A Novel Feature-Extraction For Classification of RNA Secondary Structure

Amir Hosein Kashefi [1*], Mostafa Noruzi Nashalji [2], Ali Kargarnejad[3]

1. Young researchers Club, South Tehran Branch, Islamic Azad University, Tehran, Iran
2. Young researchers Club, South Tehran Branch, Islamic Azad University, Tehran, Iran
3. Islamic Azad University, South Tehran branch, Tehran, Iran
amirdjtj@yahoo.com

Abstract: RNA has recently become the interest of scientists because of its catalytic properties, leading to an increased interest in obtaining structural information. This suggests that development of computational tools based on RNA secondary structure is essential for discovery of new non-coding RNAs and classification of their functional roles. In this paper, first we introduce a new method for feature-extraction from a RNA secondary structure sequence; next we use MLP neural networks for classification of six families from Rfam data base. Experiment results show that our represented method vs. previous works on classifying of RNA secondary structure has been improved and the structural complexity desirably has been decreased.
[Amir Hosein Kashefi, Mostafa Noruzi Nashalji, Ali Kargarnejad. **A Novel Feature-Extraction For Classification of RNA Secondary Structure**. J Am Sci. 2012;8(7):198-202]. (ISSN: 1545-1003). http://www.americanscience.org. 29

Keywords: feature extraction; RNA secondary structure; RNA classification; neural networks

1. INTRODUCTION

Nowadays, bioinformatics sciences is categorized in more than twenty different fields and because of widely improvement, some of them have been become an exclusive scientific field. Previous genetic researches lead to inspiration of genes, chromosomes, DNA, RNA, relations between inheritance and genes, their relations between protein's activities and so on. After these promotions, new problems have been defined, such that what inheritance and genetic relation exists between them and each attributes come from which gene, how genes determine the 3D proteins structure, what is the form of 3D proteins structures and the many other subjects.

Bioinformatics strictly are based on these concepts and design algorithms, software and databases and bioinformatics researches improvement. Today, researchers use computer and computer modeled data instead of using bioinformatics laboratories and examination tools. Also, many algorithms and software have been designed to analysis and scrutiny in these computer modeled bioinformatics data. For example, DNA, the genetic material of being, can be regarded as a sequence, composed of bases *a*, *t*, *c* and *g* and, saved in computer memory, then have access to all the genome (Malacinski, 2002). Ribonucleic acid (RNA) that is the subject of this study has a complicated structure with high molecule weight that have critical role to produce cellular proteins. In addition, RNA in some of the viruses plays the role of genetic carrier instead DNA. Also RNA performs a wide range of functions in the biological system. In particular, it is RNA that contains genetic information of virus such as HIV and therefore regulates the functions of such virus (Wuchty et al., 1999).

Many tasks which performed performed in this field have used support vector machine (SVM). In classifying the RNA secondary structure using SVM, there are two vital factors: first, an appropriate structure that indicate the representation of secondary structure and second, existance of a kernel function as a Reasonable Similarity criteria to measure similarity of two sequences. In 2002, Kin, Tsuda and Asai analyzed the classification of RNA secondary structure and a new kernel has been introduced that evaluate the similarity of two sequence of RNA secondary structure (Kin et al., 2002). In 2005, Karklin, Meraz and Holbrook First, defined an appropriate representation of RNA secondary structure by extending the RNA dual graph representation (Karklin et al., 2005; Gan et al., 2003), next they introduced a similarity measure between RNA secondary structures by using marginalized kernel to compare RNA molecules represented as labeled dual graphs (Kashima, 2003). In reference, Wang and Wu introduced a kernel function include both local and global information of RNA sequences and have used all these information for RNA classification and showed that their kernel function, work better than previous works on RNA classification (Wang and Wu, 2006).

In all the classification problems, first there's need to extract the features from all the patterns. Importance of feature extraction is in such a manner that whatever the achieved features are more accurate, problem classification is easier and with less complexity. In this paper, a new method for feature

extraction has been produced which improve the result of RNA secondary structure against previous works and the level of complexity has been decreased arbitrary.

AAAUGCGGCUAUCGCCCGUAUGCA

Figure 2. The first structure of an RNA.

The rest of the paper is organized as follows: Section 2 provides preliminary concepts. The new method for feature extraction and classify has been produced in Section 3. Section 4 includes the experimental results and finally concluding remarks are drawn in Section 5.

2. PRELIMINARY CONCEPTS

RNA sequences composed of bases *a* (adenine), *c* (cytosine), *g* (guanine), *u* (uracil), that have a tendency to fold back on themselves to form double-stranded structures. Watson-Crick base pairs between *a* and *u* and between *g* and *c* were formed, but sometimes less stable pairing are also possible, such as *g* with *u*. RNA structure has been presented in three different forms (Wang and Wu, 2006). First type of structure is a sequence of nucleotides which does not represent all properties of RNA sequence and just used to characterize nucleotides order in the RNA sequence. This sequence is presented with *A*, *U*, *C* and *G*, which are abbreviation of constituent RNA elements. Fig. 1 shows the RNA sequence.

The first type structure and these base pairs together with other morphological features such as bulge, hairpin, internal loops and multiple loops form structures are known as RNA secondary structures. The secondary structure of an RNA molecule determines shape of its structure and hence it has a real functional role (Wang et al., 2005). Fig. 2 shows this structure.

As it's shown in Fig. 2, nucleotide Hydrogen bonds are not adjacent to each other and they're known as pseudo knot free secondary structure. In order to create transplanted link, if *i* nucleotide clip links with j nucleotide, and also *k* and *l* nucleotides clip link to each other where $i<j$ and $k<l$, then combination of the four nucleotides would be one of these methods: $j>i>l>k$ or $k<i<j<l$.

Secondary structure with crossover links happened in special RNA that is called pseudoknot secondary structure. Fig. 3 shows these bonds. To create crossover bonds, if *i* nucleotide clip links with *j* nucleotide, and also *k* and *l* nucleotides clip link where $i<j$ and $k<l$, if the order of these four elements is such a manner that crossover each others, named crossover bonds and the relation $i<k<j<l$ is established between them (Malacinski, 2002).

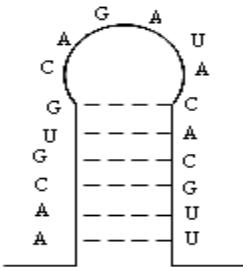

Figure 1. Pseudoknot free RNA secondary structure.

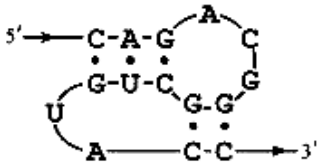

Figure 3. RNA secondary structure representation with pseudoknot.

Finally, the RNA third structure would be its space structure that composed from two or more RNA secondary structures. The operation of non-coding RNAs uniquely specified with 3D molecule structures. But creating helices and single strand loop areas are defined from the secondary structure (Zuker, 2000). A way to represent the secondary structure in linear form is parentheses representation. Each base pair in RNA sequence is represented with open and close parentheses and the non-pair bases represented with point. Fig. 4 shows this structure for a RNA sequence.

In order to mark these parentheses and determining position of bonds between the bases, some operation that known as the prediction the RNA molecule secondary structure would be necessary.

RNAs have different types which among them, the non-coding RNA is important. Non-coding RNAs do not have any role in codifiying the proteins, but have great impacts on cells (Eddy, 1999). They have many different play critical roles and exist in all the molecules throughout life (Storz, 2002). This suggests that development of computational tools based on RNA secondary structure is essential for discovery of new non-coding RNAs and classification of their functional roles.

3. THE CLASSIFIER STRUCTURE AND SIMULATION

In all of the previous works in classification of RNA secondary structure, a specified structure to represent the secondary structure has been used and in regards with this structure, the features of each RNA

sequence has been defined and then, a criteria to determine the similarity between the sequence has been produced.

TABLE II. SELECTED FAMILIES FROM RFAM

Family name	Family number
SRP_bact	1
U6	2
RRE	3
tmRNA	4
MIR159	5
MIR169_2	6

a. feature extraction

For feature extraction in Kin et al. method using context free grammars, extract the features and attached equal importance for base pairs and single bases in feature extraction (Kin et al., 2002). Also, in order to determine the adjacent between the nucleotides have used bi-gram model that is a model to determine the biological sequence. Wang and Wu (2006) in used bi-gram model and also, investigated the joint subsequence in each sequence for feature extraction. Although, base pairs locations were not determined exactly, but totally the achieved results is better than previous works.

In this study, we try to extract the features that include all the structure details. For this purpose first, RNA secondary structure sequence obtained with Vienna RNA secondary structure package (Wuchty and Fontana, 1999) then, by using each secondary structure according to Figure 4, we introduced our feature extraction. The set of our presented features include three parts:

- First set include the features in a feature vector that could present basic information about the number of paired bases of RNA sequence. So, start from the left sequence side and count the base pairs that are four pairs (G-C, C-G, A-U, U-A). Thus, the first set that is the first four dimensions of feature vector; store the information about base pairs in sequence regarding their orders.
- The second feature set in a feature vector, can present the information about the number of RNA unpaired bases type, regarding sequence beginning, sequence end and place between">>", "><", "<<" or "<>". Thus, the second set forms other 24 dimensions from feature vector, that achieved number is the result of six different state of mentioned multiplied by four unpaired bases.
- To compute the third feature set in a feature vector, such Kin et al., 2003, use the bi-gram model. This model can represent global structural information. In this paper, a bi-gram includes the combination of each of two bases set for RNA sequences, based on two single RNA to represent secondary structure of RNA sequences. Thus, third set, include other 128 dimensions from features vector, that achieved number include 8 different state ("..", "<.", ".>", ">.",".<", ">>", "<>", "<<") multiplied by 4×4 double bases. Of course, the state "<>" is computed in first set, has been removed from this section properly by tests.

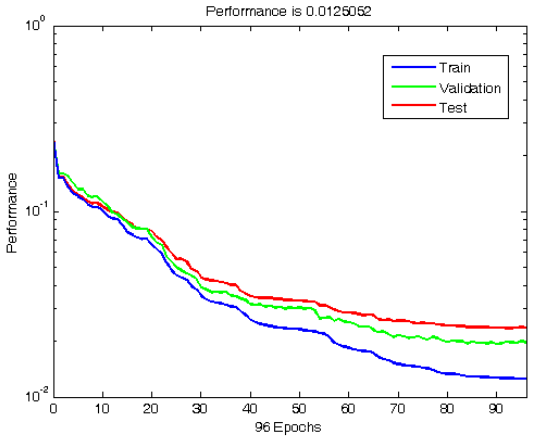

Figure 4. Classifier performance.

TABLE I. CLASSIFICATION PERCENTAGE IN TRAIN PHASE

percentage of Non-correct classification	Percentage of Correct classification	Family number
4	96	1
3.3	96.7	2
4.4	95.6	3
7.1	92.9	4
35.7	64.3	5
100	0	6
7.4	92.6	total

Thus, the features vector includes 128+24+4=156 dimensions. Therefore, for classification of secondary RNA structure, a preprocessing to extract the features has been accomplished and 156 features from the data have been extracted. These extracted features can be regarded as the input of the classifier.

b. neural networks

A neural network consists of a large number of simple processors called neurons. The input vector –in this paper called neural input- is containing r-

TABLE III. CLASSIFICATION PERCENTAGE IN TEST PHASE

percentage of Non-correct classification	Percentage of Correct classification	Family number
10.7	89.3	1
9.1	90.9	2
0	100	3
6.3	93.8	4
45.8	54.2	5
100	0	6

inputs $X=[x_1, x_2, \ldots, x_r]$. These inputs are from external source or can come from other units. Neural inputs multiplied by the weight matrix *W* to form *WX*. Each neuron has a constant input of '1' shown by x_0, x_0 is multiplied by a bias (w_0) to form w_0x_0. In each neuron WX and w_0x_0 send to the summer. The summer output, n, is $WX+w_0x_0$ ($n_i=\sum_{q=0}^{r} w_{iq}x_q \quad i=1,\ldots,s$ neurons of single layer Perceptron). The output of neuron i is $a_i=f(n_i)$, where $f(.)$ is an activation function of the neurons (Gupta et al., 2003; Hagan et al., 1996). The output of the single layer Perceptron is given as $a=[a_1, a_2, \ldots, a_p]^T$.

Network with several layers has simply cascade three Perceptron networks, each layer has its own weight matrix W, its bias vector w_0 and as output vector a. different layers can have different numbers of neurons. A layer whose output is the network output is called an output layer; the other layers are called hidden layers (Hagan et al., 1996). The network has an output layer (layer 3) and two hidden layers (layer 1 and 2). Each neuron in the first hidden layer has an input from every neural input with one additional input of 1 -bias-; each neuron in the second hidden layer has an input from every neuron in the first layer additional input like first layer. In the output layer, there is single neuron called output neuron. The output neuron has an input from every neuron in the second hidden layer and one additional input like first and second layer. In fault detection application, neural networks used measured and manipulated variables as inputs, while the output represents categories –this neural network called neural classifier- in this paper contains faulty and normal operation. Usually the output is dummy variable ('1' or '0') where '1' indicates an in-class member while '0' indicates a non-class member. Here, dummy variables as output are '1' or '-1' which the margin between the in-class members shown normal operation and non-class members shown faulty operation which is bigger than previous dummy variable.

g g g c a u c c c a

< < <. . . >>> .

Figure 5. RNA secondary structure with parentheses representation

2. EXPERIMENTAL RESULTS

In order to evaluate the performance of proposed classifier, the six non-coding RNA sequence families from Rfam data base [15, 16] were choose that are the newest Rfam database release in 2008. These families have been shown in table 1.

The used classifier is a neural network with two active layers (hidden and output layer). 70 % of data used to train the network using the Liebenberg Marguerite learning algorithm (LM), 10% of data is for validation and 20% is for testing the classifier. The performance figure of neural network is depicted in fig. 5 to raise your understanding.

Table.2 shows the result of training neural network in worst case in 30 run of neural network, and table.3 shows the result of testing the classifier.

4. CONCLUSION

As see in table.3, feature extraction of RNA secondary structure for training neural network, using neural network which are useful for classifying of RNA secondary structure. The result of first four families against the classification result of Wang (2006) that is the best classifier up to now, show that achieved result what terms of efficiency and performance and what terms of complexity has been improved because Wang used from a complex and non-deterministic software named discover, but we used a simple feature extraction and show better performance than Wang method.

REFERENCES

[1] G. M. Malacinski, Essentials of Molecular Biology, Fourth Edition, 2002.

[2] S.Wuchty, W.Fontana, I.Hofacker and P.Schuster Complete Suboptimal Folding of RNA and the Stability of Secondary Structures, Biopolymers, 49, 145–165. 1999.

[3] T. Kin, K. Tsuda and K. Asai. Marginalized Kernels for RNA Sequence Data Analysis, Proc. Genome Informatics Workshop, 2002.

[4] Y. Karklin, R. F. Meraz, S. R. Holbrook, Classification of non-coding RNA using graph representations of secondary structure, Proceedings of the Pacific Symposium on Biocomputing 10:4-15, 2005.

[5] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA

design. Nucleic Acids Res, 31(11):2926–43, 2003.

[6] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In International Conference on Machine Learning, volume 20, pages 321–328. AAAI Press, 2003.

[7] Jason T.L. Wang, Xiaoming Wu .Kernel design for RNA classification using Support Vector Machines International Journal of Data Mining and Bioinformatics (IJDMB), Vol. 1, No. 1, 2006.

[8] Liu., J., Wang, J.T., Hu, J., and Tian, B. A method for aligning RNA secondary structures and its application to RNA motif detection. BMC Bioinformatics 2005, 6:89.

[9] M. Zuker. Calculating nucleic acid secondary structure. Curr Opin Struct Biol, 10(3):303–10, 2000.

[10] S. R. Eddy. Noncoding RNA genes. Curr Opin Genet Dev, 9(6):695–9, 1999.

[11] G. Storz. An expanding universe of noncoding RNAs. Science, 296(5571):1260–3, 2002.

[12] S.Wuchty, W.Fontana, I.Hofacker and P.Schuster Complete Suboptimal Folding of RNA and the Stability of Secondary Structures, Biopolymers, 49, 145–165. 1999.

[13] M. M. Gupta, L. Jin, and N. Homma, Static and dynamic neural networks: from fundamentals to advanced theory. Wiley-IEEE, 2003.

[14] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston: PWS Publishing, 1996.

[15] Howard Hughes Medical Institute, The Rfam database of RNA alignments and CMs, http://rfam.janelia.org/index.html, December 2008.

[16] P. P. Gardner, J. Daub, John G. Tate, E. P. Nawrocki,D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn,S. Griffiths-Jones, S. R. Eddy. and A. Bateman. Rfam: updates to the RNA families database. Nucleic Acids ResearchVol. 37, Database issue, 2009.

5/24/2012