Übung: RNA Structure Prediction Tools

SS 2007 Übung zu RNA-Biochemie (RNA) (00.939)

1 Addresses

All sequences/ structures for the Übung are in `/home/sbienert/data`.

There are web interfaces for all of the secondary structure prediction tools. The `mfold` server is `http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi`. `RNAfold`, part of Vienna RNA Package is located in beautiful Austria at `http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi`. `CONTRAfold` is at `http://contra.stanford.edu/contrafold/server.html`. In a later part of the Übung, we will use `CONTRAfold` again, but from a locally installed copy in `/home/sbienert/bin`.

2 Introducing the Prediction Tools

The RNA sequences for this section are in `/home/sbienert/data/part01`.

This is the first part of an Übung which will take 3 weeks. In this part, there is no particular single sequence to use. Instead, you should try the tools on several sequences. The FASTA files in the directory above, have been chosen to show certain differences and similarities of the tools.

The main task, beside using the tools, of this part of the Übung is to compare their output in terms of (1) the predicted structures and (2) their output capabilities. Comparing the tools means considering the calculated energies and using your own method to quantify differences in the structures. For each tool, you will inspect the influence of certain parameters or methods and have a closer look at the data produced besides the structures.

2.1 Using the Tools

2.1.1 The `mfold` server

Visit the `mfold` web page and find the box for sequence input near the middle. Although the page accepts FASTA format, it loses the title. This should be put in the field above. The remaining web form provides several parameters concerning the algorithm and the output of mfold. These parameters are explained in the linked, online help.

The output of an mfold run includes many details in multiple formats. Most of the items on the output list have a link "Definition" connected to online help. The first item on the output page is the energy dot plot, which should be compared with the RNAfold plot within this Übung. The next part of the output data delivers more details about the structure and sequence values. For the Übung, the next interesting item is the structure section. mfold not only calculates the minimum free energy structure but also a few sub-optimal structures with an energy near the minimum free energy. All of these structures are listed on the result page with links for viewing in different picture formats and notations. Beside diagrams of the secondary structure, the *Vienna notation* (parenthesis notation) is important to us because RNAfold and CONTRAfold also produce this information. If there are sub-optimal folds, mfold can also compare their dot plots at the end of the result page.

2.1.2 The RNAfold server

Although the mfold server has a colourful background, the RNAfold server comes with absolutely no frills. The input mask for the sequence is similar to the mfold server. In the parameter section, the first choice is the folding algorithm. This is either a *free energy minimising* or a *partition function* algorithm. The parameters below the mask are meant to tune the thermodynamic parameters used with the folding method. The server can send results back by email, but this is only necessary for sequences longer than 400nt.

The results page of the RNAfold web server shows the predicted structure in Vienna notation at the head of the page. The energy of this particular structure is shown immediately afterwards. If the prediction was made using the partition function approach, the ensemble energy is presented next and a link to a dot plot is offered by the server. Following the information about the enthalpy of the structure, the remaining part of the page is filled with nice graphics: First, a mountain plot of the base pairing is presented, containing the curve of the minimum free energy structure and the partition function curve, if chosen as the folding method. This plot is followed by a slightly interactive picture of the minimum free energy structure.

2.1.3 The CONTRAfold server

The CONTRAfold server comes with fancy colouring and nice graphics but the least customisable interface. In exchange, everything is mostly self-explanatory and divided into 4 easy-to-follow steps. First, the sequence is entered in FASTA format (just copying the plain sequence will lead to an error message) or can be uploaded as a file. Then the output channel may be chosen and finally, the output format and base pairing behaviour can be manipulated. The end of the form presents buttons to start the folding process or reset the settings.

The result from the CONTRAfold server is just a picture of the structure and the structure presented in Vienna notation or a list of positions, nucleotides and bonding partners.

2.1.4 CONTRAfold in the Shell

In the shell CONTRAfold is invoked as "contrafold". Without any parameters, this produces a usage message. To perform a structure prediction, CONTRAfold has to be called in predict mode together with a FASTA file: `contrafold predict filename`. Important parameters for the folding are "--gamma" and "--viterbi". "--gamma" is a parameter to be explored in this Übung and "--viterbi" switches to an alternative folding method.

The output of CONTRAfold on the command line is reduced to the structure given as parenthesis string.

2.2 Assignment

1. Gather your results (make a nice tar-archive or an ASCII-file or whatever you like, but no Open Office document because this is what we do not like) and sent it to bienert@zbh.uni-hamburg.de. This will not be formally marked, but is just a check that you participated in the Übung.

2. Choose 4 sequences from the data directory and explore how the sub-optimal structures produced by mfold differ from each other. To start, you should look at the dot plot showing all structures. This will give you an overview of the base pairing and the differences of the structures. Describe and write down these differences. Aside from the comparison of the base pairing, you should compare the structural elements which the different foldings consist of. This can be done by enumerating the features together with their start and stop base. The question of interest is: Beside different base pairs, do sub optimal structures assemble to a similar shape (e.g. secondary structure elements) or are the conformations highly diverse at a similar energy level? Now you should be able to answer the last question: What is the idea of producing sub-optimal structures?

3. Take 4 sequences and run RNAfold on them. Run each sequence (1) with the minimum free energy approach and (2) with the partition function approach. Compare the base pairing and the structural features for each sequence. Write down any difference you find. Explain the technical difference of the two methods. How are they connected to the sub-optimal structure strategy of mfold?

4. Take one long sequence and try to guess what influence the γ-factor takes on the results of CONTRAfold. A glance at the original paper located on the web server might help! For this question, you have to use the command line tool. Calculate the secondary structure at a low, high and medium γ-factor. Describe the differences between the results and explain the effect of the γ-factor.

5. Design a method to compare predicted base pairings. This could be something like an edit distance. You will use this in the next question.

6. Find a sequence in the data directory which yields almost the same results with mfold, RNAfold and CONTRAfold. To compare results, you should look at the RNA structural elements (helices, loops, etc..), the calculated free energy and the base pairing. Use the measure you constructed in the previous step.

7. Find a sequence in the data directory which yields very different results for in mfold, RNAfold and CONTRAfold. Use the same criteria as in the previous question.

8. Why does CONTRAfold produce no energy for the foldings calculated?

9. What do the dot plots of RNAfold and mfold present?

10. What is described by the mountain plot of RNAfold? Why does the partition function curve have smooth changes of the slopes while the minimum free energy curve has sharp changes?