COS 341, November 10, 1997
Handout Number 14

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.

1 A Sufficient Condition for Unique Eulerian Path

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.

Construct a graph G' as follows. Add an edge (y, x) into the edge set of G. (Now the indegree of every vertex is the same as its outdegree.) Partition the edges of G into simple cycles $C_1, C_2, \cdots, C_m$ (such that all cycles are disjoint in their edges). Define $G' = (V', E')$, where $V' = \{C_1, C_2, \cdots, C_m\}$ and between two vertices C_i, C_j, there are k edges in E' if the two cycles have exactly k points in common in G. Note that depending on the choice of simple cycles, there may be several different G'.

For example, the following general digraph G has G' as shown.

The following fact was stated in Pevzner's paper.

Lemma 1 Let G be connected, and there exist vertices x, y such that $\text{outdegree}(x) - \text{indegree}(x) = 1$, $\text{outdegree}(y) - \text{indegree}(y) = -1$, and $\text{outdegree}(v) - \text{indegree}(v) = 0$ for all other vertices v. If $\text{outdegree}(v), \text{indegree}(v) \leq 2$ for all v, and if G' is a tree, then G has a unique Eulerian path.

2 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? By this, we mean firstly, how to find a $\overline{\sigma}$ that can generate exactly this spectrum, and secondly, is this a unique solution?

We shall be only concerned with the case when all the length-ℓ substrings of σ (hence also $\overline{\sigma}$) are distinct. This is a reasonable assumption when ℓ is quite a bit larger than $\log_2 n$ (such as $n = 200, \ell = 8$). Under this assumption, $|S| = n - \ell + 1$.

3 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 a Eulerian path $AT{-}TG{-}GT{-}TG{-}GC{-}CC{-}CG{-}GC{-}CA$. The string associated with this Eulerian path is $ATGTGCCGCA$.

It is clear that two different paths give two different associated strings. Also, the string associated with any Eulerian path is a string $\overline{\sigma}$ such that S is the spectrum of σ. If G_S has a unique Eulerian path, then we have found a σ and at the same time know that this is the unique solution.

Pevzner reported for the case $n = 12, \ell = 8$, a statistical experiment shows that for 94% of the strings σ, the general digraph G_S obtained from the spectrum S satisfies the conditions in Lemma 1, and hence σ can be reconstructed from the spectrum by this method.