Chiang Mai J. Sci. 2012; 39(1) : 1-7
http://it.science.cmu.ac.th/ejournal/
Contributed Paper

RNA Family Classification Using the Conditional Random Fields Model

Sitthichoke Subpaiboonkit [a, b], Chinae Thammarongtham [c], and Jeerayut Chaijaruwanich*[a, b, d]

[a] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand.
[b] Bioinformatics Research Laboratory, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand.
[c] Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology, Bangkok 10150, Thailand.
[d] Biomedical Engineering Center, Faculty of Engineer, Chiang Mai University, Chiang Mai 50200, Thailand.
*Author for correspondence; e-mail: jeerayut.c@cmu.ac.th

Received: 21 June 2011
Accepted: 6 October 2011

ABSTRACT

RNA family classification is one of the necesary tasks needed to characterize sequenced genomes. RNA families are defined by member sequences which perform the same function in different species. Such functions have a strong relationship with RNA secondary structures but not the primary sequence. Thus RNA sequences alone are not sufficient to classify RNA families. Here, we focus on computational RNA family classification by exploring primary sequences with RNA secondary structures as the selected feature to classify the RNA family using the method of conditional random fields (CRFs). This model treats RNA classification as a sequence labeling problem. Our CRFs models can classify the RNA families of the test RNA data sets with optimal F-score prediction between 98.77% - 99.32% for different RNA families.

Keywords: RNA family classification; Conditional random fields; bioinformatics; machine learning.

1. INTRODUCTION

Ribonucleic acids (RNAs) including non-coding RNAs (ncRNAs) are molecules that can play important roles in many gene regulatory mechanisms [1, 2]. These sequences have secondary structures which are conserved according to complementary base pair interactions, such as canonical base pairs (Watson-Crick complementary base pairs, with also an occasional G -U pair).

RNA families including non-coding RNA families are classified as groups of related RNA sequences where each RNA family is defined by member sequences which perform the same function in different species [3]. For such sequences with conserved functions, both RNAs [1]

and non-coding RNAs [4] show conserved secondary structures whereas the primary sequences can be highly mutated. Therefore use of the RNA sequences alone is not sufficient to determine RNA family classification.

In this paper, we propose that RNA sequences and their secondary structures can be used to classify RNA families. To examine this hypothesis, we define RNA family classification as a sequence labeling problem where RNA sequence residues and their secondary structures are labeled by proper RNA family labels. One standard method for performing such sequence labeling tasks is the conditional random fields method (CRFs). This method is a supervised machine-learning technique using an undirected graph to build probabilistic models which solves the sequence labeling task with suitable feature extraction based on the conditional approach [5]. CRFs use a finite state model with un-normalized transition probabilities. Unlike some other weighted finite-state approaches such as the convolutional neural network [6], CRFs assign a well-defined probability distribution over possible labels, trained by maximum likelihood (MLE) or maximum a posteriori (MAP) estimation. Since its loss function is convex, this guarantees convergence to the global optimum [7].

2. MATERIALS AND METHODS

2.1 Features Selection for RNA Family Classification Based on Conditional Random Fields

Let X be a random variable of RNA sequences and their secondary structural feature labels, and Y be the label random variable of the corresponding RNA family. In the discriminative framework, a conditional probability $p(Y=y \mid X=x)$ of RNA family label sequences $y = y_1,...,y_t$ are estimated from the RNA sequences and their secondary structural features $x = x_1,...,x_t$. Here, the RNA secondary structural features that are dot-parentheses format consists of 1) stem complementary base pairs denoted by "(" and ")", 2) loop nucleotides denoted by "*" and 3) non-loop unpaired nucleotides denoted by ".". The RNA family labels are the abbreviation of the RNA family chosen in the experiment data. The example of RNA sequences and their secondary structure labeling is shown in Figure 1.

A	**.**	**fiv**
A	**.**	**fiv**
U	**(**	**fiv**
G	**(**	**fiv**
G	**(**	**fiv**
G	*****	**fiv**
A	*****	**fiv**
G	*****	**fiv**
A	*****	**fiv**
C	**)**	**fiv**
C	**)**	**fiv**
G	**)**	**fiv**
A	**.**	**fiv**
A	**.**	**fiv**

Figure 1. Example of RNA sequences and their secondary structure labels. The first column shows the RNA nucleotides, the second column the RNA secondary structures, and the third column the RNA family label.

The features we choose to characterize the family members, in this case the RNA secondary structure, train the CRFs to recognize the particular characteristics of the secondary structure for each RNA family.

2.2 Predicting RNA Family Based on Conditional Random Fields

In this study, we consider CRFs only for first order dependencies of state transitions. Referring to Lafferty, McCallum, and Pereira (2001) [7], we define the conditional probability of the RNA family label y given the RNA sequences and their secondary structural features x as follows

$$p_\theta = p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T} F(y,x,t)\right) \quad (1)$$

where

$$Z(x) = \sum_{y'} \exp\left(\sum_{t=1}^{T} F(y',x,t)\right) \quad (2)$$

θ is the model parameter, $Z(x)$ is a normalized factor which sums all cases of label y' and $F(y,x,t)$ is the sum of feature frequency at position t.

We define the feature function of the entire RNA sequences and their secondary structural features x and the corresponding RNA family label y at position t as

$$F(y,x,t) = \sum_{l',l,j} \lambda^j_{l',l} f^j_{l',l}(y_{t-1}, y_t, x) + \sum_{l,j} \beta^j_l g^j_l(y_t, x) \quad (3)$$

where $f^j_{l',l}(y_{t-1}, y_t, x)$ is the transition feature function of RNA sequences and their secondary structural features x and RNA family label y at position t-1 and t equal to $l'\theta$ and l respectively in the whole states of RNA family labels; $g^j_l(y_t, x)$ is the state feature function of the RNA family label at position t and the RNA sequences and their secondary structural features x. $\lambda^j_{l',l}$, β^j_l and i.e. $\theta = (\lambda,\beta)$, are feature weights associated respectively with $f^j_{l',l}$ and g^j_l estimated from the training data.

We define

$$z_j(x, t) = \quad (4)$$

$$\begin{cases} 1 \text{ if } x \text{ at position } t-k \text{ to } t+k \text{ is } equal\ to\ (n,s) \\ 0 \text{ otherwise} \end{cases}$$

$$f^j_{l',l}(y_{t-1}, y_t, x) = \quad (5)$$

$$\begin{cases} z_j(x,t) & \text{if } y_{t-1} = l' \text{ and } y_t = l \\ 0 & \text{otherwise} \end{cases}$$

$$g^j_l(y_t, x) = \begin{cases} z_j(x,t) & \text{if } y_t = l \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Our feature selection is generated from the possible relationships between nucleotide bases and their known RNA secondary structure. From equation (4), $z_j(x, t)$ represents whether x at position t - k to t + k consisting of the nucleotides subsequence n and the RNA secondary structure feature s. We consider a reasonable local information pattern of k = ± 4 nucleotides surround the considering position t. In equation (5), and (6), l and l' are RNA family labels.

For training, the feature weight (θ) can be obtained from maximizing the log-likelihood $\mathcal{L}$ of the given training data set $\{x^{(j)}, y^{(j)}\}_{j=1\ldots N}$

$$\mathcal{L} = \sum_{j=1}^{N} \log\left(p_\theta(y^{(j)}|x^{(j)})\right) - \sum_k \frac{\theta_k^2}{2\sigma^2} \quad (7)$$

where $\sum_k \frac{\theta_k^2}{2\sigma^2}$ is a Gaussian prior computed from the feature weights θ_k and variance σ^2 that can help to reduce problems with overfitting and sparsity in the training data [8].

Parameter (θ_k) estimation of CRFs requires an iterative method to solve the convex optimization of θ_k by using quasi-Newton methods or L-BFGS [9] that are the most efficient [10].

$$\frac{\delta \mathcal{L}}{\delta \theta_k} = \left[\sum_{j=1}^{N} C_k(y^{(j)}, x^{(j)})\right] - \left[\sum_{j=1}^{N}\sum_{y} p_\theta(y^{(j)}|x^{(j)}) C_k(y, x^{(j)})\right] - \frac{\theta_k}{\sigma^2} \quad (8)$$

where $C_k(y^{(j)}, x^{(j)})$ is the sum of features f_k given y and x. The first term represents empirical data summation and the second term represents model expected value from features f_k. The last term is the first-derivative of the Gaussian prior [11].

CRFs prediction is performed by finding the most probable label sequence y^* from the training model given the observation or RNA sequences and the secondary structural features x from the testing data.

$$y^* = argmax_y p_\theta(y|x) = argmax_y \exp\left(\sum_{t=1}^{T} F(y, x, t)\right) \quad (9)$$

To generate the inferences in CRFs, a dynamic programming algorithm is applied using the Viterbi algorithm [7] to find the most probable label y^* for the observation or RNA sequences and their secondary structural features x with highest probability value.

2.3 Implementation of Conditional Random Fields

The configurations of our proposed CRFs are shown in Figure 2. We consider a range of four nucleotides forward and backward from the indicated position t. These relationships are also used for training and testing data formats.

The dataset used for training and testing in our experiment which is summarized in Table 1, was extracted from RNA STRAND which is an RNA secondary structure and statistical analysis

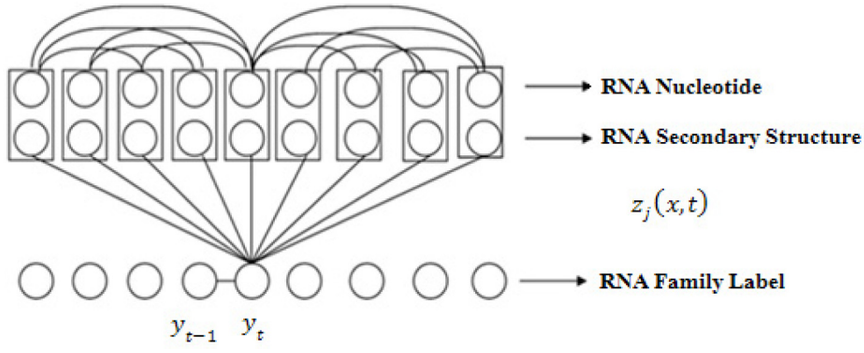

Figure 2. Graphical model of our CRFs model.

Table 1. Training and testing dataset in our experiment.

RNA families	Number of sequences
Signal Recognition Particle RNA or SRP	394
5S Ribosomal RNA or 5S RNA	161
Cis-regulatory element	41
Hammerhead ribozyme	146

database [12]. This database contains biologically feasible structure information.

The RNA families used in our experiment are Signal Recognition Particle RNA or SRP, 5S Ribosomal RNA or 5S RNA, *Cis*-regulatory element and Hammerhead ribozyme, and their represented abbreviation for using in RNA families are "srp", "fiv", "cis" and "ham" respectively. According to RNA secondary structure for each RNA family that we choose, SRP has the most complicated structure of our testing RNA families. 5S RNA has a more complicated structure than the *Cis*-regulatory element and the Hammerhead ribozyme has the simplest structure.

We performed a 10-cross-validation experiment using specificity and sensitivity for evaluation. Our training data consists of nucleotide, RNA secondary structure feature and RNA family label. The testing data consists only of RNA sequence nucleotides and their secondary structures.

In our experiment, we used the CRF++ 0.53 tool [13] which is an open source program providing a generalized CRFs learning and testing platform.

3. RESULTS AND DISCUSSION

We used the evaluation method that was developed in [2, 14] to show the performance of our CRFs classification model using sensitivity and specificity. For each RNA family, the sensitivity is represented by the percentage of true positives of test sequences of the particular RNA family. The specificity is calculated as the percentage of true positive of test sequences which are classified in the family.

Let us denote x_{ij} be the number of sequences of family i (row) that are predicted to be in family j (column) in Table 2.

Table 2. Sensitivity and specificity of the 10-cross validation experiments.

Corrected RNA Families	Classified RNA Families				Sensitivity (%)
	SRP	5S RNA	Cis-Regulatory	Hammerhead	
SRP	39	0	0	0	100
5S RNA	0	16	0	0	100
Cis-regulatory	0	0	4	0	100
Hammerhead	1	0	0	13	92.86
Specificity (%)	97.50	100	100	100	

The sensitivity for family $i = \frac{x_{ii}}{\sum_{k=1}^{4} x_{ik}}$.

The specificity for family $i = \frac{x_{ii}}{\sum_{k=1}^{4} x_{ki}}$

From Table 2, 5S RNA and *Cis*-regulatory element are correctly classified, so their sensitivity and specificity of classification are 100%. However, one Hammerhaed ribozyme was incorrectly classified to be SRP, hence the specificity of SRP and sensitivity of Hammerhead ribozyme classification are 97.50% and 92.86%, respectively. The sensitivity of SRP classification is 100% because there is no SRP classified to be other families or wrong classified. The specificity of the Hammerhead rybozyme classification is also 100% because there are no other

families classified to be Hammerhead rybozyme.

The average evaluations of 10-cross validation are shown in Table 3. The evaluations show that our RNA family classification using the CRF model with RNA secondary structures as features can correctly classify RNA families of the test RNAs with high sensitivity and specificity. Since each RNA family is defined by members which have a conserved function in different species, our data shows that the CRF function has a strong relationship to the RNA secondary structure using only the primary sequence.

Table 3. Average of specificity and sensitivity of 10-cross validation experiment.

Family	Sensitivity (%)	Specificity (%)	F-Measure
SRP	99.74	98.74	99.24
5S Ribosomal RNA	98.14	100	99.06
Cis-regulatory element	97.56	100	98.77
Hammerhead ribozyme	99.32	99.32	99.32

4. CONCLUSION

In this paper, we defined the RNA family classification problem as a sequence labeling task. We proposed a graphical discriminative model using CRFs to solve the sequence labeling problem. Then we selected RNA secondary structures with biologically feasible information from the RNA STRAND database as the feature data of the RNA sequences, and we constructed a suitable template for the CRFs to train and test the data. The Experimental results show that our CRFs model can classify RNA families of the test RNAs with high sensitivity and specificity.

ACKNOWLEDGMENTS

We would like to thank National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand Graduate Institute of Science and Technology (TGIST), National Science and Technology Development Agency (NSTDA) (Contract Grant No. TGIST 01-51-013) and The Institute for The Graduate School Chiang Mai University for financial support.

REFERENCES

[1] Wang J.T. and Wu X., Kernel design for RNA classification using support vector machines, *Int. J. Data Mining and Bioinformatics*., 2006; **1**: 57-76.

[2] Yoon B.J. and Vaidyanathan P.P., Computational identification and analysis of noncoding RNAs-unearthing the buried treasures in the genome, *IEEE Signal Processing Mag*., 2007; **24**: 64-74.

[3] Siederdissen C.H. and Hofacker I.L., Discriminatory power of RNA family models, *Bioinformatics*, 2010; **26**: 453-459.

[4] Mattick J.S. and Makunin I.V., Non-coding RNA, *Human Mol. Gen*., 2006; **15**: 17-29.

[5] Wallach H.M., *Conditional Random*

Fields: An Introduction., University of Pennsylvania: Technical Report MS-CIS-04-01; 2004.

[6] LeCun Y.B., Bengio L.Y. and Haffner P., Gradient-based learning applied to document recognition, *Proc. IEEE*., 1998; **86**: 2278-2324.

[7] Lafferty J., McCallum A. and Pereira F., Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Machine learning-International workshop then conference*., 2001; 282-289.

[8] Chen S.F. and Rosenfeld R., A Gaussian prior for smoothing maximum entropy models. School of Computer Science, Carnegie Mellon, University: Technical Report CMU-CS-99-108; 1999.

[9] Liu D.C., and Nocedal J., On the limited memory BFGS method for large scale optimization, *Math. Program*., 1989; **45**: 503–528.

[10] Sha F. and Pereira F., Shallow parsing with conditional random fields, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Edmonton, Canada: Association for Computational Linguistics*., 2003; **1**: 134–141.

[11] Phan H.X. and Le Nguyen M., FlexCRFs: Flexible Conditional Random Fields. http://flexCRF.sourceforge.net., 2004; [accessed 03/09/2010].

[12] Andronescu M., Bereg V., Hoos H.H., and Condon A., RNA STRAND: the RNA secondary structure and statistical analysis database, *BMC Bioinformatics,* 2008; **9**: 340-349.

[13] Kudo T. (2005). *CRF++: Yet another CRF toolkit*. Obtained through the Internet: http://crfpp.sourceforge.net, [accessed 15/12/2010].

[14] Baldi P., Brunak S., Chauvin Y., Andersen C.A. and Nielsen H., Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics,* 2000; **16**: 412-424.