

MS2a, Exercises Week 7

Rune Lyngsø

November 20, 2009

A RNA Secondary Structure Prediction

- a. Use Algorithm 1 and Algorithm 2 of the lecture notes on RNA secondary structure prediction to find the maximum number of base pairs for the sequence CAGGGU, and a structure with this number of base pairs. Two bases can form a valid base pair if *i*) they are separated by at least three bases in the sequence, i.e. their indices differ by at least 4, and *ii*) they form one of the three types of base pairs shown in Figure 2 in the lecture notes. For added convenience, the table you need to fill out and backtrack (cf. Figure 5 in the lecture notes) is:

		second base # →							
		0	C ₁	A ₂	G ₃	G ₄	G ₅	U ₆	
first base # ↓	C ₁								
	A ₂								
	G ₃								
	G ₄								
	G ₅								
	U ₆								
	7								

- b. Forgetting about Algorithms 1 and 2, can you find a structure with more valid base pairs than the one you found above? If so, why does Algorithm 1 fail to find this number of base pairs?

B RNA Secondary Structure Space

- a. How many distinct RNA sequences are there of length n (yes, it is that easy)?

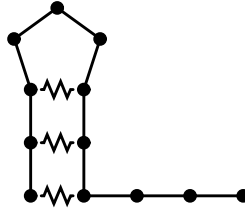


Figure 1: An RNA secondary structure with three base pairs on 12 unspecified bases.

- b. RNA secondary structures can be represented by the so called bracket, or Vienna, notation with strings over the alphabet $\{(\cdot, \cdot)\}$ (i.e. left and right parentheses and dots). A left parentheses denotes the first base in a base pair, a right parentheses the second base in a base pair, and a dot an unpaired base. For example, the structure depicted above has bracket notation $(((\cdot \cdot \cdot))) \cdot \cdot \cdot$ if the first base of the sequence is the one to the lower left. Use this notation to establish an upper bound on the number of secondary structures for sequences of length n that is less than the number of RNA sequences of length n .
- c. Find two sequences that would both have the structure depicted above as the only structure with a maximal number of canonical (i.e. C · G, A · U, and G · U) base pairs.
- d. Above we were only considering structures without pseudoknots, i.e. crossing base pairs. Assume now that pseudoknots are allowed, so any type and number of base pair crossings are allowed (a base is still restricted only to be base paired to at most one other base, though). Establish a lower bound on the number of possible pseudoknotted structures for sequences of length n that has a faster asymptotic growth than the number of sequences of length n . For convenience, you may assume that there is no minimum distance requirement between paired bases, i.e. even neighbouring bases can pair.

C Comparative Secondary Structure Prediction

- a. Algorithms 1 and 2 of the lecture notes on RNA secondary structure prediction can equally well be applied to an alignment of sequences. Instead

of finding the score of an optimal secondary structure from position i to position j for all $i \leq j$, we instead find the score of an optimal secondary structure from column i to column j in the alignment. If the score of postulating a base pair between two columns in an alignment is 1 if the two bases can form a canonical base pair for all sequences, and 0 otherwise (note that we ignore the normal requirement of bases having to be separated by three other bases to be able to form a base pair – this is purely to keep the size of this problem manageable), what is the best secondary structure you can find for the alignment

$$\begin{bmatrix} C & G & G & C & G & U & C & G \\ U & G & C & G & G & C & U & A \\ G & C & G & U & G & U & U & C \\ A & G & U & A & G & G & U & U \\ U & A & U & G & G & A & C & G \\ C & U & C & G & G & G & C & G \\ U & A & U & G & G & C & U & A \end{bmatrix}$$

You do not need to fill out the dynamic programming matrix if you can find the optimal structure in an easier way.

- b. Apart from identifying the base pairs in a resolved three dimensional structure of an RNA molecule, the other technique recognised to provide a 'gold standard' secondary structure is to identify the pairs of positions with a high *mutual information* score in a curated alignment of hundreds of homologous sequences believed to have a conserved secondary structure. The mutual information between two positions in an alignment is

$$MI_{ij} = \sum_{x_i, y_j} f_{x_i y_j} \log_2 \frac{f_{x_i y_j}}{f_{x_i} f_{y_j}}$$

where the sum is over all choices of pairs of bases (not just pairs that form canonical base pairs, but all pairs of bases – mutual information also detects non-canonical base pairs), $f_{x_i y_j}$ is the frequency with which the pair occurs in columns i and j , and f_{x_i} (f_{y_j}) is the frequency with which the first (second) base occurs in column i (j). For example, the

mutual information between the two columns in $\begin{bmatrix} A & U \\ A & U \\ U & A \\ U & A \end{bmatrix}$ and $\begin{bmatrix} A & U \\ A & A \\ U & U \\ U & A \end{bmatrix}$

is $\frac{1}{2} \log_2 \frac{1/2}{1/2 \cdot 1/2} + \frac{1}{2} \log_2 \frac{1/2}{1/2 \cdot 1/2} = 1$ and $\frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} + \frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} +$

