RNA secondary structure and internal loops

A.Yu. OGURTSOV, S.A. SHABALINA, A.S. KONDRASHOV

National Center for Biotechnology Information, NIH, 45 Center Drive, Bethesda, MD 20892-6510, USA

ogurtsov@ncbi.nlm.nih.gov

M.A. ROYTBERG

Initute of Mathematical Problems in Biology, Pushchino, Moscow Region, 142290, Russia

roytberg@impb.psn.ru

Evaluating of possible internal loops is one of key steps in predicting the optimal (i.e. possessing minimal free energy) secondary structure of an RNA molecule. The best algorithm available for this task (Lingso et al. 1999) runs in time $O(L^3)$, where L is the length of the RNA. Eppstein *et al*. (1992) investigated a closely related task of finding the optimal secondary structure that does not contain multi-branched loops, assuming that the penalty $f(s)$ for an internal loop is a convex or concave function of the total number s of unpaired nucleotides which constitute the loop. The authors proposed an elegant algorithm with run-time $O(L^2 * \log^2(L))$, or $O(M * \log^2(L))$, if only M possible combinations of paired nucleotides are considered. However, the commonly used penalty for an internal loop is a function of two variables, $F(s,d) = f(s) + g(d)$, where $f(s)$ is concave and $g(d)$ is a function of the difference d between the lengths of two unpaired regions which constitute the loop.

We propose a modification of the algorithm of Eppstein *et al* (1992) which uses $F(s,d)$ as a penalty function for internal loops and has run-time $c*M* \log^2(L)$, of the same order as the run-time of the original algorithm. We use the standard assumption that asymmetry penalty function $g(d)$ differs from a constant only at a small number of points ($g(d)$ = const when d is large enough). Then, the coefficient c is proportional to this number. Using the new algorithm, we evaluate energies of all possible internal loops during the search for optimal RNA structure in $c*M* \log^2(L)$ time, which improves time bound of Lingso et al (1999). We also propose two algorithms, with the same order of run-times that construct sets *PAIRED* and *HAIRPIN* of conditionally optimal multi-branched loop-free structures. The *(p, q)*-paired-optimal structure is an optimal multi-branched loop-free structure in which nucleotides p and q are paired. A *(p, q)*-hairpin-optimal structure is an optimal multi-branched loop-free structure in which nucleotides p and q form a pair that closes a hairpin loop. The set *PAIRED* contains a *(p, q)*-paired-optimal structure for every possible pairing (p, q), and the set *HAIRPIN* contains a *(p, q)*-hairpin-optimal structure for every possible pairing (p, q)

ACKNOWLEDGEMENTS

The work has been supported by the Russian Foundation for Basic Research (project nos. 03–04–49469, 02–07–90412), by grant from the RF Ministry for Industry, Science, and Technology 20/2002, ECO-NET program and Jumellage program CNRS-IUM. . The work was partly performed during the M.A.Roytberg visit to NCBI (January–March, 2004).

REFERENCES

1. Eppstein,D., Galil,Z., Giancarlo,R. and Italiano,G.F. (1992) Sparse Dynamic Programming II: Convex and Concave Cost Functions. Journal of ACM, **39**, 546-567.
2. Lyngsø,R.B., Zuker,M. and Pedersen,C.N. (1999) Fast evaluation of internal loops in RNA secondary structure prediction, Bioinformatics, **15,** 440-445.
3. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: *RNA Biochemistry and Biotechnology*, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers. http://www.bioinfo.rpi.edu/~zukerm/seqanal/FEBS98–html.html