Lecture 16

RNA Secondary Structure Prediction

February 29, 2000
Notes: Matthew Cary

16.1. RNA Secondary Structure

Recall from Section 1.3 that RNA is usually single-stranded in its "normal" state, and this strand folds into a functional shape by forming intramolecular base pairs among some of its bases. (See Figure 16.1 for an illustration.) The geometry of this base-pairing is known as the "secondary structure" of the RNA.

When RNA is folded, some bases are paired with other while others remain free, forming "loops" in the molecule. Speaking qualitatively, bases that are bonded tend to stabilize the RNA (i.e., have negative free energy), whereas unpaired bases form destabilizing loops (positive free energy). Through thermodynamics experiments, it has been possible to estimate the free energy of some of the common types of loops that arise.

Because the secondary structure is related to the function of the RNA, we would like to be able to predict the secondary structure. Given an RNA sequence, the *RNA Folding Problem* is to predict the secondary structure that minimizes the total free energy of the folded RNA molecule.

The prediction algorithm that will be described is by Lyngsø *et al.* [2].

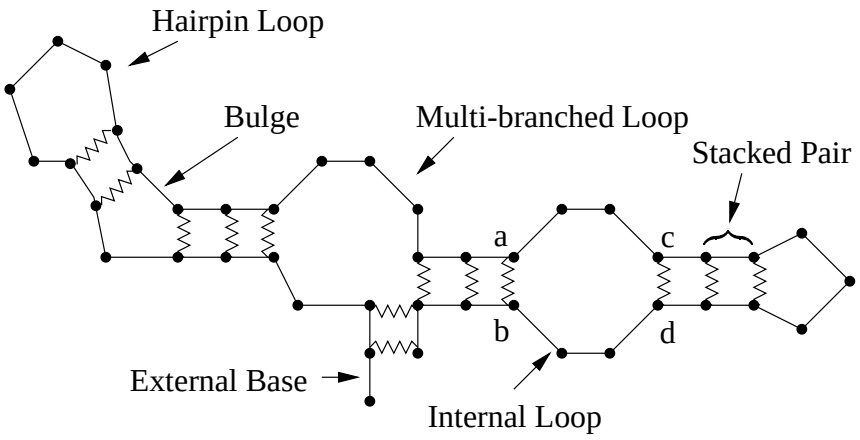

Figure 16.1: RNA Secondary Structure. The solid line indicates the backbone, and the jagged lines indicate paired bases.

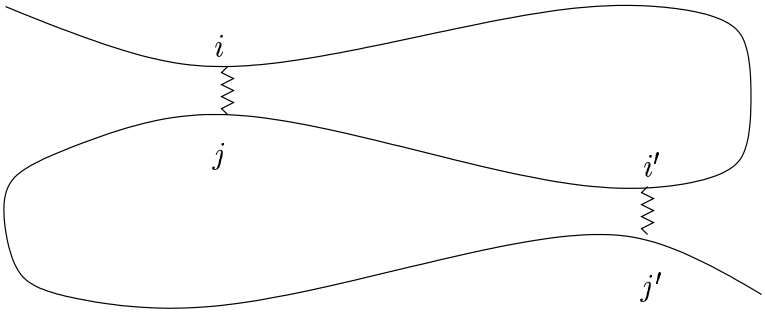

Figure 16.2: A Pseudoknot

16.2. Notation and Definitions

If $s = s_1 s_2 \ldots s_n$ is an RNA sequence and $1 \leq i < j \leq n$, then $i \cdot j$ denotes the base-pairing of s_i with s_j.

Definition 16.1: A *secondary structure* of $s = s_1 s_2 \ldots s_n$ is a set S of base pairs such that each base is paired at most once. More precisely, for all $i \cdot j \in S$ and $i' \cdot j' \in S$, $i = i'$ if and only if $j = j'$.

The configuration shown in Figure 16.2 is known as a *pseudoknot*. For the prediction algorithm that follows, we will assume that the secondary structure does not contain any pseudoknots. The ostensible justifications for this are that pseudoknots do not occur as often as the more common types of loops, and secondary structure prediction is moderately successful even if pseudoknots are prohibited. However, the real justification for this assumption is that it greatly simplifies the model and algorithm. (Certain types of pseudoknots are handled by the algorithm of Rivas and Eddy [3], but the general problem was shown NP-complete by Lyngsø and Pedersen [1].)

Definition 16.2: A *pseudoknot* in a secondary structure S is a pair of base pairs $i \cdot j \in S$ and $i' \cdot j' \in S$ with $i < i' < j < j'$.

16.3. Anatomy of Secondary Structure

Given the assumption of no pseudoknots, the secondary structure can be decomposed into a few types of simple loops, described as follows and illustrated in Figure 16.1.

Definition 16.3:

- A *hairpin loop* contains exactly one base pair.
- An *internal loop* contains exactly two base pairs.
- A *bulge* is an internal loop with one base from each of its two base pairs adjacent on the backbone.
- A *stacked pair* is a loop formed by two base pairs $i \cdot j$ and $(i+1) \cdot (j-1)$, thus having *both* ends adjacent on the backbone. (This is the only type of loop that stabilizes the secondary structure. All other loops are destabilizing, to varying degrees.)

- A *multibranched loop* is a loop that contains more than two base pairs.
- An *external base* is a base not contained in any loop.

Definition 16.4: Given a loop, one base pair in the loop is closest to the ends of the RNA strand. This is known as the *exterior* or *closing* pair. All other pairs are *interior*. More precisely, the exterior pair is the one that maximizes $j - i$ over all pairs $i \cdot j$ in the loop.

Note that one base pair may be the exterior pair of one loop and the interior pair of another.

16.4. Free Energy Functions

The assumption of no pseudoknots leads to the following related assumptions:

1. The free energy of a secondary structure is the sum of the free energies of its loops.
2. The free energy of a loop is independent of all other loops.

These assumptions imply that, to evaluate the free energy of a given secondary structure, all that is needed is a set of functions that provide the free energies of the allowable constituent loop types. These functions are the *free energy functions*, which we will assume are provided by experimentalists and are available for the algorithm's use. See `http://www.ibc.wustl.edu/~zuker/rna/energy/node2.html#SECTION20` for typical tables and formulas that can be used.

Definition 16.5: There are four free energy functions:

- $\mathrm{eS}(i, j)$. This function gives the free energy of a stacked pair that consists of $i \cdot j$ and $(i+1) \cdot (j-1)$. $\mathrm{eS}(i, j)$ depends on all the bases involved in the stack, namely s_i, s_j, s_{i+1}, and s_{j-1}. Because stacked complementary base pairs are stabilizing, eS values will be negative if both stacked base pairs are complementary. In addition to the usual complementary pairs A-U and C-G, the pair G-U forms a weak bond in RNA, and is sometimes called a "wobble pair". The eS values involving such pairs will also be negative.
- $\mathrm{eH}(i, j)$. This function gives the free energy of a hairpin loop closed by $i \cdot j$. This function depends on several factors, including the length of the loop, s_i and s_j, and the unpaired bases adjacent to s_i and s_j on the loop.
- $\mathrm{eL}(i, j, i', j')$. This function gives the free energy of an internal loop or bulge with exterior pair $i \cdot j$ and interior pair $i' \cdot j'$. Similar to eH, this function depends on $i' - i$, $j - j'$, the four paired bases, and the unpaired bases adjacent to the paired bases on the loop.
- $\mathrm{eM}(i, j, i_1, j_1, \ldots, i_k, j_k)$. This function gives the free energy of a multibranched loop closed by $i \cdot j$ with interior pairs $i_1 \cdot j_1, \ldots, i_k \cdot j_k$. This function is the least well understood at this time.

16.5. Dynamic Programming Arrays

The algorithm described by Lyngsø *et al.* [2] uses dynamic programming, the technique that was used to find optimal alignments (Section 4.1). Like the affine gap penalty algorithm of Section 5.3.3, this one fills in several tables simultaneously. The five tables used are described below.

$W(j)$**:** the free energy of the optimal structure of the first j residues, $s_1 s_2 \ldots s_j$. This is the key array: if we can compute $W(n)$ (and find its associated secondary structure), we are done.

$V(i,j)$**:** the free energy of the optimal structure for $s_i \ldots s_j$, assuming $i \cdot j$ forms a base pair in that structure.

$VBI(i,j)$**:** the free energy of the optimal structure for $s_i \ldots s_j$, assuming $i \cdot j$ closes a bulge or internal loop.

$VM(i,j)$**:** the free energy of the optimal structure for $s_i \ldots s_j$, assuming $i \cdot j$ closes a multibranched loop.

$WM(i,j)$**:** used to compute VM, in a manner to be revealed later.

Despite the similarity in their number and descriptions, it is important to understand the distinction between the free energy functions of Section 16.4 and these dynamic programming arrays. The free energy functions give the energy of a single specified loop. The arrays will generally contain free energy values for a collection of consecutive loops. For example, referring to Figure 16.1, $\mathrm{eL}(a,b,c,d)$ gives the free energy of the internal loop closed by $a \cdot b$ and $c \cdot d$, whereas $V(a,b)$ gives the total free energy of all the loops to the right of $a \cdot b$, including the stacked pairs and the hairpin.

References

[1] R. B. Lyngsø and C. N. S. Pedersen. Pseudoknots in RNA secondary structures. In *RECOMB00: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan, Apr. 2000.

[2] R. B. Lyngsø, M. Zuker, and C. N. S. Pedersen. Internal loops in RNA secondary structure prediction. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 260–267, Lyon, France, Apr. 1999.

[3] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.