CS 466 – Introduction to Bioinformatics – Lecture 13

Mohammed El-Kebir

October 9, 2019

Document history:

- 9/19/2018: Initial version.
- 10/9/2019: Minor changes.

Contents

1 RNA Secondary Structure Prediction

As described in Lecture 1, the central dogma of molecular biology states that DNA is transcribed into RNA, which in turn is translated into protein. This seems to imply that RNA is simply an intermediary molecule, whose sole role is to facilitate translation of protein by migrating from the cell nucleus (where DNA resides) to the ribosome (protein factory). Indeed, RNA molecules that are translated at a ribosome are called messenger RNAs (mRNA), reflecting their passive roles. However, other types of RNA molecules also exist that are non-coding, i.e. not translated into proteins. Such RNA molecules have complicated 3D spatial structures to achieve their biological function. In this lecture, we will introduce a problem statement for inferring RNA 2-D structure (also called secondary structure) from a 1-D RNA sequence, by means of an optimization problem. We will present the Nussinov algorithm for solving this problem.

1.1 Background

There are four RNA nucleotides, i.e. $\Sigma = \{A, U, G, C\}$. When two RNA nucleotides are close to each other in space they can form a chemical bond. We consider only two types of bonds: $A - U$ and $G - C$. The latter $(G - C)$ is slightly stronger than the former $(A - U)$ with

three hydrogen bonds instead of two. The key concept is that nucleotide pairings lead to RNA spatial structure.

1.2 Problem Statement

Let $\mathbf{v} \in \Sigma^n$ be an RNA sequence. Each nucleotide can participate in at most one pairing. Per the previous section, there are four *complementary nucleotide pairings* (A, U), (U, A), (G, C) and (C, G). The *secondary structure* P of an RNA molecule is defined by a set of non-overlapping complementary nucleotide pairings. That is, $P \subset \{(i, j) \mid 1 \leq i < j \leq n\}$ such that for all distinct pairs $(i, j), (i', j') \in P$ it holds that $i \neq i'$, $i \neq j'$, $j \neq i'$ and $j \neq j'$. Observe that $|P|$ is the number of complementary nucleotide pairings in secondary structure P.

Consider two nucleotide pairs (i, j) and (i', j') such that $i < j$ and $i' < j'$. We say that (i, j) and (i', j') are *nested* if and only if either $i < i' < j' < j$ or $i' < i < j < j'$. Conversely, (i, j) and (i', j') form a *pseudoknot* if and only if either $i < i' < j < j'$ or $i' < i < j' < j$. A secondary structure P is *pseudoknot free* provided that no two pairs in P form a pseudoknot.

We have the following problem.

Problem 1. *Given RNA sequence* $\mathbf{v} \in \Sigma^n$*, find a pseudoknot-free secondary structure* P *such that* $|P|$ *is maximum.*

1.3 Nussinov Algorithm

Let $s[i, j]$ denote the maximum number of pseudoknot-free complementary nucleotide pairings in subsequence $v_i, \ldots, v_j$. We wish to compute $s[1, n]$. Clearly, there is optimal substructure. To see this, think about an optimal secondary structure P^* with length $|P^*| = s[1, n]$. We distinguish the following cases.

1. $(1, n) \in P^*$, meaning that v_1 and v_n are complementary nucleotides and part of the optimal pseudoknot-free secondary structure P^*. In this case $s[1, n] = s[2, n-1] + 1$.

2. $(1, n) \notin P^*$, meaning that either v_1 and v_n are not complementary nucleotides or that they are complementary but not part of the optimal pseudoknot-free secondary structure P^*. In this case, we can distinguish four additional cases.

   (a) Neither v_1 nor v_n are part of complementary base pairing in P^*. In this case, $s[1, n] = s[2, n-1]$.

   (b) v_1 is not part of a complementary base pairing in P^*. In this case, $s[1, n] = s[2, n]$

   (c) v_n is not part of a complementary base pairing in P^*. In this case, $s[1, n] = s[1, n-1]$.

   (d) v_1 and v_n are both part of a complementary base pairing in P^* (but not with each other). In this case, there exists a $1 < k < n$ such $s[1, n] = s[1, k] + s[k+1, n]$. In other words, we split the sequence into two parts (i.e., a bifurcation).

We can apply the same case analysis on each of the resulting subcases. Observe, that we are "peeling off" subproblems, in line with optimal substructure.

Let $\Gamma = \{(A,U),(U,A),(G,C),(C,G)\}$. We can write down the following recurrence

$$s[i,j] = \max \begin{cases} 0, & \text{if } i \geq j, \\ s[i+1,j-1]+1, & \text{if } i < j \text{ and } (v_i, v_j) \in \Gamma, \\ s[i+1,j-1], & \text{if } i < j \text{ and } (v_i, v_j) \notin \Gamma, \\ s[i+1,j], & \text{if } i < j, \\ s[i,j-1], & \text{if } i < j, \\ \max_{i<k<j}\{s[i,k]+s[k+1,j]\}, & \text{if } i < j, \end{cases} \tag{1}$$

where $1 \leq i, j \leq n$.

Thus, we can view s as an $n \times n$ table, whose cells in the lower triangle are initialized with 0 (i.e., $s[i,j] = 0$ for all $1 \leq j \leq i \leq n$). From the recurrence, we see that each entry $s[i,j]$ where $1 \leq i \leq j \leq n$ depends on entries on the diagonals below it. Thus, we can fill out the table diagonally starting from $j = i+1$. Cell $s[1,n]$ will contain the maximum number of complementary nucleotide pairings of any pseudoknot-free secondary structure. Writing down the pseudocode for filling out the table is left as an exercise to the reader.

1.4 Traceback

How do we reconstruct the optimal pseudoknot-free secondary structure P^* from a filled out table s? We can do this recursively starting from $s[1,n]$. See slides for alternative pseudocode using a stack.

1.5 Running Time Analysis

The space complexity is dominated by storing the complete table s and is thus $O(n^2)$. To see what the running time of filling out the table is, observe that each entry $s[i,j]$ must be computed. The time required for computing $s[i,j]$ is dominated by the last case, where we scan $k \in \{i+1,\ldots,j-1\}$. Observe that each triple (i,j,k) where $1 \leq i < k < j \leq n$ will be considered exactly once when filling out s. Thus, the running time is bound by the number of such triples, which is $\binom{n}{3} = O(n^3)$.