RNA Secondary structure prediction

Stéfan Engelen and Fariza Tahi

Laboratoire La.M.I. - UMR 8042.
Université d'Evry-Val d'Essonne, CNRS, Genopole
E-mail: {sengelen,tahi}@lami.univ-evry.fr

Structural RNA are important for regulatory, catalityc or structural roles in the cells. They can act alone to catalyse RNA processing. They can also form protein-RNA complexes like the ribonuclease P which has a role in tRNA processing. The secondary structure is composed of all the Watson-Crick pairings, AU, GC, and the Wobble pairing GU. The consecutive pairings form helices and could be interleaved to make pseudoknots. The knowledge of secondary structure is essential to understand the relations between structure and function of the RNA.
One approach to predict the secondary structure of RNA is to search covarying residues which maintain the Watson-Crick pairings. This approach is called the comparative approach and consists to retrieve mutation informations from homologous sequence alignments.

We have developped an algorithm called P-DCfold implementing the comparative approach. It is based on the "divide and conquer" approach which allows a low complexity in time. This method segments the prediction in under less complex problems which underline the fact that the RNA secondary structures are also segmented in under fields. The helices are searched recursively from the "most significant" to the "less significant". Thanks to this approach and to the criteria we use (criteria of minimum length of searched helices and number of the compensatory mutations in these helices), the algorithm allows to obtain predictions of very good quality, since almost all the structure helices are predicted and no false positive helices are selected. P-DCfold algorithm allows also to predict the pseudoknots appearing in the structure. It searches for all kinds of pseudoknots in a complexity in time of $O(n^2)$ when algorithms existing in litterature are of complexities higher than $O(n^4)$ for the prediction of only certain kinds of pseudoknots.
An important point when using the comparative approach for the predicion of a RNA secondary structure is how to choose the homologous sequences to use for the comparison. More precisely, how to choose the ones which allow to obtain good predictions. A preprocessing of available homologous sequences makes it possible, by assigning scores to these sequences. We expound an algorithm, called SSCA, which measures the interest of a sequence. The measurement is based on evolutionary model in helices regions. This algorithm is in complexity in time of ($O(n^2)$).

Thus we obtain a complete system which implement efficiently the comparative approach for the RNA secondary structure prediction and which goes to the prediction of the tertiary structure, since pseudoknots can be considered as elements of the tertiary structure instead of the secondary one.

The efficiency of our system concerns its complexity in time ($O(n^2)$), its complexity in space (less than $O(n^2)$) and above all the quality of the predictions. It has been tested in several RNAs: tmRNA, RNaseP, 5S RNA, U1 RNA, SRP RNA, tRNA, 16S RNA and 23S RNA. In all these examples, more than 80 percent of the helices have been well predicted as well as all the pseudoknots and the predictions have been done in immediate time for small RNAs and less than 5 seconds for longuest ones (16S and 23S RNA, respectively of about 1500 and 2400 nucleotides).

Finally, several applications of the system could be considered: prediction of unknown structures, search for characteristic structures ("hairpin" and "hammerhead" structures), prediction of structures of certain regulating areas (5 ' and 3' areas of genes, introns ...), etc.