Determining RNA Secondary Structure

Assignment 5

Caitlin Ross
Computer Science Department, Rensselaer Polytechnic Institute
rossc3@rpi.edu

ABSTRACT

This work predicts the RNA secondary structure of multiple RNA sequences that code for prions. First MUSCLE is used to get the alignment sequences. Then the aligned sequences are used in a comparative sequence analysis to predict the secondary structure of the consensus sequence. The program developed here outputs the consensus sequence, along with the secondary structure in dot-bracket notation. This information can then be used as input into a third party tool that plots the secondary structure.

1. PROBLEM STATEMENT

The problem here is to take multiple RNA sequences and predict the secondary structure. The RNA sequences given are genes that code for prions. RNA secondary structure is important because this structure tends to be more conserved than the actually RNA sequence. Thus the secondary structure is more important than actual sequence in understanding the function of the RNA.

2. METHODS

We were given 42 RNA sequences in the PRP.fa fasta file. First we use the MUSCLE mutli-sequence aligner to align the sequences. This results in 42 alignments that are 42 bases long. Then we wrote a Python program to determine the secondary structure from the RNA alignments. After reading in the alignments with BioPython, the consensus sequence for the alignments is found. The consensus sequence is found by counting each symbol $b \in \{A, C, G, U, -\}$ in each column and taking the symbol with the most occurrences in each column.

Next we use comparative sequence analysis for the RNA secondary structure prediction. For this, we must first calculate the mutual information matrix, which looks at pairwise joint frequencies in the alignments. The mutual information matrix M is an $L \times L$ matrix, where L is the number of columns in the alignments. Then for aligned columns i and j, we have

$$M_{ij} = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}, \quad (1)$$

where $x_i, x_j \in \{A, C, G, U\}$ for columns i and j, respectively. f_{x_i} is the frequency of the base x_i observed in column i, whereas $f_{x_i x_j}$ is the pairwise frequency of one of the possible base pairs observed in columns i and j. For this, we only consider the base pairs AU, UA, CG, GC, GU, and UG to be valid base pairs for the calculation of M_{ij}. For all other base pairs, $f_{x_i x_j} = 0$. In the case where any frequencies in equation (1) are 0, we set $\log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}} = 0$. For any $M_{ij} < 0$, we set $M_{ij} = 0$.

Now we use dynamic programming to calculate the maximum mutual information secondary structure from M_{ij}. For this, we define a matrix D, which is also an $L \times L$ matrix. We start by initializing $D_{ij} = 0, \forall i, j \in L$. Then D is calculated by the following algorithm (note that we assume a starting 0-index). The end result is D, which is an upper triangular matrix.

for i = L -1..0 **do**
  for j = i+3..L **do**

$$D(i,j) = max \begin{cases} D(i+1, j) \\ D(i, j-1) \\ D(i+1, j-1) + M_{ij} \\ max_{i<k<j}[D(i,k) + D(k+1, j)] \end{cases}$$

  end for
end for

At this point, we use a back trace algorithm to trace through matrix D to find the base pairs for the RNA secondary structure. The code for the BackTrace algorithm given in the nussinov.py program on the course webpage was used here. The BackTrace() function uses the original mutual information matrix, M_{ij}, the matrix D, and the consensus sequence found earlier to find the pairs. After getting the list of pairs, the secondary structure can be determined. Here we find the dot-bracket notation of the secondary structure. Matching parentheses in positions i and j indicate a base pairing between bases i and j, otherwise there is just a dot. We determine the dot-bracket format using the function StructureFromPairs() from the nussinov.py program, which takes in the pairs found previously.

Now the secondary structure can be plotted from the dot-

bracket notation. We use the website http://nibiru.tbi.univie.ac.at/forna/ for plotting. We provide the consensus sequence with the dot-bracket notation.

3. RESULTS

The consensus sequence found by the program is CCAUGGUGGUGGCUGGGGACAGCCUCAUGGUGGU GGCUG–. The secondary structure found, in dot-bracket notation, is '...(..(((..)(..)).).)(((((..)))(..)..).....'. Figure 1 shows the RNA secondary structure plotted. Red lines show the base pairs.

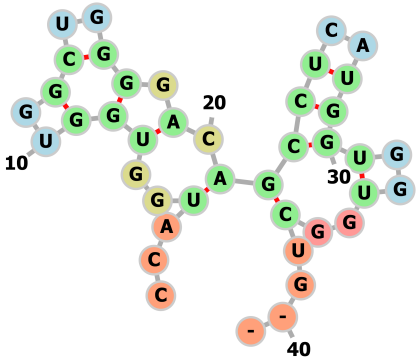

Figure 1: Plotted RNA secondary structure for the consensus sequence.