

MS2a, Exercises Week 7, Model Solution

Rune Lyngsø

November 25, 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: **(3 points for table and maximum number of basepairs, 2 points for one correct structure, 1 extra point for finding both optimal structures)**

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

The maximum number of valid base pairs that can be formed, 1, is the entry in the upper righthand corner of the triangular table. There are two different ways to backtrack this number, indicated by dashed and

solid arrows respectively. One corresponds to the secondary structure consisting of the base pair $C_1 \cdot G_5$, and the other corresponds to the secondary structure consisting of the base pair $A_2 \cdot U_6$.

- 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? (2 points for structure, 1 point for why Algorithm 1 fail to find it)

Both $C_1 \cdot G_5$ and $A_2 \cdot U_6$ are valid base pairs, and they do not share any bases. So we can form a structure with these 2 base pairs. The reason that Algorithms 1 and 2 don't find this structure is that the two base pairs are crossing. Algorithm 1 only considers structures without crossing base pairs, also known as pseudoknots.

B RNA Secondary Structure Space

- a. How many distinct RNA sequences are there of length n (yes, it is that easy)? (2 points)

For each position we have a choice of four bases, so clearly there are 4^n different sequences of length n .

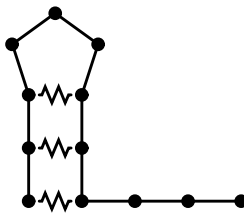


Figure 1: An RNA secondary structure with three base pairs on 12 unspecified bases.

- 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 $(((\cdot \cdot))) \cdot \cdot \cdot$ 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 . (2 points)

The bracket notation tells us that a structure on n bases can be specified by a sequence over a three letter alphabet. It follows that 3^n is an upper bound for the number of structures.

Actually, not all sequences over $\{(\cdot), \cdot\}$ corresponds to a structure as there has to be an equal number of left and right parentheses (for every first base of a base pair we need to have a second base and vice versa), and as at any point we need to have seen at least as many left parentheses as right parentheses (the first base of a base pair has to come before the second base of a base pair).

- 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. **(2 points, 1 for each correct sequence)**

The two sequences CCCAAAGGGAAA and AAACCCUUUCCC both have the structure in Fig. 1 as structure with a maximum number of base pairs. In the first sequence only the C's and G's can be base paired (A can only pair with U) and in the second sequence only the A's and U's can be base paired (C can only pair with G). In both cases the structure pairs all the bases that can form base pairs with another base. Any other structure pairing all bases that can form base pairs would have at least one pair of crossing base pairs, so this is the only maximal structure for these sequences.

- 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. **(2 points)**

As the only remaining restriction is that bases are not allowed to partake in more than one base pair, structures become equivalent to matchings. So the number of perfect matchings on n elements is a lower bound on the number of secondary structures with pseudoknots. Assuming n is even, the number of perfect matchings is $\frac{n!}{2^{n/2}(n/2)!}$ as every permutation of the n elements gives a perfect matching (pair the elements at positions $2i - 1$ and $2i$ for all $i \leq \frac{n}{2}$), but each matching

correspond to $2^{n/2}(n/2)!$ permutations (for every pair we can swap the order of the two elements in the pair, and any permutation of the pairs yield the same matching).

For $4 \uparrow n$ we get $\frac{n!}{2^{n/2}(n/2)!} = \prod_{i=1}^{n/2} (2i-1) = \prod_{i=1}^{n/4} (2i-1) \prod_{i=1}^{n/4} (\frac{n}{2} + 2i - 1) \geq 1 \cdot (\frac{n}{2})^4$ and for $n \geq 2 \cdot 4^8$ we have $(\frac{n}{2})^4 \geq (\frac{n}{2})^{\frac{n}{8}} \cdot (4^8)^{\frac{n}{8}} = (\frac{n}{2})^{\frac{n}{8}} \cdot 4^n$. Clearly $(\frac{n}{2})^{\frac{n}{8}} \rightarrow \infty$ for $n \rightarrow \infty$, so this bound grows asymptotically faster than the number of RNA sequences.

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. **(3 points)**

A quick look at the columns reveal that the only columns between which base pairs can be formed in all sequences are column 1 paired with column 8, column 3 paired with column 4, and column 5 paired with column 7. Together these three pairs form a structure with no pseudoknots and where each column is part of at most one pair. Hence, a best structure (the best structure if pseudoknots are not allowed) consists of base pairs $1 \cdot 8$, $3 \cdot 4$, and $5 \cdot 7$.

- 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 $\begin{bmatrix} A & U \\ A & A \\ U & U \\ U & A \end{bmatrix}$

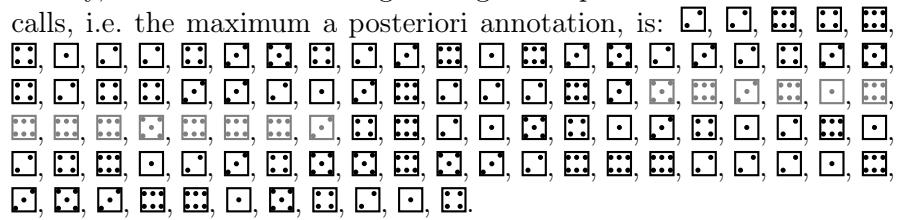
is $\frac{1}{2} \log_2 \frac{1/2}{1/2 \cdot 1/2} + \frac{1}{2} \log_2 \frac{1/2}{1/2 \cdot 1/2} = 1$ and $\frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} + \frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} + \frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} + \frac{1}{4} \log_2 \frac{1/4}{1/2 \cdot 1/2} = 0$, respectively. Compute the mutual information between the pairs of columns that you identified as base paired in Problem a and between the same number of pairs of columns not identified as base paired (your choice). **(2 points – note that only six values and not entire table are required)**

Below is the table of all pairs of mutual informations, with the entries for the three columns forming pairs that were identified above highlighted.

	1	2	3	4	5	6	7	8	
1	0	1.16	0.88	1.38	0.00	1.27	0.59	1.45	1
2		0	0.88	0.99	0.00	0.99	0.31	0.88	2
3			0	0.99	0.00	0.99	0.02	0.59	3
4				0	0.00	1.09	0.41	1.27	4
5					0	0.00	0.00	0.00	5
6						0	0.41	1.27	6
7							0	0.99	7
8								0	8

Though the mutual information for the pair 1 · 8 is the highest value in this table, clearly the data set is too small for a strong signal to emerge. One problem is that the strength of the mutual information method is also its weakness: as it can identify non-canonical base pairs, all sixteen possible types of pairs that occur more frequently than expected by random chance will boost the mutual information. If we only include contributions from canonical pairs we get the following table:

	1	2	3	4	5	6	7	8	
1	0	0.73	0.29	0.72	0.00	0.55	0.23	1.45	1
2		0	0.59	0.29	0.00	0.32	-0.02	0.35	2
3			0	0.99	0.00	0.84	0.00	0.27	3
4				0	0.00	0.23	0.12	0.58	4
5					0	0.00	0.00	0.00	5
6						0	0.00	0.29	6
7							0	0.75	7
8								0	8

Finally, the annotation having the highest expected number of correct calls, i.e. the maximum a posteriori annotation, is: 

In this particular case, the Viterbi annotation is actually slightly better than the MAP annotation in recovering the truth.

- b. Assume that we in a score based alignment problem, e.g. similarity alignment as defined on problem sheet 4, do not have two perfectly observed sequences s_1 and s_2 , but one perfectly observed sequence s_1 and a sequence s_2 observed with uncertainty such that $Pr(s_2[i] = \sigma) = p_{i,\sigma}$. How would you compute the maximum expected score of an alignment of s_1 and s_2 . (2 points)

As the score of an alignment is the sum of the scores of its columns, the expected score of an alignment is just the sum of expected scores of each column. Hence, the maximum expectation can be determined by a recursion very similar to what we saw on problem sheet 4:

$$E_{i,j} = \max\{E_{i-1,j-1} + \sum_{\sigma \in \{A,C,G,T\}} p_{j,\sigma} w(s_1[i], \sigma), E_{i,j-1} - g, E_{i-1,j} - g\}$$