computer terminal on his or her desk. It may be even harder to imagine the primitive state of computer hardware and software at the time of the recombinant DNA revolution, which dates back to about 1978. It was in this period that Russell Doolittle, using a DEC PDP11 computer and a suite of home-grown programs, began systematically searching sequences in an effort to find evolutionary and other biological relationships. In 1983 he stunned cancer biologists when he reported that a newly reported sequence for platelet derived growth factor (PDGF) was virtually identical to a previously reported sequence for the oncogene known as ν-sis.[13] This was big news, and the finding served as a wake-up call to molecular biologists: searching all new sequences against up-to-date databases is your first order of business.

Doolittle had actually begun his computer studies on protein sequences much earlier. Fascinated by the idea that the history of all life might be traceable by sequence analysis, he had begun determining and aligning sequences in the early 1960s. When he landed a job at UCSD in 1964, he tried to interest consultants at the university computer center in the problem, but it was clear that the language and cultural divide between them was too great. Because computer people were not interested in learning molecular biology, he would have to learn about computing. He took an elementary course in FORTRAN

[13] *Oncogenes* are genes in viruses that cause a cancer-like transformation of infected cells. Oncogene ν-sis in the *simian sarcoma virus* causes uncontrolled cell growth and leads to cancer in monkeys. The seemingly unrelated *growth factor* PDGF is a protein that stimulates cell growth.

programming, and, with the help of his older son, developed some simple programs for comparing sequences. These were the days when one used a keypunch machine to enter data on eighty-column cards, packs of which were dropped off at the computer center with the hope that the output could be collected the next day.

In the mid-1960s, Richard Eck and Margaret Dayhoff had begun the Atlas of Protein Sequence and Structure, the forerunner of the Protein Identification Resource (PIR) database. Their original intention was to publish an annual volume of "all the sequences that could fit between two covers." Clearly, no one foresaw the deluge of sequences that was to come once methods had been developed for directly sequencing DNA. In 1978, for example, the entire holding of the atlas, which could be purchased on magnetic tape, amounted to 1081 entries. Realizing that this was a very biased collection of protein sequences, Doolittle began his own database, which, because it followed the format of the atlas, he called NEWAT ("new atlas"). At about the same time he acquired a PDP11 computer, the maximum capacity of which was only 100 kilobytes, much of that occupied by a mini-UNIX operating system. With the help of his secretary and his younger son (eleven years old at the time), Doolittle began typing in every new sequence he could get his hands on, searching each against every other sequence in the collection as they went. This was in keeping with his view that all new proteins come from old proteins, mostly by way of gene duplications. In the first few years of their small enterprise, Doolittle & Son established a number of unexpected connections.

Doolittle admits that in 1978 he knew hardly anything about cancer viruses, but a number of chance happenings put him in touch with the field. For one, Ted Friedmann and Gernot Walter (who was then at the Salk Institute), had sought Doolittle's aid in comparing the sequences of two DNA tumor viruses, simian virus 40 (SV40) and the polyoma virus. This led indirectly to contacts with Inder Verma's group at Salk, which was studying retroviruses and had sequenced an "oncogene" called ν-mos in a retrovirus that caused sarcomas in mice. They asked Doolittle to search it for them, but no significant matches were found. Not long afterward (in 1980), Doolittle read an article reporting the nucleotide sequence of an oncogene from an avian sarcoma virus—the famous *Rous sarcoma virus*. It was noted in that article that the Salk team had provided the authors with a copy of their still unpublished mouse sarcoma gene sequence, but no resemblances had been detected. In line with his own project, Doolittle promptly typed the new avian sequence into his computer to see if it might match anything else. He was astonished to find that in fact a match quickly appeared with the still unpublished Salk sequence for the mouse retrovirus oncogene. He immediately telephoned Inder Verma; "Hey, these two sequences are in fact homologous. These proteins must be doing the same thing." Verma, who had just packaged up a manuscript describing the new sequence, promptly unwrapped it and added the new feature. He was so pleased with the outcome that he added Doolittle's name as one of the coauthors.

How was it that the group studying the Rous sarcoma virus had missed this match? It's a reflection on how people were thinking at the time. They had compared the DNA sequences of the two genes without translating them into the corresponding amino acid sequences, losing most of the information as a result. It was another simple but urgent message to the community about how to think about sequence comparisons.

In May of 1983, an article appeared in *Science* describing the characterization of a growth factor isolated from human blood platelets. Harry Antoniades and Michael Hunkapiller had determined 28 amino acid residues from the N-terminal end of PDGF. (It had taken almost 100,000 units of human blood to obtain enough of the growth factor material to get this much sequence.) The article noted that the authors had conducted a limited search of known sequences and hadn't found any similar proteins.

By this time, Doolittle had modem access to a department VAX computer where he now stored his database. He typed in the PDGF partial sequence and set it searching. Twenty minutes later he had the results of the search; human PDGF had a sequence that was virtually identical to that of an oncogene isolated from a woolly monkey. Doolittle describes it as an electrifying moment, enriched greatly by his prior experiences with the other oncogenes. He remembers remarking to his then fifteen-year old son, “Will, this experiment took us five years and twenty minutes.” As it happened, he was not alone in enjoying the thrill of this discovery. Workers at the Imperial Cancer Laboratory in London were also sequencing PDGF, and in the spring of 1983 had written to Doolittle asking for a tape of his sequence collection. He had sent them his newest version, fortuitously containing the ν-sis sequence from the woolly monkey. Just a few weeks before the *Science* article appeared, Antoniades and Hunkapiller replied with an effusive letter of thanks, not mentioning just why the tape had been so valuable to them. Meanwhile, Doolittle had written to both the PDGF workers and the ν-sis team, suggesting that they compare notes. As a result, the news of the match was quickly made known, and a spirited race to publication occurred, the report from the Americans appearing in *Science* only a week ahead of the British effort in *Nature*. Doolittle went on to make many other matches during the mid-

1980s, including several more involving oncogenes. For example, he found a relationship between the oncogene ν-jun and the gene regulator GCN4. He describes those days as unusual in that an amateur could still occasionally compete with the professionals. Although he continued with his interests in protein evolution, he increasingly retreated to the laboratory and left bioinformatics to those more formally trained in the field.