Chapter IV

Molecular Computation

A Basic concepts

A.1 Introduction to molecular computation

Molecular computation uses molecules to represent information and molecular processes to implement information processing. DNA computation is the most important (and most explored) variety of molecular computation, and it is the principal topic of this chapter. The principal motivation is that molecular computation is massively parallel (*molar* or *Avagadro-scale* parallelism, that is, on the order of 10^{26}). Data representations are also very dense (on the order of nanometers), and molecular computation can be very efficient in terms of energy.

A.2 DNA basics

The DNA molecule is a polymer (*poly* = many, *mer* = part), a sequence of *nucleotides*. A DNA nucleotide has three parts: a deoxyribose sugar, a phosphate group, and a nucleobase (C = cytosine, G = guanine, A = adenine, T= thymine).[1] See Fig. IV.1 for deoxyribose. The *DNA backbone* is a

[1] In RNA, uracil replaces thymine.

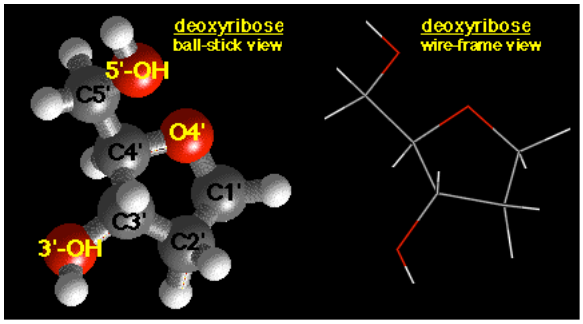

Figure IV.1: Deoxyribose. [source: wikipedia]

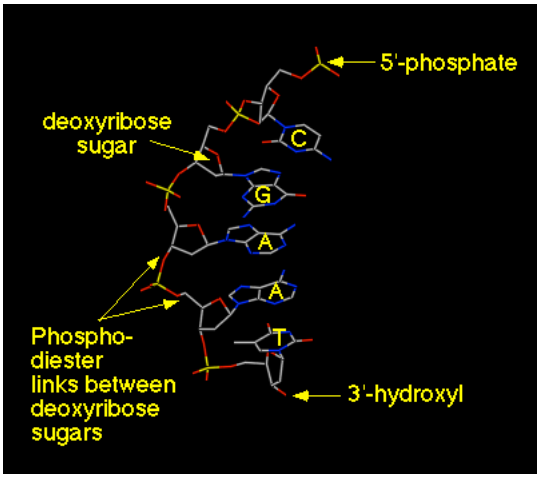

Figure IV.2: DNA backbone. [source: wikipedia]

sequence of deoxyribose sugars alternating with phosphate groups connected by covalent *phosphodiester* bonds (Fig. IV.2). The backbone connects the hydroxyl group on the $3'$ carbon (the "$3'$-hydroxyl group") of one sugar to the $5'$-hydroxyl of the next. In this way we distinguish the $5'$ and $3'$ ends of a polynucleotide: $5'$ has a terminal phosphate group, $3'$ a terminal hydroxl group. We generally write the bases in the $5'$ to $3'$ order.

Adenine and thymine each have two hydrogen-bonding sites and can bind together; similarly, guanine and cytosine each have three hydrogen-bonds and can bind together. This results in *Watson-Crick complementarity*, which means that two complementary polynucleotides can bind together. This can occur only if the two single strands are *antiparallel*, that is, run in opposite directions ($3'$ to $5'$ versus $5'$ to $3'$); see Fig. IV.3. Since we write the *sense-strand* in the $5'$ to $3'$ order, we write the *antisense-strand* in the $3'$ to $5'$ order. Complementary single stranded DNA can hybridize, leaving *sticky ends* (unmatched single strands) at either or both ends. They are called "sticky" because they are available for bonding by complementary sequences.

Fig. IV.4 gives some idea what the actual double helix looks like. It is worth keeping in mind that it is a very complex organic molecule, not a string of the letters A, C, T, G. For example, in addition to the double helix formed by the backbones of the two strands, there are two groves between the strands (Fig. IV.5). They are adjacent to the bases and therefore provide binding sites for proteins. Because the backbones are not symmetrically placed, the grooves have two sizes: *minor* (1.2nm) and *major* (2.2nm). The latter provides the best binding site.

To give some sense of the scale, here are some numbers. The nucleotides are about 3.3nm long. The chain is 2.2 to 2.6nm wide. The radius of the coil is about 1nm. The longest human chromosome (chromosome 1) is about 220×10^6bp (base pairs) long.

DNA data storage: In Aug. 2012 a team at Harvard reported converting a 53,000 word book to DNA and then reading it out by DNA sequencing.[2] This is the densest consolidation of data in any medium (including flash memory, but also compared to experimental media, such as quantum

[2]Church, George M, Gao, Yuan, and Kosuri, Sriram. "Next-generation digital information storage in DNA." *Science* **337** (28 Sept. 2012): 1628. See also *Science* (Aug. 16, 2012), `DOI: 10.1126/science.1226355` (accessed 2012-08-24). See also `http://spectrum.ieee.org/biomedical/imaging/reading-and-writing-a-book-with-dna/` (accessed 2012-08-24).

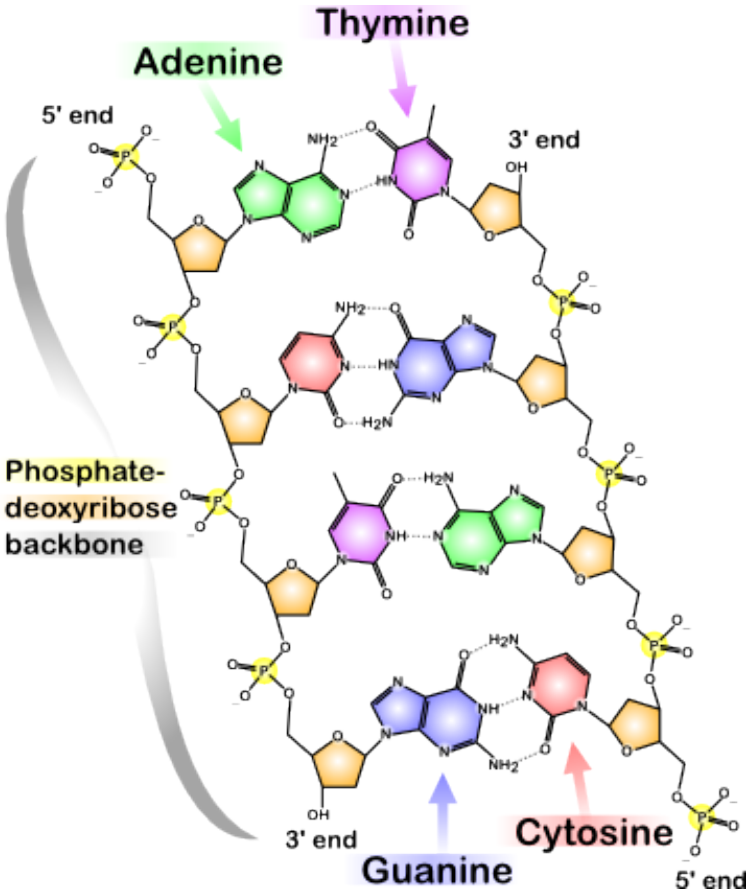

Figure IV.3: Watcon-Crick complementarity. Complementary H-bonds allow guanine to bind to cytosine and adenine to bind to thymine. [source: wikimedia commons]

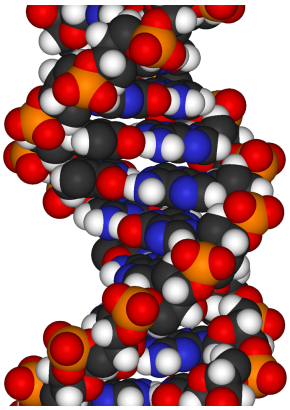

Figure IV.4: [source: wikimedia commons]

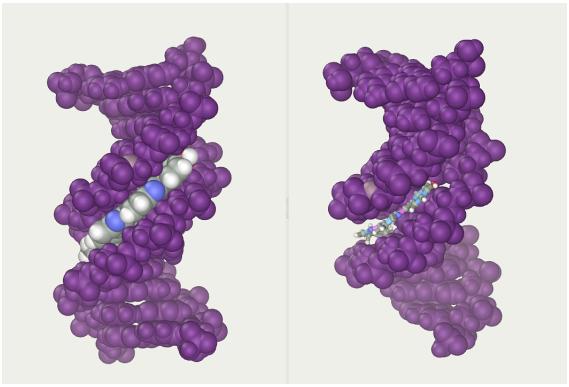

Figure IV.5: Major and minor grooves. [source: wikimedia commons]

holography). The book included 11 jpegs and a javascript program, for a total of 5.27 Mbits, and the decoded version had only 10 incorrect bits. The encoded book could be reproduced by DNA replication. This is a storage density of 5.5 petabits/mm^3 = 5.5×10^{15}bits/mm^3. The authors estimate that DNA can encode 455 exabytes (4.55×10^{20} bytes) per gram of single-stranded DNA. (This is based on 2 bits per nucleotide, and doesn't take redundancy, error correction, etc., into account, which would decrease the density by several orders of magnitude.[3]). In Jan. 2013 *Nature*, it was reported that 5 Mbits had been encoded and perfectly decoded (99.99 to 100% accuracy with quadruple redundancy. It cost $12,400 for the encoding and $220 for retrieval. The cost of DNA synthesis has been decreasing about 5× per year, and the cost of sequencing by about 12× per year, so it could become economical for archival storage. Enclosed in silica glass spheres, DNA storage could last for a million years.

A.3 DNA manipulation

There are several standard procedures for manipulating DNA that have been applied to DNA computing, which has benefitted from well-developed laboratory procedures and apparatus for handling DNA. The following sections review these procedures briefly, omitting details not directly relevant to DNA computation.

A.3.a Denaturation

Denaturation causes DNA double strands to unwind and separate into its two single strands. This occurs by breaking the H-bonds between the bases, which are weak compared to covalent bonds (e.g., those in the backbone). Thermal denaturing (heating to about 95°C) is the most common procedure, but there are also chemical denaturing agents.

A.3.b Hybridization

Hybridization is the process by which two DNA single strands combine in a sequence specific way through the formation of non-covalent H-bonds between the bases. *Annealing* is used to achieve an energetically favorable result (a minimal number mismatches).

[3]Church, *op. cit.*, Supplementary Material

A.3.c Polymerase extension

Enzymes called *polymerases* (polymerizing enzymes) add nucleotides one at a time to the 3′ end of a DNA molecule, an operation called *polymerase extension*. It can be used to fill in a sticky end, or a short sequence bound to a complementary sequence can operate as a *primer* causing the rest of the sequence to be filled in.

A.3.d Ligation

Hybridization links to strands together by H-bonds, but there is a *nick* in the DNA backbone resulting from a missing phosphodiester bond. Certain enzymes will *ligate* the strands by filling in the missing covalent bonds, making a more stable strand.

A.3.e Nuclease degradation

Nucleases are enzymes that remove nucleotides from a DNA strand. *Exonucleases* remove nucleotides from the ends (either 3′ or 5′ depending on the exonuclease), whereas *endonucleases* cut DNA strands (single or double) at specific places. *Restriction enzymes* are the most common endonucleases. They operate at specific *restriction sites*, typically 4 to 8 bases long. The cut can be *blunt* or *staggered* leaving sticky ends.

A.3.f Synthesis

Synthesis is the process of assembling single-stranded *oligonucleotides* ("oligos"), which are short strands with a desired sequence. Assembly occurs in the 3′ to 5′ direction.The process is completely automated nowadays, but due to the accumulation of errors, the current limit is about 200 nucleotides.

A.3.g Polymerase chain reaction

The *Polymerase Chain Reaction* (PCR) is a method for exponentially amplifying the concentration of a selected DNA sequence in a sample. It was a breakthrough and remains one of the most important processes in practical DNA technology. Here is the basic procedure.

The sequence to be replicated is identified by a known short sequence (about 20 bases) at the 3′ ends of both its sense and antisense strands. Short

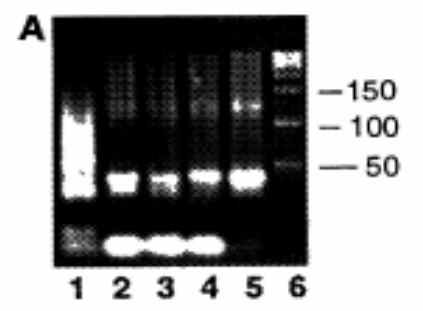

Figure IV.6: Example of result of gel electrophoresis.

DNA strands called *primers*, which are complementary to the identifying sequences, are synthesized. That is, two primers are used to bound or delimit the sequence of interest (at the $3'$ end of the sense strand and the $3'$ end of the antisense strand). (You can also have primers that bind to the $5'$ ends of the target sequence, but $3'$ is typical.)

Step 1 (add primers): The primers and DNA polymerase are added to the DNA sample (possibly containing the desired sequence and many others).

Step 2 (denaturation): The mixture is heated to about 95°C for about 30 seconds, which causes all the double strands to separate.

Step 3 (annealing): The mixture is cooled to about 68°C for a minute, which allows the primers to hybridize with the other strands.

Step 4 (polymerase extension): The mixture is warmed to about 72°C for 30 seconds or more (depending on the length of the sequence to be replicated, which goes at about about 1000 bp/min.). This allows the polymerase to extend the primer on each single strand so that it becomes a double strand. Pairs of long strands can rehybridize, but the lighter oligonucleotides tend to get there faster, so instead we get two double strands. As a result, the concentration of the desired DNA sequence has doubled.

Step 5 (repeat): Repeat Steps 2 to 4 until the desired concentration is reached. The concentration of the desired sequence increases exponentially, 2^n, in the number of cycles. There are "PCR machines" that automate the entire PCR process.

A.3.h Gel electrophoresis

Gel electrophoresis is a method of determining the size of DNA fragments in a

sample. The sample is put in a well at one end of a layer of gel, and a positive electric field is applied at the other end of the gel. Since DNA is negatively charged, it migrates through the gel, but smaller fragments migrate faster. Later the DNA can be stained and the gel imaged to determine the relative concentration of fragments of different sizes in the sample. The resulting pattern is called a *DNA ladder*; see Fig. IV.6 for several "lanes" of DNA ladders.

A.3.i Filtering

There are several ways to filter DNA samples for specific sequences. One way to select for an oligonucleotide o in a sample S is to attach the complementary sequence $\overline{o}$ to a filter. Then when S goes through the filter, the o in it will hybridize with the $\overline{o}$ in the filter, and stay there; what passes through is $S - o$. If desired, the $o\overline{o}$ on the filter can be denatured to separate the strands containing o from the $\overline{o}$ on the filter.

A.3.j Modification

Certain enzymes can be used to change subsequences of a DNA sequence.

A.3.k Sequencing

Sequencing is the process of reading out the sequence of a DNA molecule. This is traditionally accomplished by extension of a primed template sequence, but a variety of other techniques have been developed since the 1970s. The latest techniques do massively parallel sequencing of many fragments.