Introduction to DNA Computing

The lecture notes were prepared according to Leonard Adleman's seminal paper "Molecular Computation of Solutions to Combinatorial Problems" and Keith Devlin's explanatory article "Test Tube Computing with DNA".

The Hamiltonian path problem (HPP)

- Given a directed graph with two node specified as the source and the destination. The Hamiltonian path is one that starts at the source node and ends at the destination node such that each node in the graph appears once and only once on the path.
- The Hamiltonian path problem is to determine whether there is a Hamiltonian path for a directed graph and maybe furthermore, find a Hamiltonian path if one exists.
- An example of a graph of seven nodes with node 0 as the source and node 6 as the destination, which was used in Adleman's DNA experiment.

- HPP is NP-complete. Using exhaustive search, one has to generate all $n!$ possible paths (for a graph with n nodes) and check for each path if it is a Hamiltonian path. Even for a graph of only 10 nodes, the number of possible paths is $10! = 3,628,800$.
- An idea for fast algorithms: Use parallel computation to test all possible paths at the same time. The power of a parallel computer (a few thousands simultaneous computations) is not enough.

A five-step algorithm used in Adleman's experiment

- Step 1. Generate a large number of paths through the graph.
- Step 2. Keep only those that start with the source and end with the destination.
- Step 3. Keep only those of length n.
- Step 4. Keep only those that pass through each node once.
- Step 5. If there are any paths left, any of them will be a Hamiltonian path.

The algorithm is not deterministic since in Step 1 not all paths of length 7 (or n in general) are to be generated. Adleman estimated that his DNA computation generated about 10^{14} paths, so the overwhelming likelihood was that any given path of length 7 would be produced many times over.

DNA 101

- DNA is the storage medium for genetic information.
- Four kinds of *bases* that form our alphabet: adenine (A), thymine (T), guanine (G) and cytosine (C).
- A single DNA strand with b bases is a string of length b over the alphabet $\{$A,T,G,C$\}$. The two ends of a strand are marked with 3' and 5', respectively. So a DNA strand is considered oriented: 5' to 3' or 3' to 5'.
- Two complimentary pairs: A-T and G-C, caused by mutual attraction (hydrogen bonding) between A and T and between G and C.
- The Watson-Crick complementary (dual or mirror image): 5'GCTATT3' and 3'CGATAA5'.

- A DNA molecule consists of two intertwined complementary strands made up with four bases, A, T, G, C. This is called double helix structure, discovered by Watson and Crick.
- Heat (90° C) separates the double strands and cooling bonds them back.
- A typical human DNA molecule is about three billion bases long. But synthetic DNA (called oligonucleotide or oligo for short) may be ten to a hundred bases long.

Encoding the graph into DNA strands

- Nodes: For each node $i = 0, 1, \ldots, 6$, Adleman chose a random 20-base strand of DNA, O_i, to represent the node. For example,

$$O_2 = 5'\text{TATCGGATCGGTATATCCGA}3'$$

$$O_3 = 5'\text{GCTATTCGAGCTTAAAGCTA}3'$$

$$O_4 = 5'\text{GGCTAGGTACCAGCATGCTT}3'$$

- Edges: For each edge $i \to j$, where $i \neq 0$ (i.e., i is not the source) and $j \neq 6$ (i.e., j is not the destination), Adleman created a 20-base strand $O_{i \to j}$ that consisted of the last ten bases of O_i and the first ten bases of O_j. For example,

$$O_{2 \to 3} = 5'\text{GTATATCCGAGCTATTCGAG}3'$$

$$O_{3 \to 4} = 5'\text{CTTAAAGCTAGGCTAGGTAC}3'$$

In the case of $i = 0$ (the source), $O_{i \to j}$ was all of O_i followed by the first ten bases of O_j. And in the case of $j = 6$ (the destination), $O_{i \to j}$ was the last ten bases of O_i followed by all of O_j. (Why?)

- Paths: To join edges to form a path, DNA strands need to be bonded together.
  Ligation reaction: First the strands to be joined were held together temporarily by a “molecule splint”. Then they were bonded together permanently by the action of an enzyme that occurs in living cells called ligase. The second step is really a case of letting nature take its course: provided the two strands are held together for a sufficient length of time, the permanent bond will form.
  The molecule splint to ligate $O_{i\to j}$ and $O_{j\to k}$ is the Watson-Crick complementary of O_j, denoted by $\overline{O}_j$.
  What would the Hamiltonian path $0 \to 1 \to 2 \to 3 \to 4 \to 5 \to 6$ look like in the form of a DNA molecules?

The experiment step by step

Step 1. Generate paths

For each node i, get 50 pmol of $\overline{O}_i$. For each edge $i \to j$ in the graph, get 50 pmol of $O_{i \to j}$. Mix all together ($\frac{1}{50}$ tsp.) in a single ligation reaction. DNA molecules were formed that encoded a large number of random paths in the graph.

Step 2. From node 0 to node 6
This step is achieved by making many copies of the DNA molecules that encode the paths from node 0 to node 6 (instead of throwing away those paths that do not satisfy the requirement). The process is called amplification by *polymerase chain reaction* (PCR).
In general, to amplify a DNA with a strand xyz and another strand $\overline{xyz}$, heat is used to separate strands. Then primers x and $\overline{z}$ are added to the solution to start PCR. First x is bonded with $\overline{xyz}$ and $\overline{z}$ is bonded with xyz. Then the DNA polymerase, an enzyme in living cells, facilitates the growth of missing parts. At the end, two copies of the DNA molecules are formed.
To amplify (multiply) DNA molecules that encode paths from node 0 to node 6, two primers O_0 and $\overline{O}_6$ are used in the PCR process.

Step 3. Length of 7

A standard technique to separate DNA of different lengths known as *gel electrophoresis* is used. First the DNA mixture is put on one end of a sheet of sugar gel. A uniform electric charge is then applied to the two ends of the gel, negative at the end containing the DNA. The electric charge causes the DNA molecules to migrate toward the positive end. The shorter the DNA molecule, the faster it moves. After charging for a while, the DNA is separated into a spectrum ranging from the longest at the negative end and the shortest at the positive end. How can the DNA with 140 bases be identified? Do a separate, yet simultaneous gel run using a DNA mixture of only 140-base molecules, and then use the new position of this 140-base only mixture as a measuring stick to identify the 140-base molecules in the original mixture.

Finally, the part of the gel containing the DNA with 140 bases (corresponding to paths of length 7) is cut off and the DNA is extracted from the sugar gel and purified.

Step 4. Each node at least once

Since each path already contains node 0 and node 6, only five nodes, $1, 2, 3, 4, 5$, need to be checked. The following method, known as *affinity purification*, identifies those paths containing node i and should be applied five times for $i = 1, 2, 3, 4, 5$.

A $\overline{O}_i$ probe is a system that attaches the oligo $\overline{O}_i$ to a magnetic bead. First place a magnet alongside the test tube to draw and hold all the beads. Then pour in separated DNA strands (heated after Step 3). Only those strands with O_i in the sequences are drawn to the probes thus held on the magnet while the remaining contents of the tube are poured away. Next use heat to separated the DNA strands (with O_i) from the beads. Finally pour out the solution. It contains all strands with O_i in the sequences.

Step 5. Read the result
Amplify the product of Step 4 by PCR and run on a gel. The visible presence of molecules (in the form of a band) on the gel indicates that the graph does have a Hamiltonian path. Had Adleman started with a graph having no Hamiltonian path, the result of the final gel run would not have produced any band in the gel.

The dawn of a new era?

Limitations of Adleman's method

- Problem-specific: Tailored only to solve HPP
- Seven days of experiments to solve an instance that can be solved by any human in seconds
- Not suitable for numerical computation
- Results not as certain as with an electronic computer.

Advantages of Adleman's method

- Massive parallelism to run a large number of trials simultaneously
- Tremendous storage capacity (1 bit: one cubic nanometer of DNA versus one trillion cubic nanometer on a magnetic medium)
- HPP is an NP-complete problem. The computational equivalence among NP-complete problems indicates that similar DNA computation may be used to solve other NP-complete problem (Lipton's work)

A striking comparison

- 1948, Tom Kilburn, Manchester Mark I (the first programmable computer), to find the largest factor of an integer, 52 minutes, manual labor
- 1994, Len Adleman, DNA computing, to solve the HPP, 7 days, manual labor

What will DNA computing be 50 years from now?