COS 341, November 27, 2000
Handout Number 8

DNA Sequencing

The material is from a paper by P. Pevzner, "1-tuple DNA sequencing: computer analysis," *Journal of Biomoleculare Structure* **7** (1989), pp. 63-73.

The following is well known. Let G be a general digraph such that there exist vertices x, y satisfying $\text{outdegree}(x) - \text{indegree}(x) = 1$, $\text{outdegree}(y) - \text{indegree}(y) = -1$, and $\text{outdegree}(v) - \text{indegree}(v) = 0$ for all other vertices v. Then there exists at least one open Eulerian trail in G from x to y.

1 Hybridization Method for DNA Sequencing

Let $\Delta = \{A, C, G, T\}$. Define $\overline{A} = T$, $\overline{T} = A$, $\overline{C} = G$, and $\overline{G} = C$. A (single-stranded) DNA *fragment* is a string $\sigma \in \Delta^n$; n is the *length* of the fragment. Define $\overline{\sigma} = \overline{a}_1\overline{a}_2 \cdots \overline{a}_n$ if $\sigma = a_1 a_2 \cdots a_n$.

Given a DNA fragment σ of length n, the hybridization method to determine σ works as follows. Let $2 \leq \ell \leq n$ be a parameter. Construct a *chip* with 4^ℓ cells, each containing copies of one distinct string $\rho \in \Delta^\ell$. If we wash a bottle of solution containing many copies of σ over the chip, then those cells containing ρ as substrings of $\overline{\sigma}$ get some copies of σ attached to the cells. The *spectrum* of σ is the set of those activated ρ's. In other words, the spectrum is the set of all possible length-ℓ substrings of $\overline{\sigma}$.

The algorithmic question is: Given the spectrum, can we reconstruct $\overline{\sigma}$ and hence σ? That is, how to find all σ that have exactly this spectrum.

For simplicity we shall be only concerned with the special case when all the length-ℓ substrings of σ (hence also $\overline{\sigma}$) are distinct. Under this assumption, $|S| = n - \ell + 1$.

2 Hybridization and Eulerian Path

Let S be the spectrum of σ. Construct a general digraph $G_S = (V, E)$ as follows. For any string $\rho = a_1 a_2 \cdots a_\ell \in \Delta^\ell$, call $a_1 a_2 \cdots a_{\ell-1}$ and $a_2 a_3 \cdots a_\ell$ the *prefix* and *suffix* of σ. Let V be the set of all length-$(\ell - 1)$ strings that are either a prefix or a suffix of some element in the spectrum. For each element ρ in the spectrum, create an edge from its prefix to its suffix.

For any path in G_S, there is a natural associated string. We illustrate it with the following example. Let $n = 10, \ell = 3$, and S consists of $ATG, TGT, TGC, GTG, GCA, GCC$,

CGC, CCG. Then G_S has an open Eulerian trail $AT - TG - GT - TG - GC - CC - CG - GC - CA$. The string associated with this open Eulerian trail is $ATGTGCCGCA$. It happens that this is also the only open Eulerian trail.

It is clear that two different trails give two different associated strings. Also, the string associated with any open Eulerian trail is a string $\overline{\sigma}$ such that S is the spectrum of σ. Thus, the problem of determining σ is equivalent to finding all the open Eulerian trails of G_S.