Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 50–57
www.elsevier.com/locate/compbiolchem

Importance of RNA secondary structure information for yeast donor and acceptor splice site predictions by neural networks

Sayed-Amir Marashi [a,1], Hani Goodarzi [a,1], Mehdi Sadeghi [b,e,*], Changiz Eslahchi [c,e], Hamid Pezeshk [d,e]

[a] *Department of Biotechnology, Faculty of Science, University of Tehran, Tehran, Iran*
[b] *National Institute for Genetic Engineering and Biotechnology, Tehran-Karaj Highway, Iran*
[c] *Faculty of Mathematics, Shahid-Beheshti University, Tehran, Iran*
[d] *Department of Mathematics, Computer Sciences and Statistics, Faculty of Science, University of Tehran, Tehran, Iran*
[e] *Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran*

Received 5 September 2005; received in revised form 19 October 2005; accepted 19 October 2005

Abstract

Previously, Patterson et al. showed that mRNA structure information aids splice site prediction in human genes [Patterson, D.J., Yasuhara, K., Ruzzo, W.L., 2002. Pre-mRNA secondary structure prediction aids splice site prediction. Pac. Symp. Biocomput. 7, 223–234]. Here, we have attempted to predict splice sites in selected genes of *Saccharomyces cerevisiae* using the information obtained from the secondary structures of corresponding mRNAs. From Ares database, 154 genes were selected and their structures were predicted by Mfold. We selected a 20-nucleotide window around each site, each containing 4 nucleotides in the exon region. Based on whether the nucleotide is in a stem or not, the conventional four-letter nucleotide alphabet was translated into an eight-letter alphabet. Two different three-layer-based perceptron neural networks were devised to predict the 5′ and 3′ splice sites. In case of 5′ site determination, a network with 3 neurons at the hidden layer was chosen, while in case of 3′ site 20 neurons acted more efficiently. Both neural nets were trained applying Levenberg–Marquardt backpropagation method, using half of the available genes as training inputs and the other half for testing and cross-validations. Sequences with GUs and AGs non-sites were used as negative controls. The correlation coefficients in the predictions of 5′ and 3′ splice sites using eight-letter alphabet were 98.0% and 69.6%, respectively, while these values were 89.3% and 57.1% when four-letter alphabet is applied. Our results suggest that considering the secondary structure of mRNA molecules positively affects both donor and acceptor site predictions by increasing the capacity of neural networks in learning the patterns.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Splice site prediction; Gene prediction; Neural network; RNA structure; *Saccharomyces cerevisiae*

1. Introduction

Nowadays, complete genomic sequences of many eukaryotic species are available and genome projects of many other organisms are in progress. However, the huge amount of genomic information is useless unless its coding sequences are detected. Unfortunately, error-free deciphering of this code has not been achieved yet.

EST or mRNA sequences of many genes are now available and it is possible to highlight the corresponding regions on the genome. For many genes, at least a known homologous protein exists in other organisms. Using similarity searches one may detect these genes. Furthermore, assuming that coding sequences are more conserved than non-coding ones, it is possible to align genomic regions to find possible genes (Mathé et al., 2002). Unfortunately, roughly half of the genes in a novel genome are not detectable using such similarity-based approaches. This leads us to use "intrinsic" techniques to find these genes.

Finding a gene merely from the sequence requires that the program find the start site, all splice sites and the stop codon. If we were able to predict all splice sites accurately, we would be able to perform a highly reliable gene prediction (Pertea et al., 2001). Thus, improvement of splice site prediction methods directly enhances the predictive power of gene finders.

A complex of many proteins and small RNA molecules called spliceosome directs the splicing process and also contributes to

* Corresponding author. Tel.: +98 21 4458 0373; fax: +98 21 4458 0399.
E-mail address: sadeghi@nrcgeb.ac.ir (M. Sadeghi).
[1] These authors contributed equally to the manuscript.

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.10.009

other cellular processes (Staley and Guthrie, 1998; Rappsilber et al., 2002). This complex detects and interacts with different parts of a pre-mRNA molecule such as donor and acceptor splice sites and branchpoints, via which catalyzes the splicing reaction. The final aim of splice site prediction programs is to detect the splicing signals exactly as the spliceosome does.

During the last 20 years, different splice site prediction programs have been developed (Mathé et al., 2002). These programs basically rely on the accurate prediction of signal sequences and they are generally different in the way that they recognize the dependencies between different positions of a signal sequence.

Since it is generally observed that solely based on the sequence signals normally used, many splice sites are not detectable, it is believed that some additional features, not included in current models, must be essential in identifying the 3′ (and 5′) splice junctions of introns; presence of additional enhancer/suppressor elements and RNA secondary structure are of proposed candidates (Lim and Burge, 2001).

A large body of experimental evidence suggests that mRNA secondary structure can affect RNA splicing (Buratti and Baralle, 2004). Such observations led Patterson et al. (2002) to study the importance of predicted RNA structures in splice site prediction. In their work, by applying decision tree and support vector machine classifiers, they have shown that structure metrics like the secondary structure free energy and maximum helix forming probability can enhance the accuracy of prediction of acceptor sites. They reported that the role of RNA structure is not as vivid for donor sites.

Herein, we tried to predict RNA secondary structures for yeast full-length genes (i.e. starting from an AUG and ending in a stop codon). In contrast, Patterson et al. (2002) predicted the secondary structure of a very small (100-nucleotide) window around splice sites. However, this limitation clearly prevents the identification of potential long-range interaction within mRNA molecules. Fig. 1 shows the distribution of the linear distances between each pair of hybridized nucleotides when internal RNA structures of yeast genes were predicted by Mfold (see Section 2). From this histogram, it is evident that a large fraction of possible nucleotide interactions (∼41.5%) would be lost if we consider the hybridizations merely within 100-nucleoide windows. Although full-length genes may still fold differently from the real full-length mRNAs, they are clearly better approximations. Undoubtedly, the computational method is a ballpark figure itself.

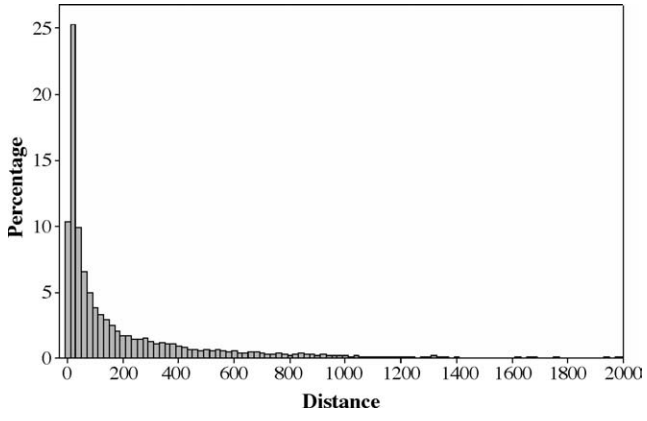

Fig. 1. Distribution of the linear distances of paired nucleotides in predicted RNA secondary structures of 154 yeast genes.

In this work, based on the states of nucleotides within predicted RNA secondary structures (stems versus loops), we translated the conventional four-letter alphabet of nucleotides into an eight-letter alphabet. It was shown that prediction of *both* donor (5′) and acceptor (3′) sites are significantly improved when this eight-letter alphabet, instead of the four-letter one, is employed to train and predict splice signals by neural networks. Besides the simplicity of this idea, application of such an eight-letter alphabet can be easily extended to other splice site prediction programs. This means that future gene prediction programs may take advantage of secondary structure prediction modules.

2. Materials and methods

2.1. Dataset

From Ares database (Grate and Ares, 2002), 154 yeast genes each containing exactly 1 intron were selected, after removing all genes with alternative splicing patterns, genes that did not encode proteins, genes with annotations that showed some degree of ambiguity, and genes containing introns starting from the beginning of sequence (i.e. 5′ UTR introns). Each gene in this dataset was started with an AUG and ended in a stop codon. From this dataset, 154 pairs of donor–acceptor site were obtained (the "real" set); in case of each gene, 2 GUs and 2 AGs other than the real sites were selected (the "decoy" set).

2.2. RNA secondary structure prediction

For all selected genes, the most stable RNA structure was predicted using Mfold Server (Zuker, 2003). When an mRNA secondary structure is predicted, it is possible to divide the nucleotides into two groups: those that take part in the base-pairing (and are located in Stems) and those that are not base-paired with other nucleotides (and are placed in Loops). Then, with the combination of {L, S} structures and the four-letter {A, U, C, G} nucleotide alphabet, all sequences were translated to an eight-letter alphabet $\{A_S, U_S, C_S, G_S, A_L, U_L, C_L, G_L\}$. These sequences and also the conventional four-letter sequences were used as the train and test datasets in the next steps. We also found the linear distances of all paired bases in the predicted secondary structures; for example, for nucleotide *i* base-paired to nucleotide *j*, $|i-j|$ was considered as the linear distance of these two nucleotides.

2.3. Computational procedure

2.3.1. Designing a neural network for 5′ splice site prediction

The complete structure of the neural network devised for 5′ splice site prediction is based on a three-layer perceptron

Download English Version:

https://daneshyari.com/en/article/15414

Download Persian Version:

https://daneshyari.com/article/15414

Daneshyari.com