1

DNA: The Molecule of Life

"All rising to great places is by a winding stair." – Francis Bacon

1.1 Introduction

Ever since ancient Greek times, man has suspected that the features of one generation are passed on to the next. It was not until Mendel's work on garden peas was recognized (see [69, 148]) that scientists accepted that both parents contribute material that determines the characteristics of their offspring. In the early 20^{th} century, it was discovered that *chromosomes* make up this material. Chemical analysis of chromosomes revealed that they are composed of both *protein* and *deoxyribonucleic acid*, or DNA. The question was, which substance carries the genetic information? For many years, scientists favored protein, because of its greater complexity relative to that of DNA. Nobody believed that a molecule as simple as DNA, composed of only four subunits (compared to 20 for protein), could carry complex genetic information.

It was not until the early 1950s that most biologists accepted the evidence showing that it is in fact DNA that carries the genetic code. However, the physical structure of the molecule and the hereditary mechanism was still far from clear.

In 1951, the biologist James Watson moved to Cambridge to work with a physicist, Francis Crick. Using data collected by Rosalind Franklin and Maurice Wilkins at King's College, London, they began to decipher the structure of DNA. They worked with models made out of wire and sheet metal in an attempt to construct something that fitted the available data. Once satisfied with their double helix model (Fig. 1.1), they published the paper [154] (also see [153]) that would eventually earn them (and Wilkins) the Nobel Prize for Physiology or Medicine in 1962.

1.2 The Structure and Manipulation of DNA

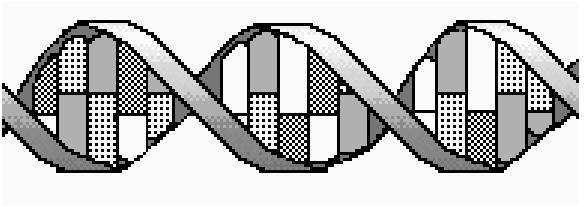

Fig. 1.1. Stylized depiction of DNA double helix

DNA (deoxyribonucleic acid) [1, 155] encodes the genetic information of cellular organisms. It consists of *polymer chains*, commonly referred to as DNA *strands*. Each strand may be viewed as a chain of *nucleotides*, or *bases*, attached to a sugar-phosphate "backbone." An n-letter sequence of consecutive bases is known as an n-mer or an *oligonucleotide*[1] of length n.

The four DNA nucleotides are adenine, guanine, cytosine, and thymine, commonly abbreviated to A, G, C, and T respectively. Each strand, according to chemical convention, has a 5' and a 3' end; thus, any single strand has a natural orientation. This orientation (and, therefore, the notation used) is due to the fact that one end of the single strand has a free (i.e., unattached to another nucleotide) 5' phosphate group, and the other end has a free 3' deoxyribose hydroxl group. The classical double helix of DNA (Fig. 1.2) is formed when two separate strands bond. Bonding occurs by the pairwise attraction of bases; A bonds with T and G bonds with C. The pairs (A,T) and (G,C) are therefore known as *complementary* base pairs. The two pairs of bases form *hydrogen bonds* between each other, two bonds between A and T, and three between G and C (Fig. 1.3).

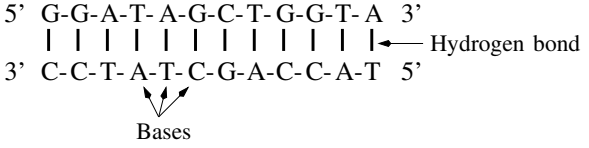

Fig. 1.2. Structure of double-stranded DNA

In what follows we adopt the following convention: if x denotes an oligo, then $\overline{x}$ denotes the complement of x. The bonding process, known as *annealing*,

[1] Commonly abbreviated to "oligo."

is fundamental to our implementation. A strand will only anneal to its complement if they have opposite polarities. Therefore, one strand of the double helix extends from 5' to 3', and the other from 3' to 5', as depicted in Fig. 1.2.

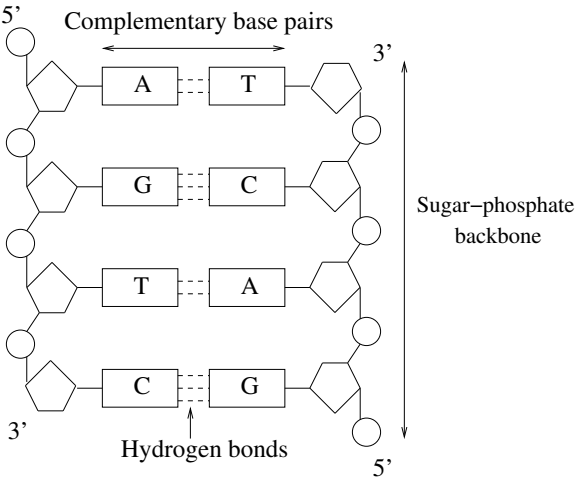

Fig. 1.3. Detailed structure of double-stranded DNA

1.3 DNA as the Carrier of Genetic Information

The *central dogma* of molecular biology [49] is that DNA produces RNA, which in turn produces proteins. The basic "building blocks" of genetic information are known as *genes*. Each gene codes for one specific *protein* and may be turned on (*expressed*) or off (*repressed*) when required.

Protein structure and function

Proteins are the working molecules in organisms, and the properties of living organisms derive from the properties of the proteins they contain. One of the most important functions proteins carry out is to act as *enzymes*. Enzymes act as specific *catalysts*; each type of enzyme catalyses a chemical reaction which would otherwise not take place at all, or at least take place very slowly. As an aside, the word "enzyme" is derived from the Greek for "in yeast", as all early work on enzymes was carried out on extracts of yeast [131]. The name of an enzyme indicates the type of reaction that it catalyses; for example, *restriction ligase* (see Sect. 1.4) catalyses the *ligation* of DNA strands (this process is described later). An organism's *metabolism* is defined as the the totality of the thousands of different chemical reactions occurring within it,

catalysed by thousands of different enzymes. Proteins also have many other different functions, such as messengers and structural components (human hair is made up of the protein keratin). So, what determines their specific properties?

This question may be answered thus: a protein's properties result from the sequence of *amino acids* that comprise it. Proteins are linear chains of amino acids, strung together rather like beads on a necklace. There are 20 different amino acids, and, given that proteins can be anything from 50 to 500 amino acids in length, the number of possible proteins is beyond astronomical. If we assume that the average protein is made up of 300 amino acids, there are 20^{300} possible protein sequences. This dwarfs the estimated number of fundamental particles in the observable universe (10^{80} [131]).

We now concern ourselves with how protein sequence determines form, or structure. Each amino acid in the chain has a particular pattern of attraction, due to its unique molecular structure. The chain of amino acids *folds* into a specific three-dimensional shape, or *conformation*. The protein *self-assembles* into this conformation, using only the "information" encoded in its sequence. Most enzymes assemble into a globular shape, with cavities on their surface. The protein's *substrate(s)* (the molecule(s) acted upon during the reaction catalysed by that protein) fit into these cavities (or *active sites*) like a key in a lock. These cavities enable proteins to bring their substrates close together in order to facilitate chemical reactions between them.

Hemoglobin is a very important protein that transports oxygen in the blood. A visualisation[2] of the hemoglobin protein [121] is depicted in Fig. 1.4. A particularly important feature of the protein is the labelled active site, where oxygen binds.

When the substrate binds to the active site the conformation of the protein changes. This can be seen in Fig. 1.5, which depicts the formation of a hexokinase-glucose complex. The hexokinase is the larger molecule; on the left a molecule of glucose is approaching an active site (the cleft) in the protein. Note that the conformation of the enzyme changes after binding (depicted on the right). This behavior, as we shall see later, may be used in the context of performing computations.

There are many modern techniques available to determine the structure and sequence of a given protein, but the problem of predicting structure from sequence is one of the greatest challenges in contemporary bioinformatics. The rules governing the folding of amino acid chains are as yet not fully understood, and this understanding is crucial for the success of future predictions.

Transcription and translation

We now describe the processes that determine the amino acid sequence of a protein, and hence its function. Note that in what follows we assume the

[2] All visualisations were created by the author, using the RasMol [138] molecular visualisation package available from http://www.umass.edu/microbio/rasmol/

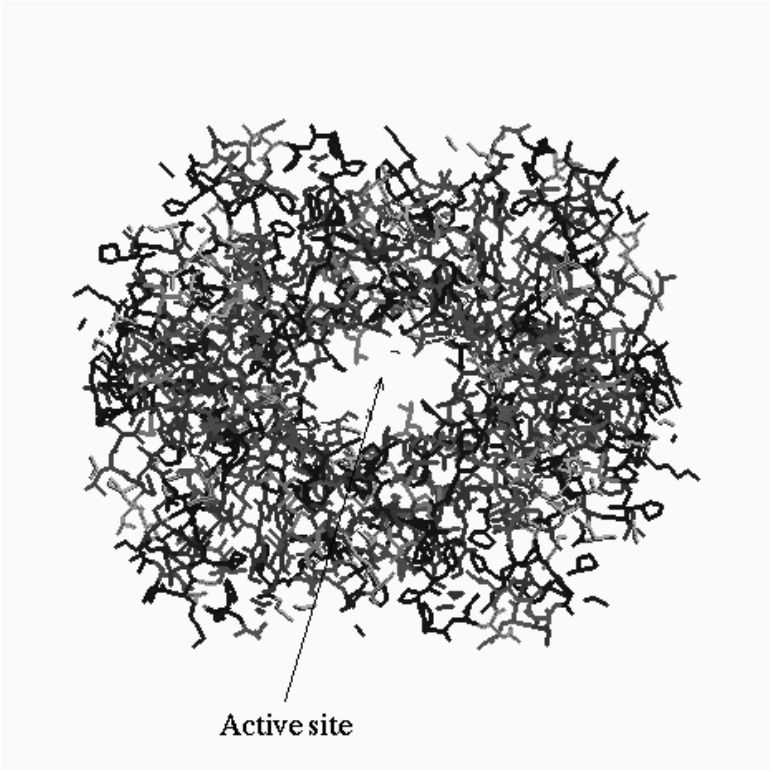

Fig. 1.4. Visualization of the hemoglobin protein

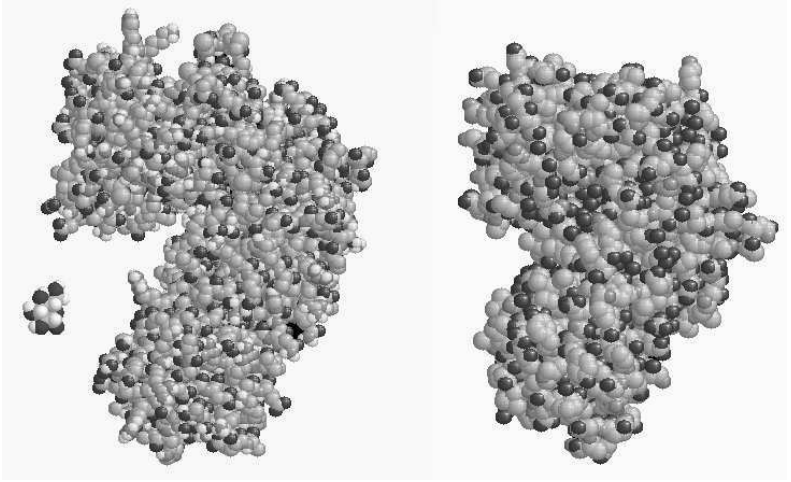

Fig. 1.5. Formation of a hexokinase-glucose complex

processes described occur in bacteria, rather than in higher organisms such as humans. In order for a DNA sequence to be converted into a protein molecule, it must be read (*transcribed*) and the transcript converted (*translated*) into a protein. Transcription of a gene produces a *messenger RNA* (mRNA) copy, which can then be translated into a protein.

Transcription proceeds as follows. The mRNA copy is synthesized by an enzyme known as *RNA polymerase*. In order to do this, the RNA polymerase must be able to recognize the specific region to be transcribed. This specificity requirement facilitates the regulation of genetic expression, thus preventing the production of unwanted proteins. Transcription begins at specific sites within the DNA sequence, known as *promoters*. These promoters may be thought of as "markers", or "signs", in that they are not transcribed into RNA. The regions that *are* transcribed into RNA (and eventually translated into protein) are referred to as *structural* genes. The RNA polymerase recognizes the promoter, and transcription begins. In order for the RNA polymerase to begin transcription, the double helix must be opened so that the sequence of bases may be read. This opening involves the breaking of the hydrogen bonds between bases. The RNA polymerase then moves along the DNA *template* strand in the $3 \rightarrow 5$' direction. As it does so, the polymerase creates an *antiparallel* mRNA chain (that is, the mRNA strand is the equivalent of the Watson-Crick complement of the template). However, there is one significant difference, in that RNA contains uracil instead of thymine. Thus, in mRNA terms, "U binds with A."

The RNA polymerase moves along the DNA, the DNA re-coiling into its double-helix structure behind it, until it reaches the end of the region to be transcribed. The end of this region is marked by a *terminator* which, like the promoter, is not transcribed.

Genetic regulation

Each step of the conversion, from stored information (DNA), through mRNA (messenger), to protein synthesis (effector), is itself catalyzed by effector molecules. These effector molecules may be enzymes or other factors that are required for a process to continue (for example, sugars). Consequently, a loop is formed, where products of one gene are required to produce further gene products, and may even influence that gene's own expression. This process was first described by Jacob and Monod in 1961 [82], and described in further detail in Chap. 6.

1.4 Operations on DNA

Some (but not all) DNA computations apply a specific sequence of biological operations to a set of strands. These operations are all commonly used by molecular biologists, and we now describe them in more detail.

Synthesis

Oligonucleotides may be synthesized to order by a machine the size of a microwave oven. The synthesizer is supplied with the four nucleotide bases in solution, which are combined according to a sequence entered by the user. The instrument makes millions of copies of the required oligo and places them in solution in a small vial.

Denaturing, annealing, and ligation

Double-stranded DNA may be dissolved into single strands (or *denatured*) by heating the solution to a temperature determined by the composition of the strand [35]. Heating breaks the hydrogen bonds between complementary strands (Fig. 1.6). Since a $G-C$ pair is joined by three hydrogen bonds, the temperature required to break it is slightly higher than that for an $A-T$ pair, joined by only two hydrogen bonds. This factor must be taken into account when designing sequences to represent computational elements.

Annealing is the reverse of melting, whereby a solution of single strands is cooled, allowing complementary strands to bind together (Fig. 1.6).

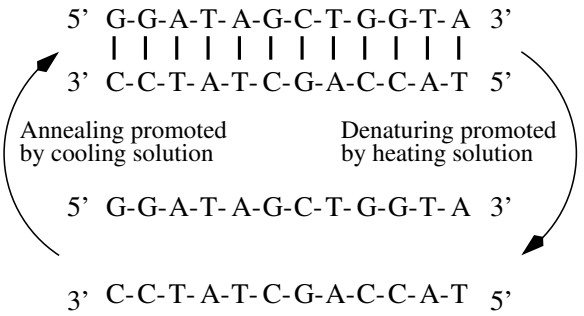

Fig. 1.6. DNA melting and annealing

In double-stranded DNA, if one of the single strands contains a discontinuity (i.e., one nucleotide is not bonded to its neighbor) then this may be repaired by DNA *ligase* [37]. This allows us to create a unified strand from several strands bound together by their respective complements. For example, Fig. 1.7a depicts three different single strands that many anneal, with a discontinuity where the two shorter strands meet. This may be repaired by the DNA ligase (Fig. 1.7b), forming a unified double-stranded complex (Fig. 1.7c).

Separation of strands

Separation is a fundamental operation, and involves the extraction from a test tube of any *single* strands containing a specific short sequence (e.g.,

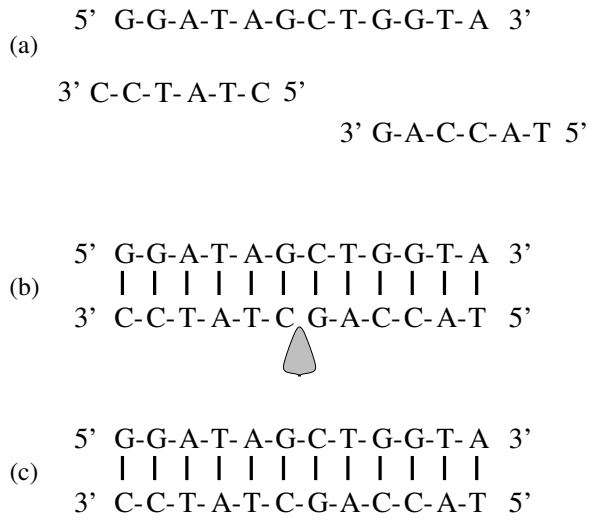

Fig. 1.7. (a) Three distinct strands. (b) Ligase repairs discontinuity. (c) The resulting complex

extract all strands containing the sequence $GCTA$). If we want to extract single strands containing the sequence x, we may first create many copies of its complement, $\overline{x}$. We attach to these oligos biotin molecules,[3] which in turn bind to a fixed matrix. If we pour the contents of the test tube over this matrix, strands containing x will anneal to the anchored complementary strands. Washing the matrix removes all strands that did not anneal, leaving only strands containing x. These may then be removed from the matrix.

Another removal technique involves the use of *magnetic bead separation*. Using this method, we again create the complementary oligos, but this time attach to them tiny magnetic beads. When the complementary oligos anneal to the target strands (Figure 1.8a), we may use a magnet to pull the beads out of the solution with the target strands attached to them (Fig. 1.8b).

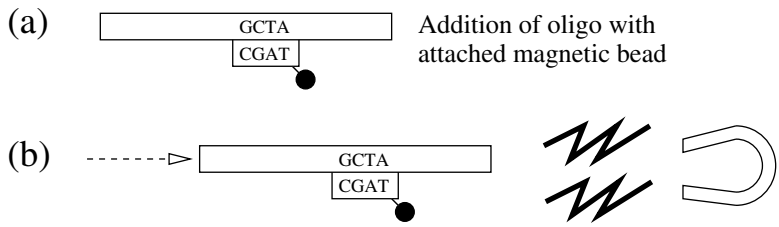

Fig. 1.8. Magnetic bead separation

[3] This process is referred to as "biotinylation".

Gel electrophoresis

Gel electrophoresis is an important technique for sorting DNA strands by size [37]. Electrophoresis is the movement of charged molecules in an electric field. Since DNA molecules carry a negative charge, when placed in an electric field they tend to migrate toward the positive pole. The rate of migration of a molecule in an *aqueous* solution depends on its shape and electric charge. Since DNA molecules have the same charge per unit length, they all migrate at the same speed in an aqueous solution. However, if electrophoresis is carried out in a *gel* (usually made of agarose, polyacrylamide, or a combination of the two), the migration rate of a molecule is also affected by its *size*.[4] This is due to the fact that the gel is a dense network of pores through which the molecules must travel. Smaller molecules therefore migrate faster through the gel, thus sorting them according to size.

A simplified representation of gel electrophoresis is depicted in Fig. 1.9. The DNA is placed in a well cut out of the gel, and a charge applied.

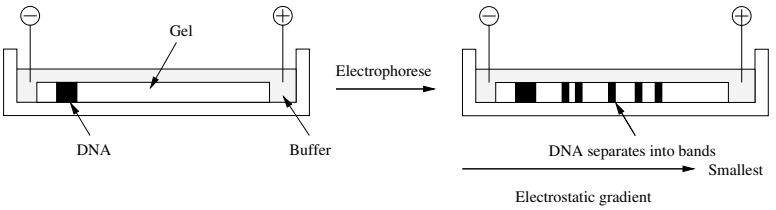

Fig. 1.9. Gel electrophoresis process

Once the gel has been run (usually overnight), it is necessary to visualize the results. This is achieved by staining the DNA with the fluorescent dye ethidium bromide and then viewing the gel under ultraviolet light. At this stage the gel is usually photographed.

One such photograph is depicted in Fig. 1.10. Gels are interpreted as follows; each *lane* (1–7 in our example) corresponds to one particular sample of DNA (we use the term *tube* in our abstract model). We can therefore run several tubes on the same gel for the purposes of comparison. Lane 7 is known as the *marker lane*; this contains various DNA fragments of known length, for the purpose of calibration. DNA fragments of the same length cluster to form visible horizontal *bands*, the longest fragments forming bands at the top of the picture, and the shortest ones at the bottom. The brightness of a particular band depends on the amount of DNA of the corresponding length present in the sample. Larger concentrations of DNA absorb more dye, and therefore

[4] Migration rate of a strand is inversely proportional to the logarithm of its molecular weight [114].

appear brighter. One advantage of this technique is its sensitivity – as little as 0.05 μg of DNA in one band can be detected as visible fluorescence.

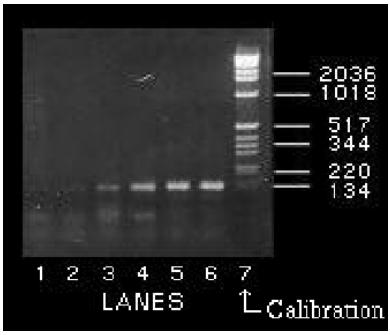

Fig. 1.10. Gel electrophoresis photograph

The size of fragments at various bands is shown to the right of the marker lane, and is measured in *base pairs* (b.p.). In our example, the largest band resolvable by the gel is 2,036 b.p. long, and the shortest one is 134 b.p. long. Moving right to left (tracks 6–1) is a series of PCR reactions which were set up with progressively diluted target DNA (134 b.p.) to establish the sensitivity of a reaction. The dilution of each tube is evident from the fading of the bands, which eventually disappear in lane 1.

Primer extension and PCR

The DNA *polymerases* perform several functions, including the repair and duplication of DNA. Given a short *primer* oligo, p in the presence of nucleotide triphosphates (i.e., "spare" nucleotides), the polymerase extends p if and only if p is bound to a longer *template* oligo, t. For example, in Fig. 1.11a, p is the oligo TCA which is bound to t, $ATAGAGTT$. In the presence of the polymerase, p is extended by a complementary strand of bases from the 5' end to the 3' end of t (Figure 1.11b).

Another useful method of manipulating DNA is the *Polymerase Chain Reaction*, or PCR [111, 112]. PCR is a process that quickly amplifies the amount of DNA in a given solution. Each cycle of the reaction doubles the quantity of each strand, giving an exponential growth in the number of strands.

PCR employs polymerase to make copies of a specific region (or *target sequence*) of DNA that lies between two *known* sequences. Note that this target sequence (which may be up to around 3,000 b.p. long) can be unknown ahead of time. In order to amplify template DNA with known regions (perhaps at either end of the strands), we first design forward and backward primers (i.e. primers that go from 5' to 3' on each strand. We then add a large excess (relative to the amount of DNA being replicated) of primer to the solution and heat

Fig. 1.11. (a) Primer anneals to longer template. (b) Polymerase extends primer in the 5' to 3' direction

it to denature the double-stranded template (Fig. 1.12a). Cooling the solution then allows the primers to anneal to their target sequences (Fig. 1.12b). We then add the polymerase. In this case, we use *Taq* polymerase derived from the thermophilic bacterium *Thermus aquaticus*, which lives in hot springs. This means that they have polymerases that work best at high temperatures, and that are stable even near boiling point (*Taq* is reasonably stable at 94 degrees Celsius). The implication of this stability is that the polymerase need only be added once, at the beginning of the process, as it remains active throughout. This facilitates the easy automation of the PCR process, where the ingredients are placed in a piece of apparatus known as a *thermal cycler*, and no further human intervention is required.

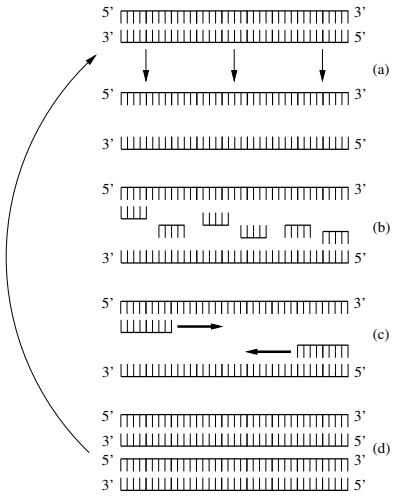

Fig. 1.12. (a) Denaturing. (b) Primer annealing. (c) Primer extension. (d) End result

The polymerase then extends the primers, forming an identical copy of the template DNA (Fig. 1.12c). If we start with a single template, then of course we now have two copies (Fig. 1.12d). If we then repeat the cycle of heating, annealing, and polymerising, it is clear that this approach yields an exponential number of copies of the template. A typical number of cycles would be perhaps 35, yielding (assuming a single template) around 68 billion copies of the *target sequence* (for example, a gene).

Unfortunately, the incredible sensitivity of PCR means that traces of unwanted DNA may also be amplified along with the template. We discuss this problem in a following chapter.

Restriction enzymes

Restriction endonucleases [160, page 33] (often referred to as *restriction enzymes*) recognize a specific sequence of DNA known as a *restriction site*. Any DNA that contains the restriction site within its sequence is cut by the enzyme at that point.[5]

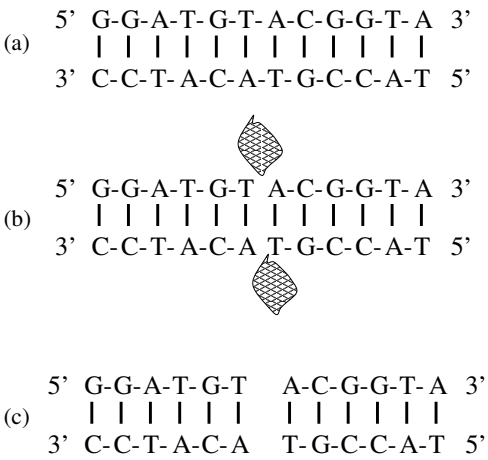

Fig. 1.13. (a) Double-stranded DNA. (b) DNA being cut by *Rsa*AI. (c) The resulting blunt ends

For example, the double-stranded DNA in Fig. 1.13a is cut by restriction enzyme *Rsa*I, which recognizes the restriction site $GTAC$. The enzyme breaks (or "cleaves") the DNA in the middle of the restriction site (Fig. 1.13b). The exact nature of the break produced by a restriction enzyme is of great importance. Some enzymes like *Rsa*I leave "blunt" ended DNA (Fig. 1.13c).

[5] In reality, only certain enzymes cut specifically at the restriction site, but we take this factor into account when selecting an enzyme.

Others may leave "sticky" ends. For example, the double-stranded DNA in Fig. 1.14a is cut by restriction enzyme *Sau*3AI, which recognizes the restriction site *GATC* (Fig. 1.14b). The resulting sticky ends are so-called because they are then free to anneal to their complement.

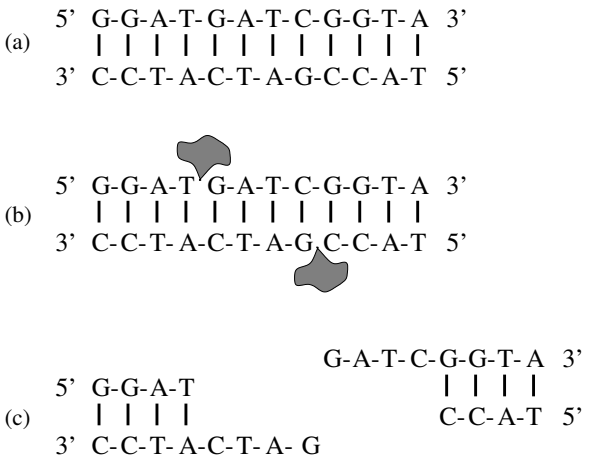

Fig. 1.14. (a) Double-stranded DNA being cut by *Sau*3AI. (b) The resulting sticky ends

Cloning

Once the structure of the DNA molecule was elucidated and the processes of transcription and translation were understood, molecular biologists were frustrated by the lack of suitable experimental techniques that would facilitate more detailed examination of the genetic material. However, in the early 1970s, several techniques were developed that allowed previously impossible experiments to be carried out (see [36, 114]). These techniques quickly led to the first ever successful cloning experiments [81, 102].

Cloning is generally defined as "... the production of multiple identical copies of a single gene, cell, virus, or organism." [130]. In the context of molecular computation, cloning therefore allows us to obtain multiple copies of specific strands of DNA. This is achieved as follows:

The specific sequence is inserted in a circular DNA molecule, known as a *vector*, producing a *recombinant DNA molecule*. This is performed by cleaving both the double-stranded vector DNA and the target strand with the *same* restriction enzyme(s). Since the vector is double stranded, restriction with suitable enzymes produces two short single-stranded regions at either end of the molecule (referred to as "sticky" ends. The same also applies to the target strand. The insertion process is depicted in Fig. 1.16. The vector and

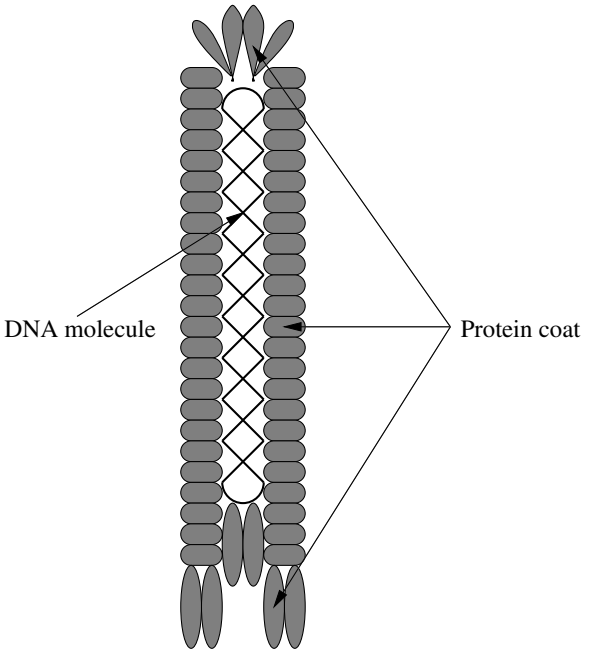

Fig. 1.15. Schematic representation of the M13 phage structure

target are both subjected to restriction; then, a population of target strands is introduced to the solution containing the vector. The sticky ends of the target bind with the sticky ends of the vector, integrating the target into the vector. After ligation, new double-stranded molecules are present, each containing the new target sequence.

In what follows, we use the *M13 bacteriophage* as the cloning vector. Specifically, we use the M13mp18 vector, which is a 7,249 b.p. long derivative of M13 constructed by Yanisch-Perron et al. [164].

Bacteriophages (or *phages*, as they are commonly known) are viruses that infect bacteria. The structure of a phage is very simple, usually consisting of a single-stranded DNA molecule surrounded by a sheath of protein molecules (the *capsid*) (Fig. 1.15).

The vector acts as a *vehicle*, transporting the sequence into a *host* cell (usually a bacterium, such as *E.coli*). In order for this to occur, the bacteria must be made *competent*. Since the vectors are relatively heavy molecules, they cannot be introduced into a bacterial cell easily. However, subjecting *E.coli* to a variety of hot and cold "shocks" (in the presence of calcium, among other chemicals) allows the vector molecules to move through the cell membrane. The process of introducing exogenous DNA into cells is referred to as *transformation*. One problem with transformation is that it is a rather inefficient process; the best we can hope for is that around 5% of the bacterial cells will

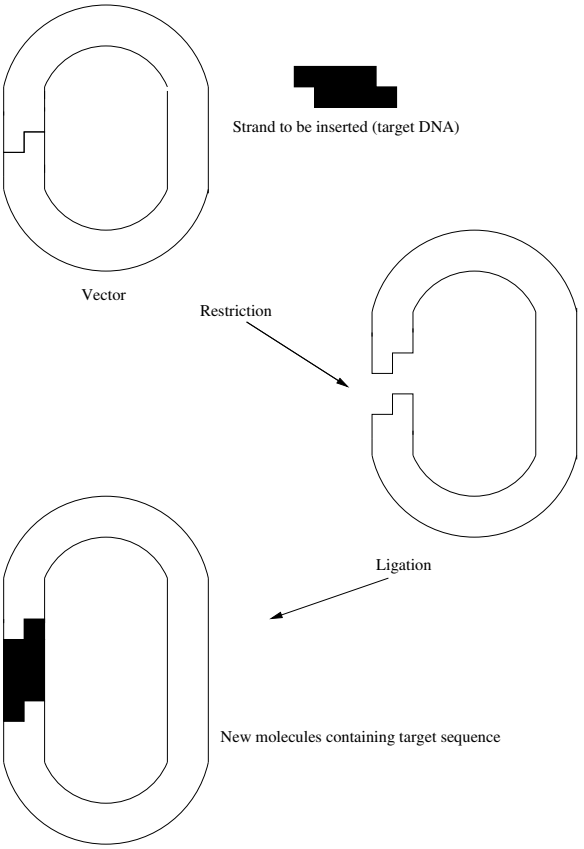

Fig. 1.16. Insertion of target strand into vector DNA

take up the vector. In order to improve this situation, we may use a technique known as *electroporation*. A high voltage pulse is passed through the solution containing the vectors and bacteria, causing the cell membranes to become permeable. This increases the probability of vector uptake. The vector then multiplies within the cell, producing numerous copies of itself (including the inserted sequence).

The infection cycle of M13 proceeds as follows. The phage attaches to a *pilus* (an appendage on the surface of the cell) and injects its DNA into the bacterium (Fig. 1.17a). The M13 DNA is not integrated into the DNA of the bacterium, but is still replicated within the cell. In addition, new phages are continually assembled within and released from the cell (Fig. 1.17b), which go on to infect other bacteria (Fig. 1.17c). When sufficient copies of the specific sequence have been made, the single-stranded M13 DNA may be retrieved from the medium. The process by which this is achieved is depicted in Fig. 1.18

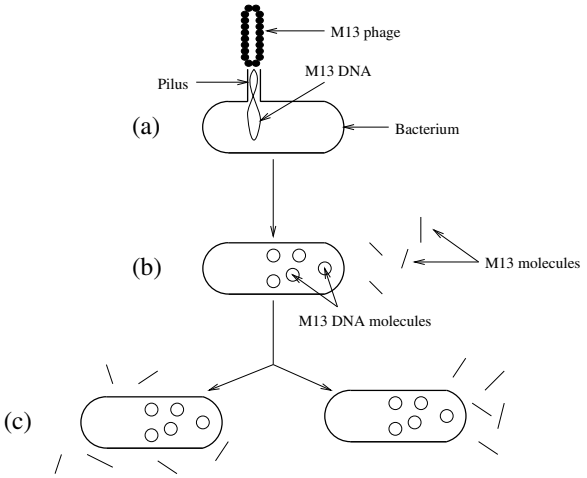

Fig. 1.17. M13 phage infection cycle

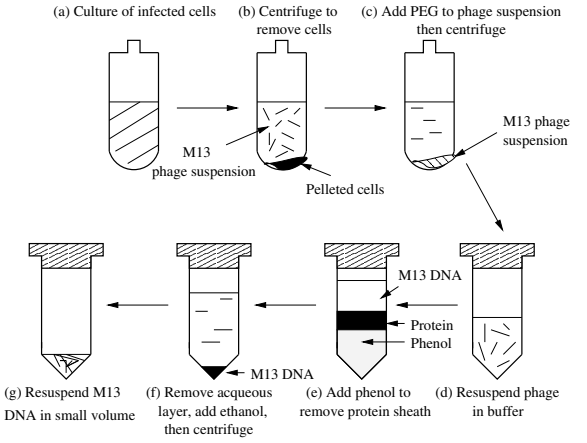

Fig. 1.18. Preparation of M13 DNA from infected culture of bacteria

(see also [104]). Once a sufficient volume of infected culture has been obtained we centrifuge it to pellet the bacteria (i.e., separate the bacteria from the phage particles). We then precipitate the phage particles with polyethylene glycol (PEG), add phenol to strip off the protein coats and then precipitate the resulting DNA using ethanol.

1.5 Summary

We described here the basic structure of DNA and the methods by which it may be manipulated in the laboratory. These techniques owe their origin to, and are being constantly improved by, the wide interests of molecular biologists working in modern areas such as the Human Genome project and genetic engineering. In Chap. 5 we show how these techniques allow us to implement a computation. Although other molecules (such as proteins) may be used as a computational substrate in the future, the benefit of using DNA is that this wide range of manipulation techniques is already available for use.

1.6 Bibliographical Notes

A definitive review of molecular genetics was co-authored by James Watson, one of the discoverers of the structure of DNA [156]. For in-depth information on molecular biology techniques, [18] is a laboratory manual that presents shortened versions of some 220 protocols selected from *Current Protocols in Molecular Biology*, the standard source in the field. For further details of the cloning process, the reader is directed to [136].

http://www.springer.com/978-3-540-65773-6

Theoretical and Experimental DNA Computation

Amos, M.

2005, XIII, 173 p., Hardcover
ISBN: 978-3-540-65773-6