



AN ALGORITHMIC FRAMEWORK FOR MAPPING FROM RNA SEQUENCE MOTIF TO GENE FUNCTIONS

Zhenghao Li

School of Computer Science and Technology, Tiangong University, China.

ABSTRACT

With the rapid development of current gene sequencing techniques such as RNA-seq (High Throughput Transcriptome sequencing), researchers have sequenced many RNA sequences, whose biological functions are unknown. The functions of many RNA are still controversial. There is also a certain relationship between different RNA sequences, and we can predict their functions by mining their relationships with each other. We design an algorithm framework to predict the biological function of RNA sequence patterns (Motif). Firstly, the patterns that frequently appear in RNA sequences are found through pattern recognition of RNA sequences. because some RNA share some similar RNA sequence patterns, their statistical significance is much higher than random similarity, and then combined with the Gene Ontology database containing functional annotations of all biological sequences, the biological functions of RNA sequences or RNA sequence fragments are predicted by means of statistical analysis.

KEYWORDS: RNA ; gene functions; algorithmic framework; statistical significance

1. INTRODUCTION:

Little is known about the biological functions performed by identified RNA sequence patterns, and some functional predictions about RNA patterns are controversial. Because of the huge amount of information and complex structure of RNA transcriptome, we can not only predict its biological function through living cell biological experiments, but also rely on bioinformatics to predict the function of RNA to a large extent. Therefore, first of all, it is necessary to design an efficient and reliable pattern recognition algorithm to mine the relationship between RNA and find out the patterns between RNA sequences, and then predict the biological function of RNA sequence patterns by combining various bioinformatics databases for functional enrichment and statistical analysis. Although the research on human genome has entered the post-genome era, pattern recognition and function prediction in RNA will still be the key research fields in bioinformatics. Among them, pattern recognition in RNA is the basis for further predicting the biological function of RNA, and related pattern recognition algorithms need to be improved; functional prediction methods for RNA sequence patterns also need to be supported by more accurate and reliable statistical models. Therefore, the research method proposed in this paper is of great significance in both theoretical research and practical application.

2. RELATED RESEARCH METHODS OF RNA FUNCTION PREDICTION:

Using support vector machine algorithm for gene function prediction: Guan et al. [1] used a large number of gene feature data as training data sets, and used three different SVM models to predict the gene function of mice, which achieved a good prediction effect. The first is a single-kernel SVM classifier, which constructs a SVM model for each GO functional class. The second is the SVM classifier which uses Bayesian network to correct the hierarchical relationship of GO function. The third method is the combination of single-kernel SVM classifier and Bayesian classifier, which divides the training feature data set into several, trains different SVM models respectively, and then uses naive Bayesian classifier to make a judgment. Finally, for each GO function, Guan selects the most appropriate model from the above three SVM models to predict. In addition to the above methods, there are many studies on the application of SVM model to predict the function of gene or protein in recent years, which shows its broad application prospect.

Using Bayesian network model to predict gene function: the MAGIC system designed and developed by Troyanskaya et al. [2] takes physical and genetic interaction data, gene expression related data and transcription factor binding site information data as training data, and then uses Bayesian network model to predict RNA gene function. Youngs et al. [3] also made use of the characteristics of Bayesian network with high speed, accuracy and strong anti-interference ability to optimize the gene function prediction method based on Gaussian random field algorithm proposed by Mostafavi et al. [4], which significantly improved the accuracy of prediction results.

Using K-nearest neighbor algorithm to predict gene function: Wong et al. [5] use the statistical features of biomedical literature to extract keywords as eigenvalues of K-NN classifier to predict GO molecular function and biological process. They believe that the combination of relevant literature text features with gene sequence and gene structure information as a data source can further improve the accuracy of gene function prediction. In addition, Bogdanov [6]

assumes that genes with similar functions have similar annotation patterns in their neighboring nodes. Based on this assumption, they also construct K-NN classifiers. Using the functions of k neighboring nodes that are most similar to the functional annotation environment of genes to be tested, they annotate the functions of the nodes to be tested, and good results are obtained.

3. AN ALGORITHMIC FRAMEWORK FOR RNA FUNCTION PREDICTION:

In this section, each process step of mapping from RNA sequence pattern to gene function algorithm proposed by the author is explained in detail one by one. Predicting the biological function of RNA sequence patterns can be divided into five steps:

1. RAP algorithm sequence alignment to find all RNA sequences similar to the target RNA pattern.
2. Extract the biological function data of similar RNA sequences from Gene Ontology
3. Statistical method of correlation between GO ID and motif
4. The function of RNA mode is predicted by the significance of GO.

3.1 Find all similar RNA sequences:

we will first use recursive pairwise sequence alignment algorithm (RAP algorithm) for sequence alignment, find out all human RNA sequences covering all RNA sequence patterns to be predicted, and give the corresponding similarity score and ranking. The specific process is as follows:

Algorithm Input alignment sequence: 1. The RNA sequence pattern (motif) generated by the sequence recognition pattern algorithm. 2. Gencode database contains all human RNA sequences file gencode.v31.transcripts.fa data set parameter

Algorithm output result: the alignment produces a file alignment_all.txt that contains all the sequence fragments in the human RNA library that are similar to the RNA pattern segment.txt to be predicted and their alignment results.

In summary, we find out all the RNA sequences covering the target motif through the RAP algorithm, and the optimal matching characters (Diff, Sim) of these sequences in the ID (ENSG), of the ensemble database.

3.2 Extraction of biological function data of similar RNA sequences:

In this section, we will find the biological basis of these similar RNA sequences (GO term, GO ID, chromosome number, multiple exon initiation and termination sites) through cross-links among Gencode database, Gene Ontology database, ensemble database and UniProtKB knowledge base in bioinformatics.

In the first step, after the result of sequence alignment is obtained, the corresponding ENSG ID, of the gene is found by comparing the corresponding gene sequence number in the result file (that is, the order of the gene in the human RNA group database file) and the ENSG ID is transformed into UniProtKB

through the UniProtKB ID mapping mapping tool.

The second step is to find the GO ID, GO term, Evidence code, and Aspect corresponding to each UniProtKB in the Gene Ontology database by using the UniProtKB obtained in the first step as the associated word.

In the third step, the ENSG ID corresponding to the gene obtained in the first step is associated with the gencode.v31.annotation.gtf file containing the relevant annotation information of all human chromosomes to find out the chromosome number of the corresponding gene exon, and give the accurate starting and ending site information on the chromosome segment.

The fourth step, data integration, defining the standard format of each column of data (Table 1)

Table 1: Data format standard for each column

Column name	Explanation
NUMBER	The sequence number of the gene is in the Genecode sequence file.
ENSG ID	ENSG ID corresponding to the gene
UniProtKB	UniProtKB corresponding to the gene
Sim	Compare the sequence with the same number of characters
Diff	Compare the number of different characters in the sequence
CHR	The chromosome number of the exon of the corresponding gene
Start	The starting site on the chromosome segment in which the exon of the gene is located
End	The termination site on the chromosome segment in which the gene exon is located.
GO ID	According to the GO term number found in Gene Ontology by UniProtKB
Aspect	Aspect indicates what kind of GO annotation category the gene belongs to
Evidence code	GO Evidence code

3.3 Statistical method of correlation between GO ID and motif:

Through the data integration and cleaning work at the end of the previous section, a standardized similar gene statistical data file is obtained. for each gene, that is, for each line in the file, we need to calculate a probability to show the significance of the gene sequence alignment. and sort the probability from small to large. The formula for calculating the correlation between genes and motif is as follows:

$$p = \sum_{k=x}^{x+y} C_{x+y}^k \cdot (p_0)^k \cdot (1 - p_0)^{x+y-k}$$

In the calculation, there are the following situations that need to be explained:

1. The alignment of each gene produces x and y, so each gene produces p_0 .
2. The computational complexity of the above formula is very high, so it is necessary to introduce some calculation skills. In the formula, $(p_0)^k$ and $(1 - p_0)^{x+y-k}$ will be repeated many times, saving the intermediate results can greatly reduce the computational complexity. For example, save it with one array $(p_0)^1, (p_0)^2, (p_0)^3, \dots, (p_0)^{10000}$ and another array $(1 - p_0)^1, (1 - p_0)^2, (1 - p_0)^3, \dots, (1 - p_0)^{10000}$.
3. p_0 is a parameter set, which reflects the probability of random alignment of RNA bases in the alignment. Because RNA is not a random sequence, the sequence correlation is very strong, which is set to 0.5 instead of 0.25.

After comparison, each motif is associated with several GO ID, and a GO ID is a biological marker or function. Through statistical significance analysis, we can reveal the possibility that a GO ID is related to a motif. the corresponding genes are sorted from small to large. The genes at the top of the list are more closely related to motif. After sequencing the genes, the significance assigned to the motif by each GO ID can be calculated. The statistical significance of a GO ID assigned to a motif is calculated using the following formula:

$$p = \sum_{k=x}^{x+y} C_{x+y}^k \cdot (p_0)^k \cdot (1 - p_0)^{x+y-k}$$

p_0 Is the background probability of each GO ID, k is the number of times that GO

ID appears in G, and x+y is the number of genes that G contains. A probability is calculated on each GO ID contained in the set G. If a probability already exists for a GO ID, a lower value is retained.

3.4 The function of RNA mode is predicted by GO significance:

In the previous section, we obtained the statistical significance of each GO ID. This section will first screen out the reliable GO, and then use the GO analysis tool to predict the biological function of the initial sequence pattern. The related work can be divided into the following two steps:

In the first step, the GO significance statistics files generated in the previous section are sorted according to the GO statistical significance probability from small to large, and the threshold is set to filter out the reliable data.

The second step is to query the relevant functional attribute description of each GO ID through the GO analysis tool: Quick GO, so as to predict the function of the initial RNA sequence pattern.

4. CONCLUSIONS:

By mining the relationship between RNA and finding out the patterns between RNA sequences, it is still necessary to combine various statistical calculations or biological experimental methods to predict the biological function of RNA sequence patterns. The functional prediction method proposed in this paper belongs to the category of general functional prediction of RNA sequence patterns. The limitation of this method is that it can not predict the differential functions of RNA sequence patterns in different biological environments. In order to overcome this limitation, future research can be specific to the functional prediction of RNA sequence patterns in a biological environment.

References:

- I. Guan Y, Myers CL, Hess DC, et al. Predicting gene function in a hierarchical context with an ensemble of classifiers[J]. Genome Biol, 2008, 9(Suppl 1): S3.
- II. Troyanskaya OG, Dolinski K, Owen AB, et al. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*) [J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(14): p. 8348-8353.
- III. Youngs N, Penfold-Brown D, Drew K, et al. Parametric Bayesian priors and better choice of negative examples improve protein function prediction[J]. Bioinformatics, 2013, 29(9): 1190-1198.
- IV. Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function[J]. Genome biology, 2008, 9(S1): S4.
- V. Wong A, Shatkay H. Protein Function Prediction using Text-based Features extracted from the Biomedical Literature: The CAFA Challenge [J]. Bmc Bioinformatics, 2013, 14.
- VI. Bogdanov P, Singh A K. Molecular function prediction using neighborhood features[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009, 7(2): 208-217.