A SIMPLE METHOD FOR FINDING STRUCTURALLY SIMILAR RNAS USING TWO-DIMENSIONAL DISCRETE CONVOLUTION

Byung-Jun Yoon

Dept. of Electrical & Computer Engineering
Texas A&M University
Collge Station, TX 77840-3128, USA
Email: bjyoon@ece.tamu.edu

ABSTRACT

As many noncoding RNA (ncRNA) families have well-conserved secondary structures, it is important to consider structural similarity when searching for RNA homologues. However, algorithms for detecting structural similarity tend to have high computational costs, making them unsuitable for large-scale genome screening. In this paper, we introduce a simple method that can find structurally similar RNAs at a low computational cost. The method uses matched filtering of base-pair matrices to identify structurally similar RNAs in a given database. As the matched filtering involves a simple two-dimensional discrete convolution, it has a relatively low complexity of $O(L^2N)$ for any kind of RNA secondary structure, where L is the length of the target RNA and N is the number of base-pairs in the reference RNA.

1. INTRODUCTION

Noncoding RNAs (ncRNAs) are functional RNAs that do not code for proteins. Extensive research on the genomes of various organisms has shown that there exist many ncRNAs that actively participate in various biological processes [6, 8]. As RNAs can directly interact with other RNA and DNA molecules, they are especially useful in regulatory mechanisms. In fact, it has been observed that ncRNAs play pivotal roles in controlling diverse gene regulatory networks in higher organisms [4, 6]. Systematic research on ncRNAs has begun relatively recently, and it is believed that there are still many ncRNAs that have not been identified yet [6, 7]. Therefore, it is important to develop efficient computational tools that can be used for identifying ncRNAs.

RNA similarity search (or RNA homology search) is an effective method for finding new ncRNAs. It tries to identify new members of a ncRNA family, by searching for RNAs that look similar to the known members in the given family. Many ncRNA families have well-conserved secondary structures that are shared by their members [2, 11]. For this reason, it is important to consider both structural similarity and sequence similarity when performing an RNA search [11]. Until now, various statistical models have been proposed for RNA similarity search, where the *covariance model* (CM) [3] and the *profile context-sensitive HMM* (profile-csHMM) [10, 13] are such examples. One practical problem of these models is the high computational cost for using them. The algorithms that can be used with these statistical models have a complexity of $O(L^3M)$ for typical RNAs, where L is the length of the target RNA and M is the number of states in the model that is proportional to the length of the reference RNA [2, 11]. For more complex RNAs, the computational complexity can be $O(L^4M)$ or even higher [13]. In order to overcome this problem, several methods have been proposed so far, mainly based on sequence-based *prescreening* [9, 12].

In this paper, we introduce a simple method that can be used to find structurally similar RNAs at a relatively low computational cost. Given a reference RNA, a matched filter matrix is constructed based on the structure of the RNA, and we perform matched filtering using this matrix to detect structurally similar RNAs. This method has the following advantages. Firstly, this method can be used for finding the structural homologues of *any* RNA family, including those with pseudoknots. Secondly, despite its generality, this method has a relatively low computational cost as it just involves a two-dimensional discrete convolution of very simple matrices. In fact, the computational complexity of the search algorithm will be only $O(L^2N)$, where L is the length of the target RNA and N is the number of base-pairs in the reference RNA. This method can also be used in combination with other sequence-based methods [9, 12] for better performance. Structurally similar RNAs that have been identified by this method may also be passed to a more complex model (such as a CM or a profile-csHMM) for further inspection, to improve the specificity of the search.

2. FINDING STRUCTURALLY SIMILAR RNAS

The structural similarity search proceeds as follows. Assume that we are given a reference RNA of length L_r, whose structure is known. Based on this base-pairing information, we construct a lower-triangular *matched filter matrix* $\mathbf{S} = \{s_{ij}\}$ of size $L_r \times L_r$ as follows

$$s_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j\ (i > j) \text{ form a base-pair.} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Now, for a target RNA with an unknown structure, we construct an $L \times L$ upper-triangular *base-pairing matrix* $\mathbf{P} = \{p_{ij}\}$ as follows

$$p_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \ (i < j) \textit{ can } \text{form a base-pair.} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Now, we compute the two-dimensional discrete convolution[1] of $\mathbf{S}$ and $\mathbf{P}$ to obtain

$$\mathbf{Y} = \mathbf{S} * \mathbf{P}. \tag{3}$$

Once $\mathbf{Y} = \{y_{ij}\}$ has been computed, we find its largest element $\lambda = \max_{i,j}(y_{ij})$, and normalize it by $\sum_{i,j} s_{ij}$ to obtain the structural similarity score σ

$$\sigma = \frac{\lambda}{\sum_{i,j} s_{ij}}. \tag{4}$$

Note that, $\sum_{i,j} s_{ij}$ is identical to the number of base-pairs N in the reference RNA, and it is the maximum possible value of λ. Therefore, we always have $0 \leq \sigma \leq 1$, where $\sigma = 1$ indicates high structural similarity between the reference and the target RNAs.

In general, the two-dimensional discrete convolution in (3) has a complexity of $O(L^2 L_r^2)$. However, $\mathbf{S}$ is a sparse matrix with only N non-zero elements, reducing the overall complexity to $O(L^2 N)$. This computational complexity is with quadratic in L, but this is practically not a problem, as it simply corresponds to the addition of N matrices of size $L \times L$. We can also use fast convolution algorithms based on multidimensional FFT to reduce the complexity [1].

3. EXPERIMENTAL RESULTS

We applied the matched filtering method elaborated in the previous section to two RNA families, HDV_RIBOZYME and TOMBUS_3_IV, in the Rfam database [5]. Since both families contain pseudoknots, CMs are incapable of handling these RNAs. Although we can use profile-csHMMs for these RNAs, the computational complexity would be $O(L^4 L_r)$, which is quite high. In our experiments, we used the RNAs in the so-called *seed alignments*, as they have relatively reliable structure annotations. For each family, we first chose a reference RNA among the members and computed the structural similarity score σ for all other members. In order to obtain a reliable estimate of the score distribution, we repeated this experiment for all members. In addition to this, we also estimated the score distribution of randomly generated RNAs for comparison.

The resulting *cumulative score distributions* are shown in Fig. 1. As we can see in this figure, the score distributions of the real RNAs are well-separated from the distributions of the random RNAs. This shows that we can use the structural similarity score σ to effectively filter out the sequences that are structurally dissimilar from the reference RNA. Let us first consider the HDV_RIBOZYME family. If we would choose a threshold of $\sigma^* = 0.42$ for prescreening the target RNAs, we would be able to filter out 98% of the unrelated RNAs at no loss of sensitivity. If this method were used for prescreening the database, where a more complex model is applied only to the RNAs that pass this prescreening stage, this would increase the average search speed by 50 (=1/0.02) times, compared to the case when the high-complexity model is directly used[2]. For TOMBUS_3_IV family, a threshold of $\sigma^* = 0.46$ would reject more than 99.5% of the unrelated RNAs without degrading the sensitivity of the search. This would make the search speed more than 200 (=1/0.005) times faster, while achieving the same prediction accuracy as using the more complex model directly.

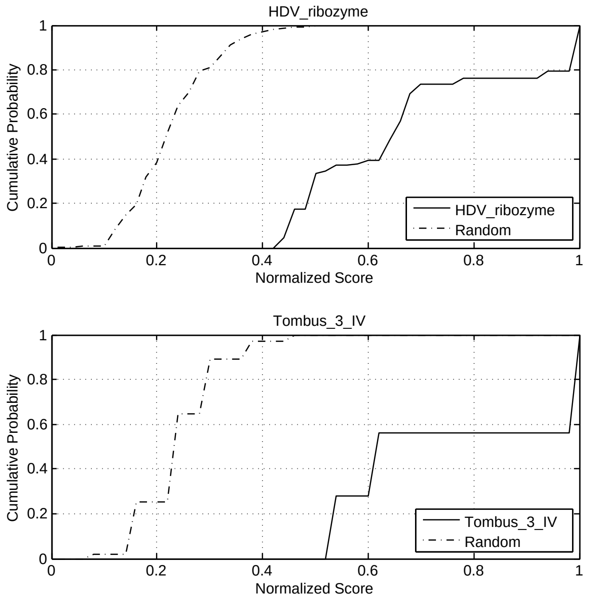

Figure 1. Cumulative distribution of the structural similarity score σ for real and randomly generated RNAs. (Top) Score distribution of the HDV_RIBOZYME family. (Bottom) Score distribution of the TOMBUS_3_IV family.

4. CONCLUSION

As shown in this paper, the matched filtering approach can effectively detect structurally similar RNAs at a low computational cost. Although matched filtering by itself cannot yield as accurate prediction results as the more powerful models (such as the CM and the profile-csHMM), we expect that it will be able to make RNA similarity searches faster when combined with these models. Future plans include the investigation of how to combine the matched filtering method with other sequence-based prescreening methods and/or a more advanced statistical model such as the profile-csHMM.

[1] Note that $*$ in (3) stands for convolution and not matrix multiplication.

[2] It is assumed that the CPU time for using the complex model is significantly longer than the CPU time for using the matched filtering method.

5. REFERENCES

[1] D. Dudgeon and R. Merserau, *Multidimensional Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[2] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.

[3] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models", *Nucleic Acids Research*, vol. 22, 2079-2088, 1994.

[4] S. Gottesman, "Stealth regulation: biological circuitswith small RNA switches", *Genes & Development*, vol. 16, pp. 2829-2842, 2002.

[5] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database", *Nucleic Acids Research*, vol. 31, pp. 439-441, 2003.

[6] J. S. Mattick, "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms", *BioEssays*, vol. 25, pp. 930-939, 2003.

[7] V. Moulton, "Tracking down noncoding RNAs", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 2269-2270, 2005.

[8] G. Storz, "An expanding universe of noncoding RNAs", *Science*, vol. 296, pp. 1260-1263, 2002.

[9] Z. Weinberg and W. L. Ruzzo, "Faster genome annotation of non-coding RNA families without loss of accuracy", *Proc. 8th Ann. Int. Conf. on Computational Molecular Biology (RECOMB)*, pp. 243-251, 2004.

[10] B.-J. Yoon and P. P. Vaidyanathan, "Profile context-sensitive HMMs for probabilistic modeling of sequences with complex correlations", *Proc. 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, May 2006.

[11] B.-J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs - Unearthing the buried treasures in the genome", *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64-74, Jan. 2007.

[12] B.-J. Yoon and P. P. Vaidyanathan, "Fast search of sequences with complex symbol correlations using profile context-sensitive HMMs and pre-screening filters", *Proc. 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, Apr. 2007.

[13] B.-J. Yoon and P. P. Vaidyanathan, "Structural alignment of RNAs using profile-csHMMs and its application to RNA homology Search: Overview and new results", *IEEE Trans. Automatic Control (Joint Special Issue on Systems Biology with IEEE Trans. Circuits and Systems: Part-I)*, vol. 53, pp. 10-25, Jan. 2008.