A Computer Scientist's Guide to Molecular Biology*

Lila Kari[1], Rob Kitto[1], Greg Gloor[2]
[1]Department of Computer Science
University of Western Ontario
London, Ontario, ON, N6A 5B7 Canada
lila, kitto@csd.uwo.ca
[2] Department of Biochemistry
University of Western Ontario
London, ON, N6A 5C1 Canada
ggloor@julian.uwo.ca

Abstract

In this paper we explain the basic structure and properties of both single- and double-stranded DNA *in vivo* (in living organisms). We also review the first *in vitro* (test tube) experiment that solved a mathematical problem, The Directed Hamiltonian Path Problem, by manipulating DNA strands. Lastly, we give a list of *bio-operations* that have so far been used in DNA computation.

KEYWORDS: *in vivo* DNA, *in vitro* DNA computing, bio-operations

1 Introduction

Single-stranded DNA is a string built of a succession of four different building blocks, called bases. The bases are abbreviated *A, C, G, T* and they are the letters of the alphabet used to write the genetic information that dictates the growth and development of virtually all living beings. In the same way that the English alphabet can be used to write either a Shakespeare play or an instruction manual for a cell-phone, the order in which the bases are arranged on a DNA strand will determine its information content. This information content will, in turn, determine whether the organism developing from it will be, say, a mouse or a human being.

*Research partially supported by Grant R2824A01 of the Natural Sciences and Engineering Research Council of Canada.

Lately, the DNA alphabet has been used for a different purpose: to encode inputs, intermediate data and outputs to computations. Indeed, the four DNA letters are enough to express any information under a suitable encoding, and in November 1994 Adleman [1] reported the first experiment that used these possibilities for computational purposes.

Adleman's experiment, solving a mathematical problem by manipulations of DNA strands in test tubes, marked the begining of a new field known under the name of *molecular computing, biomolecular computing* or *DNA computing.* Being a field at the crossroads of computer science and molecular biology, research in this field requires basic knowledge of both. The present paper is meant to assist a computer scientist interested in DNA computing. By its very nature, the paper is schematic rather than complete, and tries to compress in a few pages information that could easily fill volumes. Rather than writing an exhaustive exposition of all details, our aim has been to find the highest-level possible description that still includes the notions necessary for a basic understanding of DNA and how it works.

Section 2 explains the "syntax" of a DNA strand, i.e. the structure of single- and double-stranded DNA. Section 3 aims to explain the "semantics" of a DNA strand, i.e. how the cell extracts the genetic information out of the DNA sequences and how it uses this information to synthesize proteins. Section 4 briefly outlines Adleman's in vitro experiment. Finally, Section 5 lists the molecular biology techniques that have so far been proposed and actually used for computational purposes.

For further information on DNA computing and molecular biology the reader is referred to [20], [11], [17], [18], [16], [7], [24].

2 The DNA molecule

DNA is the genetic material that dictates the growth and development of virtually all life on this planet. Although DNA has been shaping life on Earth for billions of years, our knowledge of its form and structure has developed almost exclusively in the last half century. An early insight into the nature of our genetic material was Schrödinger's 1944 observation that our genes must essentially be an aperiodic crystal [22]. He argued that genes need a crystalline structure to give them their apparent strength and stability, but that they must also be aperiodic to be able to store the tremendous amount of information that they do. In the early 1950's, James Watson, Francis Crick and Maurice Wilkins claimed that DNA *in vivo* (in natural state) has a double-helical shape [23]. This double helix both fits Schrödinger's prediction of an aperiodic crystal and is the actual form of the molecule.

A double helix of DNA (**D**eoxyribo-**N**ucleic **A**cid) is made from two single strands of DNA, each of which is a chain of *nucleotides.* A nucleotide is an organic structure with three basic parts: a phosphate group, a 5-carbon sugar group, and a nitrogenous side group. The nitrogenous side group is more commonly called a *base.* Four different nucleotides occur in DNA: adenosine (containing an *adenine* base), guanosine (containing a *guanine* base), thymidine (containing a *thymine* base) and cytidine (containing a *cytosine* base), abbreviated A, G, T, and C respectively. The four DNA nucleotides differ only in their bases - the sugar and phosphate groups are the same for each. Nucleotides can be joined together in a linear chain to form a single strand of DNA. A short single strand of DNA consisting of up to 20 nucleotides is called an *oligonucleotide*.

An oligonucleotide has a backbone of alternating sugar and phosphate groups with one of the four bases bound to each sugar group. The sugar group is made up of atoms joined in a ring-like pentagonal structure. One corner on this ring contains an oxygen atom, while the other four contain carbon atoms, labelled $1'$ through $4'$ clockwise starting from the oxygen atom. The $4'$ carbon atom on the ring is bound to a 5th carbon atom that is labelled $5'$. To summarize, the carbon atoms are labelled $1'$ through $5'$ depending on their positions. New molecules bind to carbon atoms at different locations and, therefore, the labels are also used to show where a particular molecule binds to the sugar ring. For example, a $5'$ phosphate group is a phosphate group bound to the $5'$ carbon atom of the sugar ring. The backbone of alternating sugar and phosphate groups gives an oligonucleotide a *polarity*, i.e. it has two distinct ends. One end is a $5'$ phosphate group, and the other end is a sugar group. In DNA, the sugar group is $2'$ deoxyribose (a deoxyribose which does not have a hydroxyl group OH linked to the $2'$ carbon). The last sugar group in the nucleotide chain contains a $3'$ hydroxyl group and, therefore, this end of a DNA strand is called the $3'$ end.

An oligonucleotide can be formally represented by its sequence of bases and its polarity, i.e. by a string of letters from the alphabet $\{A, G, C, T\}$. For example, $5'CAG$ is a DNA strand where nothing is bound to the phosphate group of a cytidine molecule, the sugar group of the cytidine molecule is bound to the phosphate group of an adenosine molecule, the sugar group of that adenosine is bound to the phosphate group of a guanosine molecule, and nothing, except a guanine base, is bound to the sugar group of the guanosine molecule. Although this oligonucleotide could be represented equivalently by $3'GAC$, the convention is to use the $5'$-sequence representation.

Each nucleotide in DNA has a complement: A and T are complementary, and G and C are complementary. Oligonucleotides also have complements, resulting from the complementarity of their respective nucleotides. A given oligonucleotide is complementary to another oligonucleotide with opposite polarity in which opposing bases are complementary. Thus, for example, the complement of $5'CAG$ is $3'GTC$, and the complement of $5'TGGC$ is $3'ACCG$.

Complementary oligonucleotides are attracted to each other. Under the right conditions (temperature, PH, oligonucleotide length etc.), two complementary oligonucleotides form a double-stranded DNA molecule. The two strands of DNA are held together by bonds between the different bases, and the two sugar-phosphate backbones wind around the outside of the bound bases in a double helix in which the two backbones run in opposite, but parallel directions. The bonds holding the complementary nucleotides together are a combination of Van der Waals (i.e. hydrophobic) forces and hydrogen bonds, with two hydrogen bonds forming between each A-T pair, and three hydrogen bonds between each C-G pair.

Ribonucleic acid (RNA) is another genetic molecule which is similar to DNA. RNA is also a chain of nucleotides, and a strand of RNA has both a $3'$ end and a $5'$ end. The sugar group of an RNA nucleotide contains one more oxygen atom than the sugar group of a DNA nucleotide, located at the $2'$ position. The phosphate group in an RNA nucleotide is identical to the phosphate group in a DNA nucleotide, and three of the four RNA bases, adenine, cytosine, and guanine, are identical to their DNA counterparts. The fourth RNA base is *uracil*, abbreviated U, which is similar to, but distinct from, thymine.

In RNA, C is complementary to G, and U is complementary to A and G. Furthermore, a strand of RNA can bind to a strand of DNA in the same way that two complementary strands of DNA bind, except that in RNA U takes the place of T as the complement of A, and is also complementary to G. RNA is important for understanding how meaning is extracted from DNA, which will be explained in the next section.

In eukaryotic cells, DNA is found in the nucleus of a cell - a small region towards the centre of the cell, separated from the rest of the cell by a thin membrane. Prokaryotic cells do not have a separate membrane-bound nucleus. RNA is found both in the nucleus and in the cytoplasm - the region of the cell where protein synthesis occurs.

3 The natural information content of DNA

The previous section outlined the structure of the DNA molecule, and noted that a DNA molecule can be represented by a string over the alphabet $\{A, G, C, T\}$ with a polarity. Although strings over $\{A, G, C, T\}$ (with a polarity) do unambiguously represent DNA molecules, this representation doesn't really show the natural meaning or function of a strand of DNA. How a cell actually uses its DNA is a complicated procedure, and we will only give a high-level overview of the process here.

According to the latest theories, there are two broad classes of DNA interspersed along a DNA strand representing a human chromosome: *spacer* DNA and *coding* DNA. Although roughly 95% of our DNA is spacer DNA, we will concentrate on the coding DNA. The reason is that the coding DNA is the region where the genetic information used to produce proteins or RNA molecules is stored, whereas at least some of the spacer DNA is used to regulate the flow of this information. To use the analogy presented in [18], DNA can be viewed as a text, where the contents of this text describe how an organism will develop. The words of this text are the coding DNA, and the spacer DNA works as molecular punctuation, dictating which words will be read and translated at a particular time and place. Using this metaphor, we can say that spacer DNA is also the the binding, end pieces, chapter headings and list of abbreviations of the book.

The meaning of a word of this text (i.e. of a section of coding DNA) is given by the sequence of the RNA or protein molecule it encodes. These molecules fold into three dimensional shapes that determine their function. Since translating a section of DNA into a protein molecule involves making a strand of RNA, we will focus on explaining how DNA is translated into a protein molecule. Proteins control virtually all inter- and intra- cellular activity, which in turn govern the development and maintenance of all cell-based organisms. Consequently, understanding how proteins are made from the templates given by the sections of coding DNA is essential for understanding the meaning and function of DNA *in vivo*.

We will describe this process in eukaryotic cells. The basic mechanisms are similar in prokaryotes. There are three main phases involved in translating a section of DNA into its associated protein. These phases are *transcription*, *editing and binding the transcript*, and *translation*. Before transcription can take place, it is necessary to activate the *regulatory region* beside the gene.

3.1 Transcription

The regulatory region is non-coding DNA which is activated by the binding of a specific combination of proteins. Without these proteins, the gene remains silent (i.e. cannot be translated). However, if the correct regulatory proteins are present in the cell, they bind to the regulatory region and this initiates the translation process.

The first stage is transcription, which essentially involves making a disposable copy, or *transcript*, of the gene, out of RNA. Once the regulatory region has been activated, the RNA polymerase[1] can begin its job of making a transcript.

The RNA polymerase starts at one end of the gene, and twists the double helix slightly to unwind it and expose about 12 bases, separating them from their counterparts on the other strand. This unwinding allows the RNA polymerase to begin forming a chain of RNA bases complementary to one of the two DNA strands by stepwise addition of nucleotides to the $3'$ end of the growing RNA strand. The RNA polymerase is structured such that it can travel along the gene, unwinding the double helix in front of it, and allowing the DNA to return to its normal shape behind it.

In this manner, the RNA polymerase moves along the gene, creating a single strand of RNA (the transcript) which is complementary to one of the strands of DNA in the gene. At the end of each gene there is a stop sequence, which forces the RNA polymerase to detach itself both from the DNA, which winds back into a complete double helix, and from the newly created transcript, which is now completed. This transcript therefore contains a copy of the sequence information which was present in the original gene.

3.2 Editing

While the transcript is still in the nucleus of the cell, it needs to be edited and bound, a process which turns it into *messenger RNA* (mRNA). The reason for

[1] RNA polymerase is an enzyme, already present in the nucleus of the cell, which makes the transcript.

this stage is that a given gene, and hence any transcript created from it, contains some spacer DNA, and may also contain some other information which is not useful for that particular cell. For example, a gene which encodes a protein necessary for muscle development might contain DNA sequences necessary for making several different types of muscle mass. This is an efficient use of DNA, since much of the information is the same regardless of what type of muscle is to be made, but unless the unnecessary parts are removed, the protein required by a particular cell is not made.

In the nucleus of each cell there is a collection of RNA-protein complexes called SNRPs (Small Nuclear Ribonucleo-Proteins). Each of these SNRPs binds to a specific RNA sequence and removes it, rejoining afterwards the two remaining pieces of the transcript. Thus, returning to the muscle example, the SNRPs would remove all the information specific to making proteins for other types of muscle development, as well as the parts of transcript which were copied from the spacer DNA in the original gene. The SNRPs would leave intact the generic protein coding information and the RNA sections relevant for muscle development in that particular cell at that particular time. This also shows that the eventual meaning of a section of DNA depends both on its actual sequence of base pairs, and on the environment (i.e. the proteins present) in the nucleus of its cell.

Before the SNRPs have edited out the sections of the transcript that are not useful in that particular cell, other proteins in the nucleus add a $5'$-chemical cap to one end of the transcript. Then after the editing, a string of As (a tail) is added to the other end of the transcript. This process of adding the cap and tail to the transcript is called *binding the transcript*. At this point, the edited and bound transcript, called messenger RNA (mRNA), can be moved out of the nucleus into the cytoplasm for the actual translation to begin.

The nucleus of the cell is separated from the cell body by a semi-permeable membrane. By itself, the mRNA cannot pass through this membrane into the cytoplasm. However, once it has been bound, certain proteins can attach to the cap and tail of the mRNA and move it through the membrane into the cytoplasm for translation.

3.3 Translation

The cytoplasm uses three types of molecules for translation: *ribosomes*, *amino acids*, and *transfer RNAs* (tRNA). In any healthy cell there is an ample supply of each of these molecules.

Translation is the cell's method for using the *genetic code* to form the protein (sequence of amino acids) specified by the mRNA. There are 64 possible three-nucleotide sequences of RNA (4x4x4), and each of these is called a *codon*. Three of these codons are stop codons, which indicate the end of the mRNA, and each of the other 61 codons correspond to one of the 20 amino acids. Thus, on average, there are about three different codons corresponding to each amino acid. This correspondence between the codons and their associated amino acids is called the *genetic code*.

The ribosome is the macro-molecule that assembles the chain of amino acids into the protein molecule which corresponds to a strand of mRNA. By themselves, amino acids do not bind to the mRNA; the transfer RNA (tRNA) is necessary to form a temporary bond between codons and their associated amino acids as the protein is synthesized by the ribosome.

Each tRNA binds simultaneously to a codon and one amino acid[2]. Many tRNA molecules can recognize more than one codon, but each codon corresponds to the same amino acid. Indeed, the enzyme *aminoacyl-tRNA synthetase* binds a tRNA to its associated amino acid, and it performs this binding in such a way that exactly three bases of the tRNA (called an *anticodon*) are exposed. The amino acid and the tRNA to which it is bound are referred to as an *aminoacyl tRNA*.

When the mRNA enters the cytoplasm, it is enveloped by the ribosome such that 6 mRNA bases (2 codons) are exposed. The ribosome holds the mRNA like this until two anticodons (on two aminoacyl tRNA molecules) have bound to each of these codons. When the anticodons on two tRNAs have bound to the two exposed codons on the mRNA, a bond forms between the amino acid molecules that are attached to each of the tRNAs. The chemical bond is formed by *peptidyl transferase* between the two amino acids such that the amino acid closer to the mRNA cap (the $5'$ end) is displaced from its tRNA and bound to the amino acid closer to the mRNA tail (the $3'$ end). The tRNA closer to the tail now joins the dipeptide (a chain of amino acids of length two) to the mRNA.

At this point, the ribosome shifts three bases along the mRNA, towards the mRNA tail. Because only 6 bases of the mRNA are exposed at any one time by the ribosome, this shift dislodges the first tRNA (which is no longer attached to an amino acid) from the mRNA, and exposes the next codon on the mRNA. The two amino acids are still bound to each other, and one of them is also attached to its tRNA, which is still attached to the mRNA. This is the beginning of the chain of amino acids which is being formed.

After the empty tRNA is detached from the mRNA, the anticodon on a third tRNA, which is attached to the appropriate amino acid, binds to the newly exposed codon on the mRNA. After the third tRNA binds to the mRNA, peptidyl transferase forms a bond between the dipeptide and the amino acid bound to the third tRNA which is closer to the mRNA tail. This forms a tripeptide bound to this tRNA.

The ribosome continues shifting along the mRNA, one codon at a time, until it encounters one of the three stop codons. Until then, the process is the same at each step: shifting one codon along the mRNA, dislodging one tRNA and exposing a new codon for a new tRNA to bind to. As the new tRNAs bind to the mRNA, their amino acids also bind to the growing chain of amino acids, and even when the tRNA is dislodged by the ribosome, its amino acid remains part of the chain. As the chain of amino acids grows, it begins folding into the

[2] Thus there are distinct tRNAs which correspond to the 61 coding codons and they ensure that the correct amino acid is bound to the codon specified by the genetic code.

three-dimensional shape which ultimately is the meaning of the original gene.

When the ribosome encounters one of the three stop codons in the mRNA (UGA, UAA or UAG), the end of translation is signaled. At this time, the chain of amino acids is detached from the tRNA, from the ribosome, and from the mRNA, and finishes folding into its three-dimensional shape. This chain of folded amino acids constitutes the protein that represents the actual meaning of the gene that was being translated, in the specific environment or context of the cell it was in. Depending on the protein created, it may stay in the cell where it was made and serve some purpose there, or it may travel to some other cell. In either case, once the string of amino acids detaches itself from the ribosome and folds into its shape, the translation of the gene is complete.

3.4 Summary

To briefly summarize, the three stages involved in extracting meaning from DNA *in vivo* are transcription, editing and binding the transcript, and translation. Before transcription can occur, certain regulatory proteins must bind to the regulatory region beside the gene. Once they do, RNA polymerase makes a temporary copy, or transcript, of the gene out of RNA. This transcript contains more information than is useful for any particular cell at a particular time, and certain sections must be removed. This is accomplished by enzymes in the cell nucleus, and constitute the editing and binding stage.

Finally, the edited and bound transcript, which is now mRNA, is moved out of the nucleus into the cytoplasm for translation. In this stage, the ribosome synthesizes a chain of amino acids which corresponds to the mRNA as specified by the genetic code. Once this chain of amino acids is complete, it separates from the ribosome, and finishes folding itself into a three-dimensional shape. This folded protein is the meaning of the original gene, in that environment.

4 Adleman's in vitro experiment

One of the most important motivations for computing with DNA comes from its size: a single nucleotide consists of approximately 50 atoms, and therefore a huge amount of information could be stored in very little DNA. Indeed, the DNA from a single human cell, which has a total length of 2 meters, is compactly packed in the $10^{-5}m$ diameter cell [7]. The human body consists of billions of individual cells and DNA is the control-centre of each of them. (In [5] it is estimated that a DNA-based memory of about 50 grams of DNA would have a memory capacity comparable to that of the human brain.)

The practical possibilities of encoding information in a DNA sequence and of manipulating DNA strands in vitro (in the test tube) were used in [1] to solve a 7-node instance of the Directed Hamiltonian Path Problem. A directed graph G with designated vertices v_{in} and v_{out} is said to have a Hamiltonian path if and only if there exists a sequence of compatible one–way edges $e_1, e_2, \ldots, e_z$

(that is, a path) that begins at v_{in}, ends at v_{out} and enters every other vertex exactly once.

The following (nondeterministic) algorithm solves the problem:

Step 1. Generate random paths through the graph.

Step 2. Keep only those paths that begin with v_{in} and end with v_{out}.

Step 3. If the graph has n vertices, then keep only those paths that enter exactly n vertices.

Step 4. Keep only those paths that enter all of the vertices of the graph at least once.

Step 5. If any paths remain, say YES; otherwise say NO.

In the following we shortly describe the DNA experiment that implements this algorithm. A detailed explanation of the molecular procedures involved will follow in Section 5.

To implement Step 1, each vertex of the graph was encoded into a random 20–nucleotide strand (20–letter sequence) of DNA that was *synthesized*. Then, for each (oriented) edge of the graph, a DNA sequence was synthesized consisting of the second half of the sequence encoding the source vertex and the first half of the sequence encoding the target vertex.

By *mixing* together single strands encoding the edges and single strands encoding complements of vertices, DNA sequences corresponding to compatible edges were linked together. Indeed, by construction, a complement of a vertex strand would bind to both a strand encoding an edge entering the vertex, and a strand encoding an edge exiting the vertex. Hence, the *ligation reaction* resulted in the formation of DNA molecules encoding random paths through the graph.

To implement Step 2, the product of Step 1 was amplified by *polymerase chain reaction (PCR)*. Thus, only those molecules encoding paths that begin with v_{in} and end with v_{out} were amplified.

For implementing Step 3, a technique called *gel electrophoresis* was used, that makes possible the separation of DNA strands by length. Thus, only molecules encoding paths of the desired length (any candidate to the Hamiltonian Path has to pass through each vertex and therefore has to have a certain length) were retained.

Step 4 was accomplished by iteratively using a process called affinity purification. This process permits the *extraction* from a pool of DNA strands of only those strands that contain a given pattern as a subsequence. By repeatedly applying this process, strands were retained that contained as a subsequence the encoding for the first vertex, second vertex, and so on, until only those paths that pass through all vertices remained.

To implement Step 5, the presence of a molecule encoding a Hamiltonian path was checked. This was done by amplifying the result of Step 4 by PCR and then determining the DNA sequence of the amplified molecules.

5 DNA as a computational tool

The preceding section showed how a mathematical problem can be solved by encoding information in DNA strands and using some molecular biology procedures to perform computational steps. In the following we list some of the molecular biology techniques that have so far been proposed or used for performing computations ([14], [12], [25], [10]):

– *Synthesizing* a desired polynomial–length strand, [1], [10]. In standard solid–phase DNA synthesis, a desired DNA molecule is built up nucleotide by nucleotide on a support particle in sequential coupling steps. For example, the first nucleotide (monomer), say A, is bound to a glass support. A solution containing C is poured in, and the A reacts with the C to form a two–nucleotide (2–mer) chain AC. After washing the excess C solution away, one could have the C from the chain AC coupled with T to form a 3–mer chain (still attached to the surface) and so on.

– *Mixing*: pour the contents of two test tubes into a third one to achieve union [2], [10]. Mixing can be performed by rehydrating the tube contents (if not already in solution) and then combining the fluids together into a new tube, by pouring and pumping for example.

– *Annealing*: bind together two single–stranded complementary DNA sequences by cooling the solution. (See [6], [26], [10].) Annealing in vitro is also known as *hybridization*.

– *Melting*: break apart a double–stranded DNA into its single–stranded complementary components by heating the solution. (See [6], [26], [10].) Melting in vitro is also known under the name of *denaturation*.

– *Extracting* those strands that contain a given pattern as a substring by using affinity purification [1], [10]. This process permits single strands containing a given subsequence v to be filtered out from a heterogeneous pool of other strands. After synthesizing strands complementary to v and attaching them to magnetic beads, the heterogeneous solution is passed over the beads. Those strands containing v anneal to the complementary sequence and are retained. Strands not containing v pass through without being retained.

– *Separating* the strands by size using gel electrophoresis [1], [10]. The molecules are placed at the top of a wet gel, to which an electric field is applied, drawing them to the bottom. Larger molecules travel more slowly through the gel. After a period, the molecules spread out into distinct bands according to size.

– *Cutting* DNA double strands at specific sites by using restriction enzymes. One class of enzymes, called *restriction endonucleases*, recognizes a specific short sequence of DNA, known as a *restriction site*. Any double-stranded DNA that contains the restriction site within its sequence is cut by the enzyme at that location. (See [8], [19], [10].)

– *Ligating*: paste DNA strands with compatible sticky ends by using DNA ligase. A DNA double-strand can either have blunt ends, i.e. be fully double-stranded or can be partially double-stranded, i.e. it can have single-stranded overhanging ends (called sticky ends) at one or both of its extremities. The enzyme *DNA ligase*, joins together, or ligates, the end of a DNA strand to another strand.

DNA ligase either ligates two blunt-ended double strands or two strands with compatible sticky ends. (See [8], [26], [10].)

– *Marking* single strands by hybridization: complementary sequences are attached to the strands, making them double–stranded. The reverse operation is *unmarking* of the double strands by denaturing, that is, by detaching the complementary strands. The marked sequences are double–stranded while the unmarked ones are single–stranded [4], [15], [21], [10].

– *Destroying* the marked strands by using exonucleases [15], [10]. Using enzymes called *exonucleases*, either double-stranded or single-stranded DNA molecules may be selectively destroyed. The exonucleases chew up DNA molecules from the end in, and exist with specificity to either single-stranded or double-stranded form. Another method for destroying marked strands is by cutting all the marked strands with a restriction enzyme and removing all the intact strands by gel electrophoresis [4], [10].

– *Amplifying (copying)*: make copies of DNA strands by using the Polymerase Chain Reaction (PCR), that uses the *DNA polymerase* enzyme [1], [10]. The *DNA polymerases* perform several functions including replication of DNA. The replication reaction requires a guiding DNA single-strand called *template*, and a shorter oligonucleotide called *primer*, that is annealed to it. Under these conditions, DNA polymerase catalyzes DNA synthesis by successively adding nucleotides to one end of the primer. The primer is thus extended in one direction until the desired strand that starts with the primer and is complementary to the template is obtained.

PCR is an in vitro method that relies on DNA polymerase to quickly amplify specific DNA sequences in a solution. PCR involves a repetitive series of temperature cycles, with each cycle comprising three stages: denaturation of the guiding template DNA to separate its strands, then cooling to allow annealing to the template of the primer oligonucleotides, which are specifically designed to flank the region of DNA of interest, and, finally, extension of the primers by DNA polymerase. Each cycle of the reaction doubles the number of target DNA molecules, the reaction giving thus an exponential growth of their number. (See [2], [4], [13].)

– *Substituting*: substitute, insert or delete DNA sequences by using *PCR site-specific oligonucleotide mutagenesis* (see [9], [10]). The process is a variation of PCR in which a change in the template can be induced by the process of primer modification. Namely, one can use a primer that is only partially complementary to a template fragment. (The modified primer should contain enough bases complementary to the template to make it anneal despite the mismatch.) After the primer is extended by the polymerase, the newly obtained strand consist of the complement of the template in which a few nucleotides have been substituted by other, desired ones.

– *Detecting* and *Reading*: given the contents of a tube, say YES if it contains at least one DNA strand, and NO otherwise [1], [2], [10]. PCR may be used to amplify the result and then a process called *sequencing* is used to actually read the DNA strands in solution. The basic idea of the most widely used sequencing method is to use PCR and gel electrophoresis. Assume we have a homogeneous solution, that is, a solution containing mainly copies of the strand we wish to sequence, and very few contaminants (other strands). For detection of the positions of As in the target strand, a blocking agent is used that prevents the templates from being extended beyond As during PCR. As a result of this modified PCR, a population of subsequences is obtained, each corresponding to a different occurrence of A in the original strand. Separating them by length using gel electrophoresis reveals the positions where A occurs in the strand. The process can then be repeated for each of C, G, and T, to yield the sequence of the strand. Recent methods use four different fluorescent dyes, one for each base, which allows all four bases to be processed simultaneously. As the fluorescent molecules pass a detector near the bottom of the gel, data are output directly to an electronic computer.

The procedures listed above (know also under the name of *bio–operations*), and possibly others, can then be used to write *biomolecular programs* which receive a tube containing DNA strands as input and return as output either YES or NO or a set of tubes. A computation consists of a sequence of tubes containing DNA strands.

Besides the novelty of the approach, and in spite of the technical difficulties that arise from the error rates of bio-operations [10], there are several reasons why computing with DNA might have advantages over electronic computing. These include memory capacity, massive parallelism, and power requirements. Indeed, one gram of DNA, which when dry would occupy a volume of approximately one cubic centimetre, can store as much information as approximately one trillion CDs [3]. Moreover, computing with DNA provides enormous parallelism. In Adleman's experiment, which was carried out in one fiftieth of a teaspoon of solution, [1], approximately 10^{14} oriented edges were simultaneously concatenated in about one second, [3]. It is not clear whether the fastest available supercomputer is capable of such a speed. Finally, as far as energy efficiency is concerned, in principle one joule is sufficient for approximately 2×10^{19} ligation operations, while existing supercomputers operate in the significantly smaller range of 10^9 operations per joule [3].

These practical incentives and the fascination of being able to perform computations with biological means have inspired many researchers to pursue the challenging topic of DNA computing. It is anticipated that the pioneer research in this field of intersection between computation and biology will have great significance in many aspects of science and technology. Indeed DNA computing sheds new light onto the very nature of computation, and opens vistas for computability models totally different from the classical ones. In an optimistic way, one may think of an analogy between the work of researchers in this area and the work on finding models of computation carried out in the 30s, which has laid the foundation for the design of today's electronic computers.

References

[1] L.Adleman. Molecular computation of solutions to combinatorial problems. *Science* v.266, Nov.1994, 1021–1024.

[2] L.Adleman. On constructing a molecular computer. [14], 1-22.

[3] L.Adleman. Computing with DNA. *Scientific American*, August 1998, 54–61.

[4] M.Amos, A.Gibbons, D.Hodgson. Error–resistant implementation of DNA computation. [12], 87–101.

[5] E.Baum. Building an associative memory vastly larger than the brain. *Science*, vol.268, April 1995, 583–585.

[6] D.Boneh, R.Lipton, C.Dunworth, J.Sgall. On the computational power of DNA. *Discrete Applied Math.*, 71(1996), 76-94.

[7] C.R.Calladine, H.R. Drew. *Understanding DNA: the molecule and how it works.* Academic Press, New York, 1999.

[8] T.Head. Formal language theory and DNA: an analysis of the generative capacity of recombinant behaviors. *Bull. Math. Biology*, 49(1987) 737–759.

[9] L.Kari, G.Thierrin. Contextual insertions/deletions and computability. *Information and Computation*, 131, 1(1996), 47-61.

[10] L.Kari. DNA computing - the arrival of biological mathematics. *The Mathematical Intelligencer*, vol.19, 2(1997), 9-22.

[11] E. Lander and M.S.Waterman, Eds. *Calculating the Secrets of Life*, National Academic Press, 1995.

[12] L.F.Landweber, E.B.Baum, Eds.: *DNA Based Computers II*, Proceedings of a DIMACS workshop, Princeton, 1996, American Math. Soc., 1998.

[13] T. Leete, M. Schwartz, R. Williams, D. Wood, J. Salem, H. Rubin. Massively parallel DNA computation: expansion of symbolic determinants. [12], 49–66.

[14] R.J.Lipton, E.M.Baum. Eds.: *DNA Based Computers I*, Proceedings of a DIMACS Workshop, Princeton, 1995, Amer. Math. Society, 1996.

[15] Q.Liu, Z.Guo, A.Condon, R.Corn, M.Lagally, L.Smith. A surface–based approach to DNA computation.[12], 206–216.

[16] J. Monod. *Chance and Necessity*, Alfred A. Knopf, 1971.

[17] D. Prescott and L. Goldstein, Eds. *Cell Biology: A Comprehensive Treatise - Volume 2: The Structure and Replication of Genetic Material*, Academic Press, 1979.

[18] R. Pollack. *Signs of Life*, Houghton Mifflin Company, 1994.

[19] P.Rothemund. A DNA and restriction enzyme implementation of Turing machines. [14], 75-120.

[20] G.Paun, G.Rozenberg, A.Salomaa. *DNA Computing: New Computing Paradigms*. Springer Verlag, Berlin, 1998.

[21] S. Roweis, E. Winfree, R. Burgoyne, N. Chelyapov, M. Goodman, P. Rothemund, L. Adleman. A sticker based architecture for DNA computation. [12], 1–27.

[22] E. Schrodinger. *What is Life?*. Cambridge, Eng.: Cambridge University Press, 1944.

[23] J.D. Watson, F.H.C.Crick. A structure for deoxyribose nucleic acid. *Nature*, April 25, 1953, 737-38.

[24] J.D.Watson, N.H.Hopkins, J.W.Roberts, J.Steitz, A.M.Weiner. *Molecular biology of the gene*, 5th Edition, Addison Wesley Longman, Harlow, 1998.

[25] E.Winfree, D.Gifford, Eds.: *DNA Based Computers V*. Proceedings of a DIMACS workshop, Amer. Math. Soc. Press. In press, 1999.

[26] E.Winfree. On the computational power of DNA annealing and ligation. [14], 199-210.