A Study of RNA Structure Automatic Classification by Using Neural Network

Chuen-Der Huang

Abstract

Structure Classification Of RNA (SCOR) is an evolving resource that will continue growing as more RNA structures become available. The RNA structure classification task must be done carefully with experimental process to complete the final result. It is not only a hard but time-consuming job to complete this work as other works in bioinformatics. Recognizing the importance of the difficulty of the subject and propose the machine learning method to solve this problem, the study has been made.

The RNA sequences data we used come from the SCOR database and has been grouped according to their functions by using machine learning methods with a simple neural network. The radial based function neural networks (RBFN) are chosen here to accomplish the task. With the method we applied, the prediction results can reach to higher than 83 %. The SCOR database comes from the open web site of SCOR.

Key words: SCOR, RNA, neural network, motif.

Chuen-Der Huang, Associate Professor, Department of Electrical Engineering ,HIT

利用類神經網路作RNA結構自動分類之研究

黃淳德

摘 要

RNA結構分類資料庫（SCOR）在RNA的結構分類上扮演一個很重要的角色，這個資料庫隨著其他實驗的完成，依然持續成長中。RNA的結構分析，如同生物資訊的其他議題般，是一個辛苦且繁複的工作，尤其是實驗室的作業，更是費時。本研究將利用機器學習的方式，對這個問題提供輔助的結果以縮短時程。

研究中所用的資料均來自SCOR這個資料庫，利用編碼及類神經網路的學習方式對所選的資料作學習以建立一個適合這個問題的神經網路，我們採用徑向基底函數的神經網路，經過實驗預測正確率可達83%。

關鍵詞：SCOR, RNA, neural network, motif。

黃淳德：修平技術學院電機工程系副教授

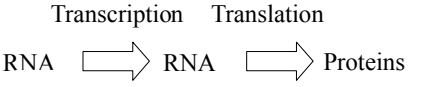
I. Introduction

Both in the Protein Data Bank (PDB) and in the Nucleic Acid Database (NDB), the numbers of RNA structures, whose coordinates are available, are substantial and rapidly growing. In the past, the great majority of the structures are made up of only one helical stack. Recently, such as the hammerhead ribozyme, the determination of structures contains two or more helical stacks. The structures provide a large amount of information about RNA structural motifs. These motifs have also been studied extensively [1,3,16]. In order to organize this information and make it available to the non-specialist, to discover new features of RNA structure and relationships to sequence and function, and to enumerate and classify substructures for model building and RNA engineering, Klosterman and his colleagues have developed a database for the structural classification of RNA called SCOR.

The establishers of SCOR examined of 259 PDB entries, cataloged and classified all of the internal and external loops in a comprehensive collection of RNA structures contained in the PDB and NDB [8,9,10,14, 15,16].

II. SCOR,StructureClassification Of RNA

SCOR, Structure Classification Of RNA, is established in 2002. It is a well known database of RNA. RNA (ribonucleic acid) is the important materials which are the templates to form the proteins in life, the different functions of RNAs such as message RNA, ribosomal RNA, transfer RNA and RNA polymerases are join together to synthesize the proteins [3]. In normal cells the messages of heredity can be summarized as the following.



The data we used come from this data bank at the web site http://scor.lbl.gov. SCOR is an evolving resource that will continue to grow as more RNA structures become available. It is divided into several classes, naturally Occurring RNA, evolved RNA, synthetic RNA, and the RNA that structure without classified.

In this study, the structural functions of data have been adopted and the total number and each number of groups are illustrated in the Table 1. Examining at Table 1, it can be easy to find that the classes of Ribozymes and SnRNA consist just a few number of RNA,

three and five in the table; therefore, in our study we do not consider these two kinds of RNA.

Table 1. RNA number of data in SCOR.

Naturally Occurring RNA	Transfer RNA	33
	Ribosomal RNA	29
	Ribozymes	18
	SnRNA	3
	SRPRNA	5
	Genetic Control	31
	Viral Packaging	17
Evolved RNA		20
Synthetic RNA		28
Structure Without Classified		75
Total Number		259

The database structure of RNA is somewhat in common with the database structure of protein, but indeed there are some differences between proteins and RNA in fundamental, bio-structures and functions etc.. The properties of differences will affect the design of database structure in RNA. One of the classified differences of these two databases in structure can be described as follow. The protein structure is considered to be modular at domain level while RNA is considered to be modular at the motif level. In addition to this difference, the sequence of RNA within helical regions can easily co-vary without affecting structure. Nonetheless, the proteins can not appear this phenomenon; therefore, the features we choice in RNA database is less than in protein database.

III. Database and Features

(1)Training dataset

According to their function, the RNAs are classified into ten classes in the SCOR database. Table 1 is the contents of the RNA in SCOR. As we mentioned above, the database of SCOR is growing in number, but here we take the first version of them to be our data to prove our method.

From Table 1, it can be found easily that SnRNA and SRPRNA have only few numbers in the database and then we do not consider these two kinds of classes under our machine learning experiments. Because it is known that too few of data can not achieve satisfied results in machine learning. In our experiments, machine learning method, we separated each class into two parts, one is used for training and the rest part is used as testing. Besides, for a few number of data, the cross validation method is often applied to make sure the effect the experiments. With this consideration, we divided the obtained data into seven groups; therefore the ratio of training data and testing data is six to one.

The jack knife test has been done in our experiments. By this way, we can obtain seven sets of training data for training to make the results of our experiments more reliable. Meanwhile, each one of the training data set contains 331 RNA in amount. All the data will be fed into a specified supervised network in order to be classified into 8 groups by their functions.

(2)Testing dataset

Since we have divided all of the data into seven groups, the testing data we used to test the network is one seventh of each group. We divided the data into seven groups from the original database by random. By this method, we have seven sets of testing data which are corresponding to the training data; because the numbers of data are limited therefore the jack knife testing has to do here to prove the accuracy of prediction. After grouped, each one of the testing data set contains 54 RNA data approach in number. The testing data will be fed into the trained network in order to test the accuracy of the prediction.

IV. Feature Vector extraction

It is known that in machine learning the features vectors extraction is a very important task; different features vectors extraction may lead to different results, better or worse. The structures of RNA are based on the sequence of four compositions, four bases. The bases of RNA and their symbols are symbolized as A (adenine), U (uracil), C (cytosine) and G (guanine). In our study, we use three kinds of descriptors denoted by K, T and D to represent our features. The represent symbol K is the percentage of bases; T is the percentage frequency with which symbol A is followed by symbol B or symbol B is followed by symbol A. The third descriptor is symbolized by D which represents the distribution of the property is described by 5 chain length (in percent), that is the first, 25%, 50%, 75% and 100%. In our case there are five bases in the item of K, that is A, U, C, G and X; and the item contains ten possible transpositions.

The character X denotes that unknown or uncertain composition of the sequence [4,5,6,7,12,13].

V. Kernel of Machine Learning Algorithm

While mention about the machine learning algorithm, neural network (NN) will be discussed without doubt. Neural network has been developed for many years and was

used well and widely in many fields. Recently, the techniques and concepts are introduced into the field of bioinformatics rapidly. One of the advantages of NN is that the NN can do nonlinear, multi-classes and high performance work in machine learning under different kind of structures. Since the middle of twenty century started from Widrow-Hoff, there are many kinds of neural networks have been proposed both in structures and algorithms. The radial basis function network (RBFN) is a kind of hybrid network of NN which combined self-organize-map (SOM) and back-propagation. In RBFN, the hidden layer nodes could show the coordinate of training sample clusters. This network, suggested by Moody and Darken in 1989, is very suitable to be used as classifier. Considered about the characteristics of RBFN, we chose the RBFN to be the network in our experiments.

The learning algorithm of RBFN is also a kind of hybrid network. The learning of RBFN is two phases, during the first phase, called as the unsupervised learning phase. There are three steps will be made in this phase: (1) determinate the Eulier distance (d_k), (2) find out the winner node and (3) the winner takes all. Only the weight of winner node should be modified and it would be adjusted. Eulier distance (d_k) is the distance between hidden layers unit and input vectors (X_i). It can be described as the sentence.

$$d_k=\sqrt{\sum_i (X_i - W_{ik})^2} \quad (1)$$

The weights W_{ik} are adjusted by the equation.

$$\Delta \mathbf{W}_{ik}=\eta_1 * (\mathbf{X}_i - \mathbf{W}_{ik}) \quad (2)$$

Where η_1 is the unsupervised learning rate.

In the second phase of learning, the supervised learning phase, the average weight rule is taken. At first, count out the distance between hidden layer and training sample. Secondly, calculate the output value of hidden layer by the equation:

$$H_k = \exp(-\frac{d_k^2}{2s^2}) \quad (3)$$

Next, calculate the output value of output layer by the equation:

$$Y_j = \frac{\sum_k W_{kj} H_k}{\sum_k H_k} \quad (4)$$

Final step, adjust the weights between output layer and hidden layer by the equation shown:

$$\Delta W_{kj} = h_2 d_j \frac{H_k}{\sum_k H_k} \quad (5)$$

where σ is called the smoothing parameter

$$d_j = (T_j - Y_j) f'(net_j)$$

and h_2 is the supervised learning rate.

Table 2. The RNA and their ID code in SCOR

tRNA	rRNA	Ribozymes	SnRNA	SRPRNA	Genetic Control	Viral Packaging	Evolved RNA	Synthetic RNA	Structure Without Classified	
1yfg	364d	1ajf	1urn	1cq5	1biv	17ra	1koc	157d	100d	1nao
2tra	1elh	1ajl	1a9n	1duh	1mnb	1a1t	1am0	1f5g	104d	1nxr
1ehz	354d	1c0o	1u2a	28sr	1esh	1bn0	1raw	1mis	124d	1ofx
1evv	1a4d	1eor		1dul	1slo	1esy	1kod	1osu	161d	1oka
1tn2	1d6k	1gid		1d4r	2bj2	1aq3	1fmn	1qes	168d	1pb
1tra	1dfu	1grz			1akx	1d0t	1ull	1qet	169d	1qc0
4tra	353d	1guc			1anr	1d0u	484d	1yfv	176d	1qcu
4tna	361d	1hlx			1arj	1dzs	1f1t	280d	1a34	1qln
6tna	1a3m	1tlr			1qd3	1zdh	1nem	333d	1ac3	1r3x
1c0a	1byj	1hmh			397d	1zdi	1ec6	402d	1al5	1rau
1efw	1pbr	1mme			1bvj	1zdj	5msf	405d	1apg	1rmv
1euy	1bgz	299d			1csl	1zdk	6msf	406d	1av6	1rna
1exd	1rng	359d			1duq	1rht	7msf	409d	1b2m	1rnk
1gtr	1zif	379d			1ebq	1tfn	1exy	420d	1b7f	1rxa
1gts	1zig	488d			1ebr	1bau	1eht	433d	1bmv	1sdr
1qrs	1zih	1ato			1ebs	1f5u	1tob	438d	1byx	1vtm
1qtq	1wts	1cx0			1etf	462d	2tob	332d	1c2q	216d
1qf6	1dk1	1b36			1kis		2ldz	485d	1c9s	219d
1qu2	1g1x				1aqo		429d	247d	1cgm	222d
1asy	1uuu				1a4t		1hvu	1br3	1cvj	246d
1asz	1afx				1qfq			1egk	1cwp	259d
1eiy	1mms				1cn8			1c4l	1cx5	2a8v
1ser	1qa6				315d			165d	1d87	2bbv
1b23	1ffk				1ekz			1qbp	1d96	2tmv
1ttt	1jj2				2a9l			205d	1d9d	2tpk
2fmt	1vop				1ei2			255d	1dno	310d
1ikd	483d				1qc8			413d	1dhh	377d
434d	1scl				1a60			283d	1di2	398d
464d	430d				3php				1drr	404d
1bz2					1aud				1dxn	410d
1bz3					1dz5				1efo	419d
1bzt									1evp	421d
1bzu									1f27	435d
1kos									1fix	437d
									1gsg	468d
									1gtc	479d
									1kaj	8drh
									1kpd	

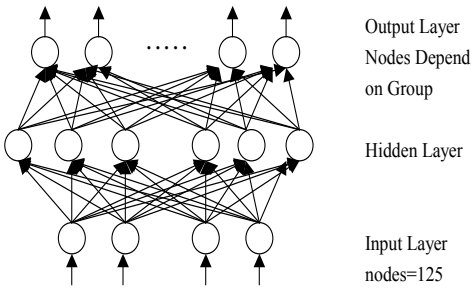


Figure 1. Illustration of the basic RBFN structure.

In figure 1 the basic concept of RBFN is illustrated. The RBFN has only one hidden layer but the nodes of hidden layer may increase with the operation of cost function. In these experiments, the RBF network is chosen [11].

VI. Experiments and Results

In our experiments, RBF network is chosen to do the classification. First, we separated the data into two parts with random, one is used as training data and the other is used for testing. The random picked testing data to form the set is according to the rate of seven to one. By using jack-knife testing we can obtain seven results. Table 3 shows the results we made. In Table 3, it can be found that we do not train the network to one hundred percent of learning. In fact, we have done it carefully to avoid the over learning problem. Every experimental result we made, the accuracy is larger than 72% even up to 94%. We obtained the result by arithmetic average.

The accuracy measurement in our method is very simply and clearly. The accuracy can be described as follow. If the amount of testing proteins is n_i, it should be classed into the F_i^{th} fold but in fact the outputs of the classifiers only class the amount of c_i into the F_i^{th} folds, the accuracy rate is simply to be calculated as c_i / n_i for the F_i^{th} fold and so on. In addition, besides calculate the individual accuracy rate, the total accuracy rate can be calculated easy too. The following relationship can be easy to understand.

$$N = n_1 + n_2 + ... + n_7 \quad (6)$$

$$C = c_1 + c_2 + ... + c_7 \quad (7)$$

$$Q = \frac{C}{N} \quad (8)$$

Where N is the amount of testing.
C is the amount of corrective prediction.
Q is the accuracy of prediction.

In this method we made, the measurement of results are very clearly; the so called true positive and false positive will not occur in our method, for

the networks we used are multi-output network each one has its position and one only [1,2].

Table 3. The results of prediction by using the RBFN in the experiments.

Training	NO. of Correct	Accuracy (%)	Testing	NO. of Correct	Accuracy (%)
Training_Subset_1	331	100	Testing_Subset_1	45	83.3
Training_Subset_2	330	99.7	Testing_Subset_2	51	94.4
Training_Subset_3	328	99.1	Testing_Subset_3	47	87.0
Training_Subset_4	330	99.7	Testing_Subset_4	46	85.2
Training_Subset_5	331	100	Testing_Subset_5	45	83.3
Training_Subset_6	330	99.7	Testing_Subset_6	36	72.2
Training_Subset_7	330	99.7	Testing_Subset_7	43	79.6
Average		99.7	Average		83.6

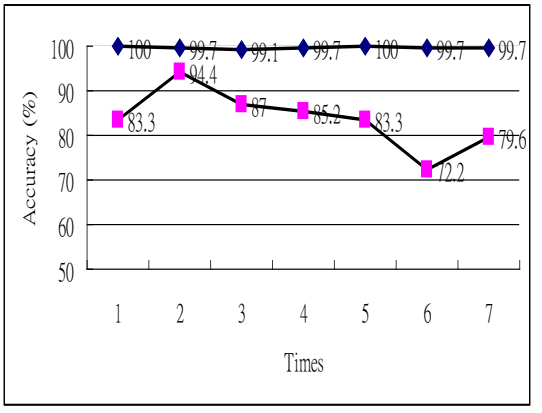


Figure 2.The results of prediction of each time by using RBFN.2

VII. Conclusions

In this paper, we refer and studied several paper of RNA and based on the paper to propose the machine learning algorithm to SCOR. We survey the results we have, NN based; the results show that the accuracy of prediction is satisfied, 83.6%. With the results we can recognize that the NN could be used for automatic classify of SCOR.

VIII. References

[1] Baldi, P. and Brunak, S. (1998) Bioinformatics: the Machine Learning Approach. MIT Press, Cambridge, MA.

[2] Baldi, P *et al*., (2000) Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics* ,16, issue 5, 412-424.

[3] Butcher, S. E., Dieckmann, T. and Feigon, J. (1997) Solution structure of a GAAA tetraloop receptor RNA. *EMBO* J., 16, 7490 – 7499.

[4] Chou, K.C. and Zhang, C.T. (1995) Prediction of protein structural classes. *Mol .Bio*., 30,275-349.

[5] Dubchak Inna, and Sung-Hou Kim, et al. (1999) Recognition of a Protein Fold in the Contex of the SCOP Classification,PROTEINS: Structure, *Function, and Genetics,* vol. 35, 401-407.

[6] Hobohm, U. and Sander, C. (1994)

Enlarged representative set of *p*rotein structures. *Protein Science* 3, 522-524.

[7] Jaakkola, T. Diekhans, M. and Haussler, D. (2000) A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology,* 7 (1,2): 95—114.

[8] Jucker, F. M. and Pardi, A. (1995) Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. *Biochemistry,*34,14416 – 14427.

[9] Klosterman P.S., Hendrix D.K., Tamura M., Holbrook S R, and Brenner S E. (2004)
Three-Dimensional Motifs from the SCOR: Structural Classification of RNA Database - Extruded Strands, Base Triples, Tetraloops, and U-turn. *Nucleic Acids Res.* 32. 2342-2352.

[10] Klosterman, P. S., Tamura, M. , Holbrook, S. R. and Brenner, S. E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.,* 30, 392 – 394.

[11] Lin C. T. and Lee C. S. G. (1996) *Neural fuzzy systems: a neural-fuzzy synergism to intelligent systems,* Englewood Cliffs, Prentice-Hall., N.J.

[12] Loredana Lo Conte, Steven E. Brenner, *et al.,* (2002) SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Research,* vol. 30, No. 1, 264-267.

[13] Lo Conte et al., (2000) SCOP: a structural classification of proteins database. *Nucl Acid Res,* 28, 257-259.

[14] Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H., Seeman, N. C. and Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science,* 185, 435 – 440.

[15] Tamura, M., Hendrix, D. K., Klosterman, P. S. , Schimmelman, N. R., Brenner, S. E. and Holbrook, S. R. (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.,* 32, 182 – 184.

[16] Tamura M. and Holbrook S.R. (2002) Sequence and Structural Conservation in RNA Ribose Zipper. *J. Mol . Biol.* 320. 455-474.