CHINESE JOURNAL OF BIOTECHNOLOGY
Volume 24, Issue 7, July 2008
Online English edition of the Chinese language journal

Cite this article as: Chin J Biotech, 2008, 24(7), 1140−1148.

RESEARCH PAPER

Consensus RNA Secondary Structure Prediction Based on Support Vector Machine Classification

Yingjie Zhao, and Zhengzhi Wang

College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China

Abstract: The comparative sequence analysis is the most reliable method for RNA secondary structure prediction, and many algorithms based on it have been developed in last several decades. This paper considers RNA structure prediction as a 2-classes classification problem: given a sequence alignment, to decide whether or not two columns of alignment form a base pair. We employed Support Vector Machine (SVM) to predict potential paired sites, and selected covariation information, thermodynamic information and the fraction of complementary bases as feature vectors. Considering the effect of sequence similarity upon covariation score, we introduced a similarity weight factor, which could adjust the contribution of covariation and thermodynamic information toward prediction according to sequence similarity. The test on 49 Rfam-seed alignments showed the effectiveness of our method, and the accuracy was better than many similar algorithms. Furthermore, this method could predict simple pseudoknot.

Keywords: comparative sequences analysis; RNA secondary structure; support vector machine; similarity weight factor

Introduction

Similar to other biological macromolecules, the secondary structure of RNA is essential for understanding its function. The interest in the role of RNA has increased dramatically because of many studies on functional RNA. Generally, there are two approaches for RNA structure determination: the biological experiment and the computational prediction. To deal with the sharply increasing sequence data, it is unpractical to determine the secondary structure by using the experimental method because it is expensive and time-consuming. The computational method, profiting from its briefness and efficiency, has become the preferred approach. As one of the classical issues in computational biology, although many algorithms have been proposed during the last decades, the RNA secondary structure prediction is still an open problem. Most proposed algorithms can be classified into three kinds: (1) the minimum free energy method (MFE)[1–3] based on the thermodynamic theory; (2) the statistical learning method[4–7]; (3) and the comparative sequence analysis method[8–11] based on phylogeny. Method 1, as the major approach to predicting the structure from a single sequence, tolerates the hypothesis that the structure which holds minimum free energy is most steady. The parameters of MFE come from experimental determination and extrapolation. Limited by the existing measuring technique and method, it is difficult to improve the prediction accuracy of this method. Furthermore, the energy of the actual structure is sometimes not the minimum. Method 2 employs machine learning approaches to model structure a formation, from the database of a known structure, and then predicts an unknown structure from the sequence, using this model. This method includes stochastic context-free grammars (SCFG)[4], the genetic algorithm (GA)[5], and the statistical sampling algorithm[7]. However, intensive

Received: November 23, 2007; Accepted: December 27, 2007

Corresponding author: Yingjie Zhao. Tel: +86-731-4574991; E-mail:matriz@163.com

computation is the main disadvantage of this method. When a set of evolutionary or structure-related sequences are used, method 3 has proved to be the most reliable. It predicts the formation of a base pair through detecting numbers of covariation in a sequence alignment. The bottleneck of this method is the requirement of a certain amount and a similarity of sequences, and moreover, the existence of a common structure in the alignment is a prerequisite. Consequently, many combined plans, which combine thermodynamic and phylogenetic information (the mutual information score or the covariation score), have been proposed, to improve the prediction accuracy, such as, Hxmatch[12], construct[13,14], contrafold[15], hxplot[16], and ILM[17]. These combined methods educe a candidate base pair matrix in the first step, whose entry is 1 if the corresponding columns of the alignment are expected to form a base pair, otherwise it is 0. Subsequently they assemble in a final secondary structure by their various approaches. Mutual mutation is still the foundation of these methods.

The comparative sequence analysis methods fall into three general categories[18]: the first one predicts common structure from a given alignment, which has been known before prediction, including: Pfold[5,19], RNAalifold[20], ILM[17], and KnetFold[21,22]; the second one utilizes the so-called Sankoff algorithm[23], which deals with sequence alignment and structure prediction simultaneously, including: Foldalign[24,25], Dynalign[26], and PMcomp[27]; the last one initially predicts the structure of every sequence in a considered family, and then extracts the common structure by aligning those structures, including: RNAforester[28] and MARNA[29,30]. Most comparative methods belong to the first class, and they can attain alignment by using normal sequence alignment programs or adopting reference structures from the database directly. High quality alignment is required when using this method. The drawback of the other classes is their intensive computation.

Given an aligned RNA sequence family, the common secondary structure prediction can be considered as a classification problem, to judge whether any two columns in the alignment correspond to a base pair, using the information provided by the alignment. The authors' method, based on the first class of comparative sequence analysis, predicts the potential paired sites, by employing the Support Vector Machine (SVM). The feature vectors of the classifier are composed of the covaration score, the fraction of the complementary nucleotides, and the consensus probability matrix. Then the common secondary structure is assembled from those sites using the stem combining rules.

1 Algorithms

Given a sequence alignment, the authors first compute the feature vectors of every pair of columns, and then predict the potential paired sites using the trained model in advance, and come up with a base pair probability matrix, finally, building up the secondary structure according to the base pairing rules and the stem combining rules.

1.1 Covariation score

Mutual information score[31,32] is generally adopted to detect complementary mutation in the alignment quantitatively, but it does not work for the single conserved pair (Fig. 1). Therefore, the covariation score that Hofacker[20] defined, to find the complementary mutation, is borrowed:

$$C(i,j) = \sum_{XY,X'Y'} f_{i,j}(XY) D(XY,X'Y') f_{i,j}(X'Y') \qquad (1)$$

Where $D(XY,X'Y')$ is the Hamming distance between XY and $X'Y'$, and the sum has taken over all the complementary base pairs. The covariation score distinguishes the conserved pairs, the pairs with consistent mutations, and the pairs with compensatory mutations. In addition, Hofacker has introduced an inconsistent sequence penalty [20] for increasing the signal-to-noise ratio:

$$q(i,j) = 1 - f_{i,j}^{comp} - f_{i,j}(-\bullet-) \qquad (2)$$

Where $f_{i,j}^{comp}$ is the frequency of complementary base pairs in columns i and j. This penalty is usually subtracted from $C(i,j)$. It has been observed that this improves the structure prediction for the small number of sequences.

The single conserved pair in the evolutionary process is shown as Fig.1. The ith column corresponds to a conserved G (or U), the jth column corresponds to an alternating C and U (or A and G). In this case, the mutual information of (i,j) is 0, but the covariation score is 4. Obviously, the latter can detect the bias toward the complementary base pairs better than the former.

$$\begin{array}{c} \qquad\quad i \qquad\quad j \\ \times\times\times\times G \times\times\times U \times\times\times\times \\ \times\times\times\times G \times\times\times C \times\times\times\times \\ \times\times\times\times G \times\times\times U \times\times\times\times \\ \times\times\times\times G \times\times\times C \times\times\times\times \end{array}$$

Fig. 1 Schematic of a single conserved pair[20]

1.2 Weighted average base pair probability matrix

Although the covariation score can find complementary mutation in the alignment, it fails to find the conserved base pairs, because the covariation score is 0 for those sites. This is also a collective disadvantage of all comparative sequence analysis methods. In most combined methods, it is common to introduce the thermodynamic character to improve the prediction accuracy. Besides utilizing the thermodynamic parameters decided by the experiments directly, the base pair probability matrix of a given sequence, calculated with