# *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.

## 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.

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.

## 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.

### 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.

## 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.

## 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.

## 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.

## 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.

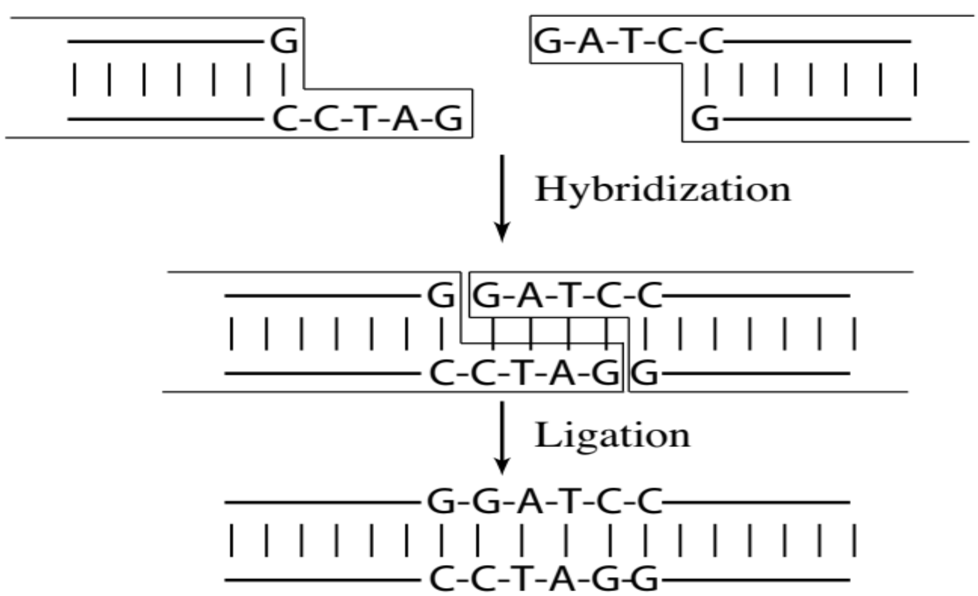


**Figure 3.6** Cutting and pasting two fragments that have sticky ends (created by the restriction enzyme *Bam*HI). After hybridization, the bonds in the same DNA strands remain unfixed. The ligation step patches these bonds.

## 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.

## 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.

## 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 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.

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.

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.

# *Exhaustive Search*

Exhaustive search algorithms require little effort to design but for many problems of interest cannot process inputs of any reasonable size within your lifetime. Despite this problem, exhaustive search, or *brute force* algorithms are often the first step in designing more efficient algorithms.

We introduce two biological problems: *DNA restriction mapping* and *regulatory motif finding*, whose brute force solutions are not practical. We further describe the *branch-and-bound* technique to transform an inefficient brute force algorithm into a practical one. In chapter 5 we will see how to improve our motif finding algorithm to arrive at an algorithm very similar to the popular CONSENSUS motif finding tool. In chapter 12 we describe two randomized algorithms, GibbsSampler and RandomProjections, that use coin-tossing to find motifs.

## Restriction Mapping

Hamilton Smith discovered in 1970 that the *restriction enzyme* *Hin*dII cleaves DNA molecules at every occurrence, or *site*, of the sequences GTGCAC or

GTTAAC, breaking a long molecule into a set of *restriction fragments.* Shortly thereafter, maps of restriction sites in DNA molecules, or *restriction maps*, became powerful research tools in molecular biology by helping to narrow the location of certain genetic markers.

If the genomic DNA sequence of an organism is known, then construction of a restriction map for *Hin*dII amounts to finding all occurrences of GTGCAC and GTTAAC in the genome. Because the first bacterial genome was sequenced twenty-five years after the discovery of restriction enzymes,

for many years biologists were forced to build restriction maps for genomes without prior knowledge of the genomes' DNA sequence.[1]

Several experimental approaches to restriction mapping exist, each with advantages and disadvantages. The distance between two individual restriction sites corresponds to the length of the restriction fragment between those two sites and can be measured by the *gel electrophoresis* technique described in chapter 2. This requires no knowledge of the DNA sequence. Biologists can vary experimental conditions to produce either a *complete* digest [fig. 4.1 (a)] or a *partial* digest [fig. 4.1 (b)] of DNA.[2] The restriction mapping problem can be formulated in terms of recovering positions of points when only pairwise distances between those points are known.

To formulate the restriction mapping problem, we will introduce some notation. A *multiset* is a set that allows duplicate elements (e.g., $\{2, 2, 2, 3, 3, 4, 5\}$ is a multiset with duplicate elements 2 and 3). If $X = \{x_1 = 0, x_2, \ldots, x_n\}$ is a set of $n$ points on a line segment in increasing order, then $\Delta X$ denotes the multiset of *all* $\binom{n}{2}$ pairwise distances[3] between points in $X$:

$$\Delta X = \{x_j - x_i : \ 1 \leq i < j \leq n\}.$$

For example, if $X=\{0, 2, 4, 7, 10\}$, then $\Delta X=\{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$, which are the ten pairwise distances between these points (table 4.1). In restriction mapping, we are given $\Delta X$, the experimental data about fragment lengths. The problem is to reconstruct $X$ from $\Delta X$. For example, could you infer that the set $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$ was derived from $\{0, 2, 4, 7, 10\}$? Though gel electrophoresis allows one to determine the lengths of DNA fragments easily, it is often difficult to judge their multiplicity. That is, the number of different fragments of a given length can be difficult to determine. However, it is experimentally possible to do so with a lot of work, and we assume for the sake of simplifing the problem that this information is given to us as an input to the problem.

This *Partial Digest problem*, or PDP, is sometimes called the *Turnpike problem* in computer science. Suppose you knew the set of distances between every (not necessarily consecutive) pair of exits on a highway leading from one town to another. Could you reconstruct the geography of the highway from this information? That is, could you find the distance from the first town to each exit? Here, the "highway exits" are the restrictions sites in DNA; the lengths of the resulting DNA restriction fragments correspond to distances between highway exits. Computationally, the only difference between the Turnpike problem and the PDP is that the distances between exits are given in miles in the Turnpike problem, while the distances between restriction sites are given in nucleotides in the PDP.

We remark that it is not always possible to uniquely reconstruct a set $X$ based only on $\Delta X$. For example, for any integer $v$ and set $A$, one can see that $\Delta A$ is equal to $\Delta(A \oplus \{v\})$, where $A \oplus \{v\}$ is defined to be $\{a + v : a \in A\}$, a *shift* of every element in $A$ by $v$. Also $\Delta A = \Delta(-A)$, where $-A = \{-a : a \in A\}$ is the *reflection* of $A$. For example, sets $A = \{0, 2, 4, 7, 10\}$, $\Delta(A \oplus \{100\}) = \{100, 102, 104, 107, 110\}$, and $-A = \{-10, -7, -4, -2, 0\}$ all produce the same partial digest. The sets $\{0, 1, 3, 8, 9, 11, 12, 13, 15\}$ and $\{0, 1, 3, 4, 5, 7, 12, 13, 15\}$

present a less trivial example of this problem of nonuniqueness. The partial digests of these two sets is the same multiset of 36 elements[4]:
$\{1_4, 2_4, 3_4, 4_3, 5_2, 6_2, 7_2, 8_3, 9_2, 10_2, 11_2, 12_3, 13, 14, 15\}$.

| | 0 | 1 | 3 | 4 | 5 | 7 | 12 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | | 1 | 3 | 4 | 5 | 7 | 12 | 13 | 15 |
| 1 | | | 2 | 3 | 4 | 6 | 11 | 12 | 14 |
| 3 | | | | 1 | 2 | 4 | 9 | 10 | 12 |
| 4 | | | | | 1 | 3 | 8 | 9 | 11 |
| 5 | | | | | | 2 | 7 | 8 | 10 |
| 7 | | | | | | | 5 | 6 | 8 |
| 12 | | | | | | | | 1 | 3 |
| 13 | | | | | | | | | 2 |
| 15 | | | | | | | | | |

| | 0 | 1 | 3 | 8 | 9 | 11 | 12 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | | 1 | 3 | 8 | 9 | 11 | 12 | 13 | 15 |
| 1 | | | 2 | 7 | 8 | 10 | 11 | 12 | 14 |
| 3 | | | | 5 | 6 | 8 | 9 | 10 | 12 |
| 8 | | | | | 1 | 3 | 4 | 5 | 7 |
| 9 | | | | | | 2 | 3 | 4 | 6 |
| 11 | | | | | | | 1 | 2 | 4 |
| 12 | | | | | | | | 1 | 3 |
| 13 | | | | | | | | | 2 |
| 15 | | | | | | | | | |

In general, sets $A$ and $B$ are said to be *homometric* if $\Delta A = \Delta B$. Let $U$ and $V$ be two sets of numbers. One can verify that the multisets

$$U \oplus V = \{u + v : u \in U, v \in V\}$$

and

$$U \ominus V = \{u - v : u \in U, v \in V\}$$

are homometric (a problem at the end of this chapter). The "nontrivial" nine-point example above came from $U = \{6, 7, 9\}$ and $V = \{-6, 2, 6\}$. Indeed $U \oplus V$=$\{0, 1, 3, 8, 9, 11, 12, 13, 15\}$ while $U \ominus V$=$\{0, 1, 3, 4, 5, 7, 12, 13, 15\}$ as illustrated below:

| $U \oplus V$ | -6 | 2 | 6 |
|---|---|---|---|
| 6 | 0 | 8 | 12 |
| 7 | 1 | 9 | 13 |
| 9 | 3 | 11 | 15 |

| $U \ominus V$ | -6 | 2 | 6 |
|---|---|---|---|
| 6 | 12 | 4 | 0 |
| 7 | 13 | 5 | 1 |
| 9 | 15 | 7 | 3 |

While the PDP is to find one set $X$ such that $\Delta X = L$, biologists are often interested in *all* homometric sets.

## Impractical Restriction Mapping Algorithms

The algorithm below, BRUTEFORCEPDP, takes the list $L$ of $\binom{n}{2}$ integers as an input, and returns the set $X$ of $n$ integers such that $\Delta X = L$. We remind the reader that we do not always provide complete details for all of the operations in pseudocode. In particular, we have not provided any subroutine to calculate $\Delta X$; problem 4.1 asks you to do fill in the details for this step.

BRUTEFORCEPDP$(L, n)$

$M \leftarrow$ maximum element in $L$
**for** every set of $n-2$ integers $0 < x_2 < \cdots < x_{n-1} < M$
  $X \leftarrow \{0, x_2, \ldots, x_{n-1}, M\}$
  Form $\Delta X$ from $X$
  **if** $\Delta X = L$
    **return** $X$
**output** "No Solution"

BRUTEFORCEPDP is slow since it examines $\binom{M-1}{n-2}$ different sets of positions, which requires about $O(M^{n-2})$ time.[5]

One might question the wisdom of selecting $n-2$ *arbitrary* integers from the interval 0 to $M$. For example, if $L$ does not contain the number 5, there is really no point in choosing any $x_i = 5$, though the above algorithm will do so. Indeed, observing that all points in $X$ have to correspond to *some* distance in $\Delta X$, we can select $n-2$ distinct elements from $L$ rather than the less constrained selection from the interval $(0, M)$. Since $M$ may be large, even with a small number of points, building a new algorithm that makes choices of $x_i$ based only on elements in $L$ yields an improvement in efficiency.

## Regulatory Motifs in DNA Sequences

Fruit flies, like humans, are susceptible to infections from bacteria and other pathogens. Although fruit flies do not have as sophisticated an immune system as humans do, they have a small set of *immunity genes* that are usually dormant in the fly genome, but somehow get switched on when the organism gets infected. When these genes are turned on, they produce proteins that destroy the pathogen, usually curing the infection.

One could design an experiment that is rather unpleasant to the flies, but very informative to biologists: infect flies with a bacterium, then grind up the flies and measure (perhaps with a DNA array) which genes are switched on as an immune response. From this set of genes, we would like to determine what triggers their activation. It turns out that many immunity genes in the fruit fly genome have strings that are reminiscent of TCGGGGATTTCC, located upstream of the genes' start. These short strings, called NF-$\kappa$B *bind-*

*ing sites*, are important examples of *regulatory motifs* that turn on immunity and other genes. Proteins known as *transcription factors* bind to these motifs, encouraging RNA polymerase to transcribe the downstream genes. Motif finding is the problem of discovering such motifs without any prior knowledge of how the motifs look.

Ideally, the fly infection experiment would result in a set of upstream regions from genes in the genome, each region containing at least one NF-$\kappa$B binding site. Suppose we do not know what the NF-$\kappa$B pattern looks like, nor do we know where it is located in the experimental sample. The fly infection experiment requires an algorithm that, given a set of sequences from a genome, can find short substrings that seem to occur surprisingly often.

"The Gold Bug", by Edgar Allan Poe, helps to illustrate the spirit, if not the mechanics, of finding motifs in DNA sequences. When the character William Legrand finds a parchment written by the pirate Captain Kidd, Legrand's friend says, "Were all the jewels of Golconda awaiting me upon my solution of this enigma, I am quite sure that I should be unable to earn them." Written on the parchment in question was

```
53++!305))6*;4826)4+.)4+);806*;48!8`60))85;]8*:+*8!
83(88)5*!;46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5*
-4)8`8*; 4069285);)6!8)4++;1(+9;48081;8:8+1;48!85;4
)485!528806*81(+9;48;(88;4(+?34;48)4+;161;:188;+?;
```

Mr. Legrand responds, "It may well be doubted whether human ingenuity can construct an enigma of the kind which human ingenuity may not, by proper application, resolve." He notices that a combination of three symbols—`; 4 8`—appears very frequently in the text. He also knows that Captain Kidd's pirates speak English and that the most frequent English word is "the." Proceeding under the assumption that `; 4 8` encodes "the," Mr. Legrand deciphers the parchment note and finds the pirate treasure. After making this substitution, Mr. Legrand has a slightly easier text to decipher:

```
53++!305))6*THE26)H+.)H+)TE06*THE!E`60))E5T]E*:+*E!
E3(EE)5*!TH6(TEE*96*?TE)*+(THE5)T5*!2:*+(TH956*2(5*
-H)E`E*T H0692E5)T)6!E)H++T1(+9THE0E1TE:E+1THE!E5TH
)HE5!52EE06*E1(+9THET(EETH(+?3HTHE)H+T161T:1EET+?T
```

You might try to figure out what the symbol ")" might code for in order to complete the puzzle.

Unfortunately, DNA texts are not that easy to decipher, and there is little doubt that nature has constructed an enigma that human ingenuity cannot entirely solve. However, bioinformaticians borrowed Mr. Legrand's method, and a popular approach to motif finding is based on the assumption that frequent or rare words may correspond to regulatory motifs in DNA. It stands to reason that if a word occurs considerably more frequently than expected, then it is more likely to be some sort of "signal," and it is crucially important to figure out the biological meaning of the signal.

This "DNA linguistics" approach is at the heart of the *pattern-driven* approach to signal finding, which is based on enumerating all possible patterns and choosing the most frequent (or the most statistically surprising) among them.

## Profiles

Figure 4.2 (a) presents seven 32-nucleotide DNA sequences generated randomly. Also shown [fig. 4.2 (b)] are the same sequences with the "secret" pattern $P$ = ATGCAACT of length $l = 8$ implanted at random positions. Suppose you do not know what the pattern $P$ is, or where in each sequence it has been implanted [fig. 4.2 (c)]. Can you reconstruct $P$ by analyzing the DNA sequences?

We could simply count the number of times each $l$-mer, or string of length $l$, occurs in the sample. Since there are only $7 \cdot (32 + 8) = 280$ nucleotides in the sample, it is unlikely that any $8$-mer other than the implanted pattern appears more than once.[9] After counting all $8$-mer occurrences in figure 4.2 (c) we will observe that, although most $8$-mers appear in the sample just once (with a few appearing twice), there is one $8$-mer that appears in the sample suspiciously many times—seven or more. This overrepresented $8$-mer is the pattern $P$ we are trying to find.

Unlike our simple implanted patterns above, DNA uses a more inventive notion of regulatory motifs by allowing for mutations at some nucleotide positions [fig. 4.2 (d)]. For example, table 4.2 shows eighteen different NF-$\kappa$B motifs; notice that, although none of them are the consensus binding site sequence TCGGGGATTTCC, each one is not substantially different. When the implanted pattern $P$ is allowed to mutate, reconstructing $P$ becomes more

CGGGGCTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAACCAAAGCGGACAAA
GGGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCTC
CTGCTGTACAACTGAGATCATGCTGCTTCAAC
TACATGATCTTTTGTGGATGAGGGAATGATGC

(a) Seven random sequences.

CGGGGCTATGCAACTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAATGCAACTCCAAAGCGGACAAA
GGATGCAACTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGATGCAACTCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCATGCAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCAACTTTCAAC
TACATGATCTTTTGATGCAACTTGGATGAGGGAATGATGC

(b) The same DNA sequences with the implanted pattern ATGCAACT.

CGGGGCTATGCAACTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAATGCAACTCCAAAGCGGACAAA
GGATGCAACTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGATGCAACTCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCATGCAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCAACTTTCAAC
TACATGATCTTTTGATGCAACTTGGATGAGGGAATGATGC

(c) Same as (b), but hiding the implant locations. Suddenly this problem looks difficult to solve.

CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA
GGATGgAtCTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCtTGgAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCcAtTTTCAAC
TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC

(d) Same as (b), but with the implanted pattern ATGCAACT randomly mutated in two positions; no two implanted instances are the same. If we hide the locations as in (c), the difficult problem becomes nearly impossible.

**Figure 4.2** DNA sequences with implanted motifs.

**Table 4.2** A small collection of putative NF-$\kappa$B binding sites.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | C | G | G | G | G | A | T | T | T | C | A |
| | A | C | G | G | G | G | A | T | T | T | T | T |
| | T | C | G | G | T | A | C | T | T | T | A | C |
| | T | T | G | G | G | G | A | C | T | T | T | T |
| | C | C | G | G | T | G | A | T | T | C | C | C |
| | G | C | G | G | G | G | A | A | T | T | T | C |
| | T | C | G | G | G | G | A | T | T | C | C | T |
| | T | C | G | G | G | G | A | T | T | C | C | T |
| | T | A | G | G | G | G | A | A | C | T | A | C |
| | T | C | G | G | G | T | A | T | A | A | A | C |
| | T | C | G | G | G | G | G | T | T | T | T | T |
| | C | C | G | G | T | G | A | C | T | T | A | C |
| | C | C | A | G | G | G | A | C | T | C | C | C |
| | A | A | G | G | G | G | A | C | T | T | C | C |
| | T | T | G | G | G | G | A | C | T | T | T | T |
| | T | T | T | G | G | G | A | G | T | C | C | C |
| | T | C | G | G | T | G | A | T | T | T | C | C |
| | T | A | G | G | G | G | A | A | G | A | C | C |
| A: | 2 | 3 | 1 | 0 | 0 | 1 | 16 | 3 | 1 | 2 | 4 | 1 |
| T: | 12 | 3 | 1 | 0 | 4 | 1 | 0 | 9 | 15 | 11 | 5 | 6 |
| G: | 1 | 0 | 16 | 18 | 14 | 16 | 1 | 1 | 1 | 0 | 0 | 0 |
| C: | 3 | 12 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 5 | 9 | 11 |
| | T | C | G | G | G | G | A | T | T | T | C | C |

complicated, since the 8-mer count does not reveal the pattern. In fact, the string ATGCAACT does not even appear in figure 4.2 (d), but the seven mutated versions of it appear at position 8 in the first sequence, position 19 in the second sequence, 3 in the third, 5 in the fourth, 31 in the fifth, 27 in the sixth, and 15 in the seventh.

In order to unambiguously formulate the motif finding problem, we need to define precisely what we mean by “motif.” Relying on a single string to represent a motif often fails to represent the variation of the pattern in real biological sequences, as in figure 4.2 (d). A more flexible representation of a motif uses a profile matrix.

Consider a set of $t$ DNA sequences, each of which has $n$ nucleotides. Select one position in each of these $t$ sequences, thus forming an array s =

```
                  CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTT...
      TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA
                     GGATGgAtCTGATGCCGTTTGACGACCTA...
                    AAGGAaGCAACcCCAGGAGCGCCTTTGCTGG...
AATTTTCTAAAAAGATTATAATGTCGGTCCtTGgAACTTC
  CTGCTGTACAACTGAGATCATGCTGCATGCCAtTTTCAAC
            TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC
```

(a) Superposition of the seven highlighted 8-mers from figure 4.2 (d).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Alignment** | | A | T | C | C | A | G | C | T |
| | | G | G | G | C | A | A | C | T |
| | | A | T | G | G | A | T | C | T |
| | | A | A | G | C | A | A | C | C |
| | | T | T | G | G | A | A | C | T |
| | | A | T | G | C | C | A | T | T |
| | | A | T | G | G | C | A | C | T |
| **Profile** | **A** | 5 | 1 | 0 | 0 | 5 | 5 | 0 | 0 |
| | **T** | 1 | 5 | 0 | 0 | 0 | 1 | 1 | 6 |
| | **G** | 1 | 1 | 6 | 3 | 0 | 1 | 0 | 0 |
| | **C** | 0 | 0 | 1 | 4 | 2 | 0 | 6 | 1 |
| **Consensus** | | A | T | G | C | A | A | C | T |

(b) The alignment matrix, profile matrix and consensus string formed from the 8-mers starting at positions $\mathbf{s} = (8, 19, 3, 5, 31, 27, 15)$ in figure 4.2 (d).

**Figure 4.3** From DNA sample, to alignment matrix, to profile, and, finally, to consensus string. If $\mathbf{s} = (8, 19, 3, 5, 31, 27, 15)$ is an array of starting positions for 8-mers in figure 4.2 (d), then $Score(\mathbf{s}) = 5 + 5 + 6 + 4 + 5 + 5 + 6 + 6 = 42$.

$(s_1, s_2, \ldots, s_t)$, with $1 \le s_i \le n - l + 1$. The $l$-mers starting at these positions can be compiled into a $t \times l$ *alignment matrix* whose $(i, j)$th element is the nucleotide in the $s_i + j - 1$th element in the $i$th sequence (fig. 4.3). Based on the alignment matrix, we can compute the $4 \times l$ *profile matrix* whose $(i, j)$th element holds the number of times nucleotide $i$ appears in column $j$ of the alignment matrix, where $i$ varies from $1$ to $4$. The profile matrix, or *profile*, illustrates the variability of nucleotide composition at each position for a particular choice of $l$-mers. For example, the positions $3$, $7$, and $8$ are highly conserved, while position $4$ is not. To further summarize the profile matrix, we can form a *consensus string* from the most popular element in each column of the alignment matrix, which is the nucleotide with the largest entry in the profile matrix. Figure 4.3 shows the alignment matrix for $\mathbf{s} = (8, 19, 3, 5, 31, 27, 15)$, the corresponding profile matrix, and the resulting consensus string ATGCAACT.

By varying the starting positions in $\mathbf{s}$, we can construct a large number of different profile matrices from a given sample. We need some way of grading them against each other. Some profiles represent high conservation of a pattern while others represent no conservation at all. An imprecise formulation of the Motif Finding problem is to find the starting positions $\mathbf{s}$ corresponding to the most conserved profile. We now develop a specific measure of conservation, or strength, of a profile.