Chapter IV

Molecular Computation

A Basic concepts

A.1 Introduction to molecular computation

¶1. Molecular computation uses molecules to represent information and molecular processes to implement information processing.

¶2. DNA computation is the most important (and most explored) variety of molecular computation.

¶3. The principal motivation is that molecular computation is massively parallel.
Molar or *Avagadro-scale* parallelism.

¶4. Data representations are also very dense (on the order of nanometers).

¶5. Molecular computation can be very efficient in terms of energy.

A.2 DNA basics

¶1. The DNA molecule is a polymer (*poly* = many, *mer* = part), a sequence of *nucleotides*.

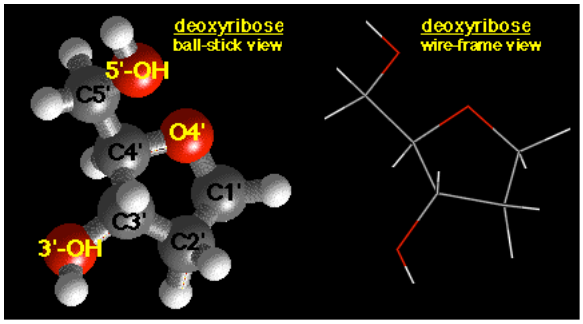

Figure IV.1: Deoxyribose. [source: wikipedia]

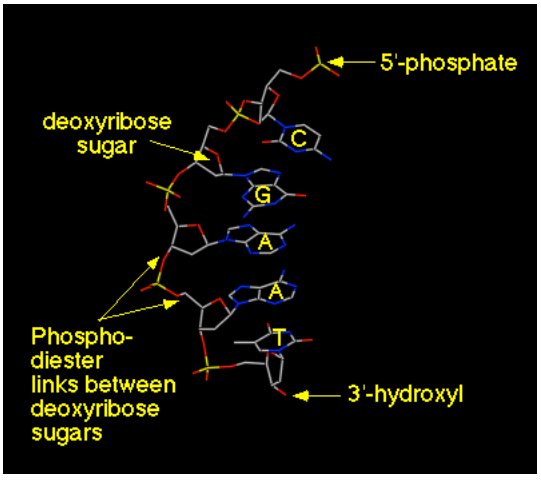

Figure IV.2: DNA backbone. [source: wikipedia]

¶2. **Nucleotide:** 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.

¶3. **DNA backbone:** The *DNA backbone* is a sequence of deoxyribose sugars alternating with phosphate groups connected by covalent *phosphodiester* bonds.
See Fig. IV.2.

¶4. It connects the hydroxyl group on the 3′ carbon (the "3′-hydroxyl group") of one sugar to the 5′-hydroxyl of the next.

¶5. 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.

¶6. **Watson-Crick complementarity:** A and T each have two H-bonding sites and can bind together.
G and C each have three H-bonds and can bond together.
H-bonds are weak compared to covalent bonds.

¶7. As a consequence, two complementary polynucleotides can bond together.
This can occur only if the two strands are *antiparallel*, i.e., run in opposite directions (3′ to 5′ versus 5′ to 3′).
See Fig. IV.3.

¶8. Since we write the *sense-strand* in the 5′ to 3′ order, we write the *antisense-strand* in the 3′ to 5′ order.

¶9. **Sticky ends:** Complementary single stranded DNA can hybridize, leaving *sticky ends* at either or both ends.
This is because they are available for bonding by complementary sequences.

¶10. 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.

[1] In RNA, uracil replaces thymine.

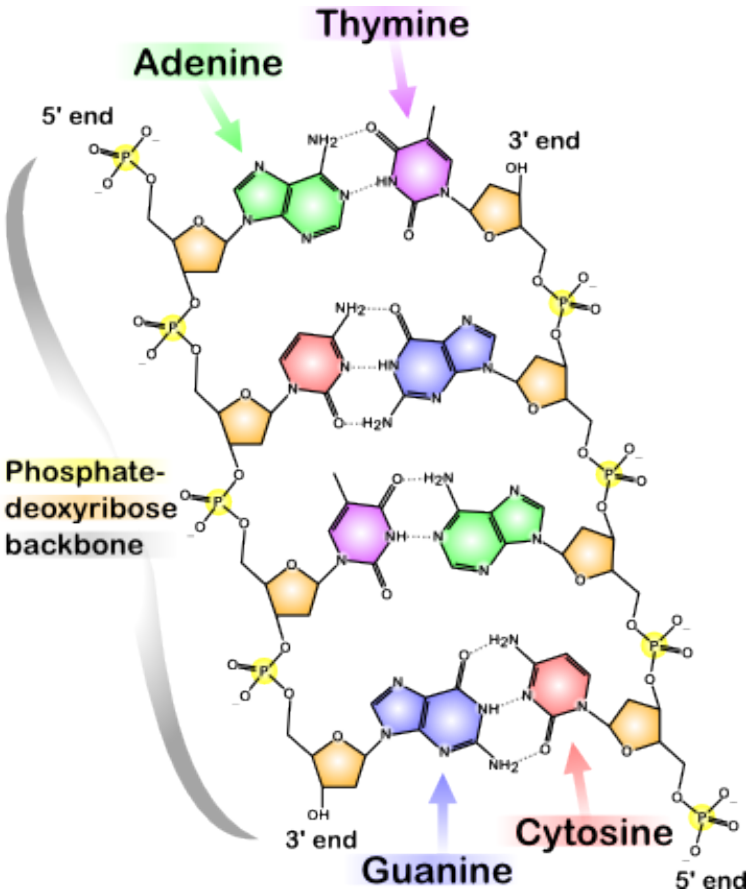

Figure IV.3: [source: wikimedia commons]

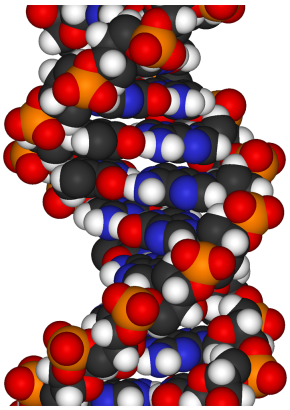

Figure IV.4: [source: wikimedia commons]

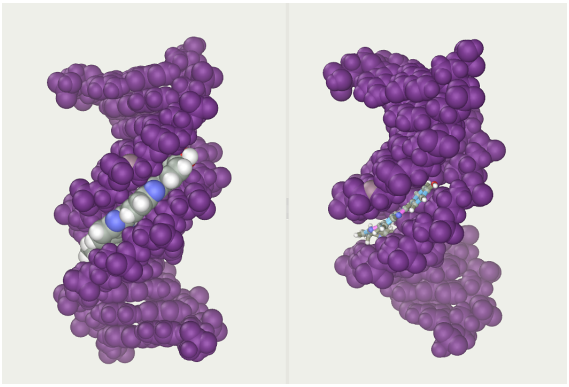

Figure IV.5: Major and minor grooves. [source: wikimedia commons]

¶11. **Major and minor grooves:** 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.

¶12. **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 $\sim 220\times10^6$bp (base pairs) long.

¶13. **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 experimental media, such as quantum holography).
The book included 11 jpegs and a javascript program, for a total of 5.27 Mbits.
The decoded version had only 10 incorrect bits.
5.5 petabits/mm^3 = 5.5×10^{15}bits/mm^3. The book could be reproduced by DNA replication.
(The cost of DNA synthesis has been decreasing about 5× per year, and the cost of sequencing by about 12× per year.)

¶14. **Potential:** DNA can encode 455 exabytes (4.55×10^{20} bytes) per gram of single-stranded DNA.
(This is for 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])

---

[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).

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

A.3 DNA manipulation

A.3.a Denaturation

¶1. **Denaturation:** causes DNA to unwind and separate into its two strands.
This occurs by breaking the H-bonds between the bases.

¶2. Thermal denaturing is most common (about 95°C), but there are also chemical denaturing agents.

A.3.b Hybridization

¶1. **Hybridization:** the process by which two DNA strands combine in a sequence specific way.

¶2. Formation of non-covalent H-bonds.

¶3. **Annealing:** used to achieve an energetically favorable result (minimal mismatches).

A.3.c Polymerase extension

¶1. Enzymes called *polymerases* (polymerizing enzymes) add nucleotides to the 3′ end of a DNA molecule.

¶2. It can be used to fill in a sticky end.

¶3. 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

¶1. Hybridization links to strands together by H-bonds, but there is a *nick* in the DNA backbone resulting from a missing phosphodiester bond.

¶2. Certain enzymes will *ligate* the strands by filling in the missing covalent bonds.

A.3.e Nuclease degradation

¶1. **Nucleases:** enzymes that remove nucleotides from a DNA strand.

¶2. **Exonucleases:** remove nucleotides from the ends (either 3′ or 5′ depending on the exonuclease).

¶3. **Endonucleases:** cut DNA strands (single or double) at specific places.

¶4. **Restriction enzymes:** are the most common.

¶5. They operate at specific *restriction sites*, typically 4 to 8 bases long.

¶6. The cut can be *blunt* or *staggered* leaving sticky ends.

A.3.f Synthesis

¶1. **Synthesis:** the process of assembling single-stranded *oligonucleotides* ("oligos") with a desired sequence.

¶2. Assembly occurs in the 3′ to 5′ direction.

¶3. The process is completely automated.

¶4. Due to the accumulation of errors, the current limit is about 200 nucleotides.

A.3.g Polymerase chain reaction

¶1. The *Polymerase Chain Reaction* (PCR) is a method for exponentially amplifying the concentration of a selected DNA sequence in a sample.

¶2. **Primers:** The sequence to be replicated is identified by a known short sequence (about 20 bases) at its beginning (5′ end).
Short DNA strands called *primers*, which are complementary to the identifying sequence, are synthesized.
Two primers are used to bound the sequence of interest (at the 3′ end of the sense strand and the 3′ end of the antisense strand).

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

¶4. **Step 2 (denaturation):** The mixture is heated to about 90°C, which causes all the double strands to separate.
E.g., for about 30s.

¶5. **Step 3 (annealing):** The mixture is cooled to about 55°C for a minute.
This allows the primers to hybridize with the other strands.

¶6. **Step 4 (polymerase extension):** The mixture is warmed to about 72°C for 30s or more (depending on the length of the sequence to be replicated). 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 desired DNA has doubled.

¶7. **Step 5 (repeat):** Repeat Steps 2 to 4 until the desired concentration is reached.

¶8. **Exponential amplification:** The concentration of the desired sequence increases exponentially, 2^n, in the number of cycles.

¶9. **PCR machines:** There are "PCR machines" that automate the entire PCR process.

A.3.h Gel electrophoresis

¶1. *Gel electrophoresis* is a method of determining the size of DNA fragments in a sample.

¶2. The sample is put in a well at one end of a layer of gel.

¶3. A positive electric field is applied at the other end of the gel.

¶4. Since DNA is negatively charged, it migrates through the gel, but smaller fragments migrate faster.

¶5. Later the DNA can be stained and the gel imaged to determine the relative concentration of fragments of different sizes in the sample.

¶6. **DNA ladder:** the resulting pattern.

A.3.i Filtering

¶1. There are several ways to filter DNA samples for specific sequences.

¶2. To filter for an oligonucleotide o in a sample S, attach the complementary sequence $\overline{o}$ to a filter.

¶3. Then when S goes through the filter, the o in it will hybridize with the $\overline{o}$ in the filter, and stay there.

¶4. What passes through is $S - o$.

¶5. If desired, the $o\overline{o}$ on the filter can be denatured into o and $\overline{o}$.

A.3.j Modification

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

A.3.k Sequencing

¶1. **Sequencing:** The process of reading out the sequence of a DNA molecule.
This is usually accomplished by extension of a primed template sequence.

¶2. A variety of techniques have been developed since the 1970s.
The latest techniques do massively parallel sequencing of many fragments.