ALGORITHM FOR PREDICTING THE EVOLUTIONARILY CONSERVED SECONDARY STRUCTURES OF RNA

Vorobiev D.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: denis@bionet.nsc.ru

Key words: RNA, secondary structure, genetic algorithm, phylogenetic approach

Resume

Motivation: Phylogenetic algorithms for predicting the RNA secondary structure (SS) offer advantage over thermodynamic methods in the presence of representative samples of isofunctional sequences of RNA.
Results: In this paper, a new phylogenetic algorithm for constructing the RNA SS from a multiple alignment of sequences using a genetic algorithm approach is presented.
Availability: the program is available from the author.

Introduction

The RNA SS is predicted using approaches referring to two groups: thermodynamic and phylogenetic algorithms. The former are most extensively used (Mathews et al., 1999; McCaskill, 1990; Gultyaev et al., 1995) because their operation needs no other data than RNA sequences. The experimental evidence for either the paired or free states of separate bases simplifies calculation of the RNA secondary structure. However, these data are not always available. The phylogenetic algorithms are used more rarely because they need a consistent sampling of RNA sequences with a similar function and structure.

The errors of thermodynamic algorithms usually occur for one of three reasons: the inaccuracy of energy rules, the impossibility of taking into account the energetics of either tertiary or RNA-protein interactions and the disregard of the peculiarities of folding kinetics. The last problem was partly solved by the method of RNA ensemble kinetics modelling (see, Mironov et al., 1986).

The phylogenetic algorithms use not only RNA thermodynamics but also information on the phylogenetic conservatism of RNA SS that perform a similar function (Eddy, Durbin, 1994; Gorodkin et al., 1997; Chen et al., 2000; Hofacker, Stadler, 1999). This offers them a fundamental advantage and in many cases, allows them to get around the drawbacks typical of purely thermodynamic methods.

In this paper, a pilot variant of the algorithm for predicting the evolutionarily stable RNA is proposed. Actually, it is similar to genetic algorithm (GA) used to predict the SS of a single RNA molecule (Vorobiev et al., this volume). However, it allows one to take into account the conservatism of the primary (high homology of sample sequences) and secondary (coadaptive substitutions) structures.

Methods and Algorithms

Assume that we have a multiple alignment of N sequences of RNA of length L (problems on the construction of such alignment will be discussed below). The RNA SS is unambiguously given by a set of helices. In the unit RNA sequence, the helix is represented by the three <x,y,l> (where x is the coordinate of the left end of the helix arm, y is that of the right end of the helix arm and l is the helix length) such that the bases at the sequence positions x+a and y-a (for all 0 ≤ a <1) form the complementary pairs (AU, GC, or GU).

In the first step of the algorithm, we make up the list \{h\} of helices common for the sequences from the alignment. The helix is included in the list \{h\} if a continuous (without gaps) helix with the same coordinates is present in more than Tshare% of sequences.

In the second step, the list \{h\} is send to the input of the genetic algorithm identical to that used by us to predict the SS of the single RNA molecule (Vorobiev et al., this volume). On the output of the genetic algorithm there is a set of helices some of which cannot exist in all sequences. For these sequences, the forbidden helices are broken either partially or completely.
For each sequence, we can optionally complete the construction of helices, reducing the energy of its SS, by the steepest descent technique.

**Implementation and Results**

The algorithm has been realized in language C of standard ANSI. It was tested using RNA of HIV-2 rev response elements (Fig.) and 5S RNA (Chen et al., 2000). In these cases, more than 90% of nucleotide pairs forming a real structure were predicted.

**Discussion**

The main problem of our approach as of all other similar algorithms is the need for preliminary construction of a multiple alignment of RNA sequences. This restricts the range of approach applicability to the cases of high homology of RNA sequences. The algorithms like ClustalW (Thompson et al., 1994), usually used to construct the alignment, are adequate to the problem of SS consensus construction only in the case of high sequence homology in a sample. Therefore, it is of prime importance that our algorithm is supplemented with a method for construcing multiple alignment that takes into account the conservatism by both the primary and secondary structures.
Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grant № 01-07-90376), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), National Institutes of Health USA (№ 2 Ro1-HG-01539-04A2), the Department of Energy USA (№ 535228 CFDA 81.049). The author is grateful to N.A.Kolchanov and I.I.Titov for helpful discussions and to G.A.Ilyina for assistance in translation.

References