Determine ncRNA structure shape using context-free grammar and support vector machine

Rujira Achawanantakun
Michigan State University
rujiraa@gmail.com

Yuan Zhang
Michigan State University
zhangy72@msu.edu

Abstract

Non-coding RNA molecules perform their cellular roles through their primary sequences as well as secondary structures. A model of secondary structure is required to determine mechanism of actions of ncRNA sequences. Secondary structure prediction is mainly determining all the possible base pairs of a given ncRNA sequence. In this work, we first make use of a structure prediction tool, Mfold to predict structures for each sequence of our benchmark dataset. Then we represent each of the predicted structures with a simple string notation. To achieve this goal, we introduce a notation representing ncRNA structures in the parameter space of a context-free grammar. Then we extract the structure pattern from each predicted structure. We call these patterns structure shapes or abstract shapes. SVM is used to classify all the predicted shapes into positive and negative classes. By keeping a track of the shapes, we can determine the consensus shape for each dataset. We compare our results with an existing tool called RNAcast. Our ncRNA notation using context-free grammar is simpler and more effective than the tree structure used in RNAcast. Moreover our prediction accuracy on Balibase datasets is much higher than RNAcast.

1 Introduction

Non-coding RNAs (ncRNAs) refer to RNA molecules that are not translated to proteins. NcRNAs play important roles in a variety of cellular processes such as RNA splicing and gene regulation. An ncRNA performs its biological functions via not only its primary structure which is its linear sequence of nucleotides, but also its secondary structure. Figure 1 is an example of a transfer RNA sequence as well as its structure.

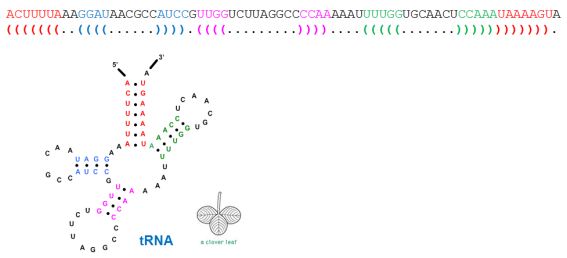

Figure 1: Transfer RNA structure

It is of great importance to accurately determine ncRNA secondary structures in order to understand the process they perform their functions through their structures. Unlike coding RNA molecules, ncRNA molecules have their secondary structures conserved through evolution. To understand the mechanism of actions of a ncRNA sequence, a model of the structure is required. Since experimental determination of ncRNA structures, including NMR and X-ray crystallography, is time consuming and expensive, computational approaches are of great interest to ncRNA structure prediction. However, the number of possible structures based on thermodynamic models is much larger than the number of input sequences. Moreover, many of these predicted structures are not similar or close to the native ones. To improve the accuracy of structure

prediction tools, two problems need to be addressed. The first one is how to use a simple and effective representation for ncRNA secondary structures. The other is to propose a mechanism which can choose true structures that are most likely to represent the native structures. In this work, we propose a ncRNA secondary structure representation that encodes an ncRNA's sequence and secondary structure in the parameter space of context-free grammar (CFG). We define a feature set and apply support vector machine (SVM) to find possible structure shapes for the datasets. For a dataset with more than one possible shapes we will select the one with the minimum value of average free energy over all the predicted structures contributing to the shape.

2 Related Works

In recent years, many methods have been proposed for de novo RNA secondary structure deviation. The first category is to use experimental methods, such as NMR and X-ray, to determine ncRNA secondary structure. These methods are time consuming and dependent on the initial models. Another category is comparative method, which has high accuracy in prediction. However, this method requires a lot of homologous sequences which in some cases are not available. Another important category of methods is dynamic programming techniques. Base pair maximization method by dynamic programming tries to find the best score by aligning bases and chooses the path which produces the optimal alignment score. This method suffers from several problems. For one thing, the structure which has the best alignment score doest not necessarily lead to the most stable structure. For the other, base pair maximization is comparative to sequence alignment which is not biological reasonable in this case.

Free energy minimization (FEM) is also used in ncRNA secondary structure prediction and has achieved an average of 73% accuracty Mathews (2004). It applies dynamic programming to maximize the score taking into account thermodynamics based on the theory that the most probable strucutre corresponds to the most stable one. In our project we make use of an implementation of FEM method called Quikfold to predict possible secondary structures for our sequences from the test datasets.

Robert (2004) formalized the concept of abstract shape. An abstract shape of an RNA molecule comprises a class of similar structures. And each shape has a representative structure which has the minimum free energy within the class. Different levels of shapes give different levels of abstraction. The higher level the shape is, the more abstract the shape is. Figure 2 is an example of different levels of abstract shape for the same secondary structure. Figure 3 is an example of different secondary structures sharing the same abstract shape.

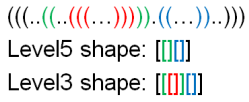

Figure 2: An example of different levels of abstract shapes

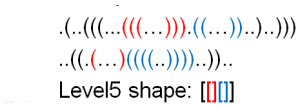

Figure 3: An example of different structures sharing the same shape

Shape analysis was implemented in the program *RNAshapes*. They represented each predicted structure using the tree data structure. The *RNAsubopt* from the Vienna RNA package Vienna (1994) was used as a structure prediction tool for complete suboptimal folding. And then they applied *RNAshapes* to the prediction of optimal and suboptimal abstract shapes of several RNAs. Using a tree to represent an ncRNA structure led to the complication of transformation between input structure and representative structure. Also it is computational expensive to compare two secondary structures represented by tree structures. Jens (2005) used abstract shape proposed by Robert (2004) in RNA consensus structure prediction. Jens (2005) showed that the choice of abstraction level should be the less abstract level at which the information of stems in the shape is retained. Compared with the strongest abstraction that does not account for bulges at all, the less abstract level tends to output a true shape in the list of near-optimal shapes. This method has been implemented in the program *RNAcast*. The algorithm starts with

running *RNAshapes* on each sequence in a set of homologous sequences. They evaluated each shape using a scoring function. Then, the one with the highest score is returned as the consensus shape. Since this algorithm relied on *RNAshape*, it still suffered from complicated representation of a tree data structure. In addition, the scoring function only considered one feature, sum of free energies. Other possible features such as number of structures which share one shape are not taken into consideration. This may lead to low accuracy in detecting the true structure.

3 Methodology

3.1 NcRNA notation

We introduce an ncRNA representation in the parameter space of context-free grammar (CFG). Using CFG, the representation of both ncRNA sequence and secondary structure are as simple as a one-dimension string. Compared to the annotation in RNAcast this notation is simpler and can more effectively facilitate further manipulation. In order to represent ncRNA structures, we define six production rules as well as the encoding algorithm. A comparison between our annotation and that used in RNAcast is shown in Figure 4.

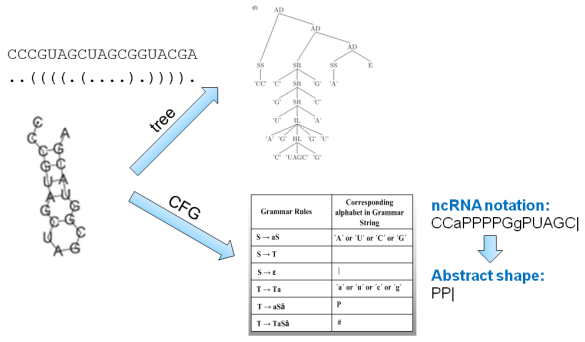

Figure 4: Comparison between our ncRNA notation and RNAcast

We propose the notation with a set of characters, {A,C,G,U,a,c,g,u,♯,|}. Each character corresponds to each production rule. Traditionally, there are two steps in ncRNA notation process. First, we need to create a sequence of production rules to generate an ncRNA sequence and its secondary structure. Second, we need to transform the sequence of production rules into an ncRNA notation according to the definition of predefined characters. In our project, we propose an effective dynamic programming algorithm to convert the input ncRNA sequence and its secondary structure into the grammar annotation as shown in Figure 5. This improves the performance of the transformation process.

```
void parse(i, j)
{
if i >= j
        print '|';
        return;
else if Xi is a single stranded base
        print uppercase of Xi;
        i++;
        parse(i,j);
else if  Xj is a single stranded base
        print lowercase of  Xj;
        j--;
        parse(i,j);
else if  Xi and Xj form a base pair
        print 'P';
        i++ and j--;
        parse(i,j);
else
        print '#';
        k = the position that forms a base pair with Xj;
        parse(i,k-1);
        parse(k,j);
}
```

Figure 5: The algorithm which converts the input sequence and secondary structure into ncRNA notation

We transform an ncRNA sequence and its structure to an ncRNA notation with time complexity $O(L^2)$ where L is the length of ncRNA sequence. As an example, a sequence *CCCGUAGCUAGCGGUACGA* and its secondary structure *..((((.(....).)))).* can be encoded as *CCaPPPPGgPUAGC|* with the production rules.

3.2 Support Vector Machine

In this project we make use of Support Vector Machine (SVM) to distinguish correct abstract shapes from incorrect ones. SVM is a powerful tool for supervised learning. It basically builds a hyperplane to separate data points from different classes with the largest margin or boundary. This hyperplane will be used to do classification on the test set. Figure 6 SVM (2011) is an example of SVM classification.

In the classification stage, each data point corresponds to a predicted abstract shape for each ncRNA

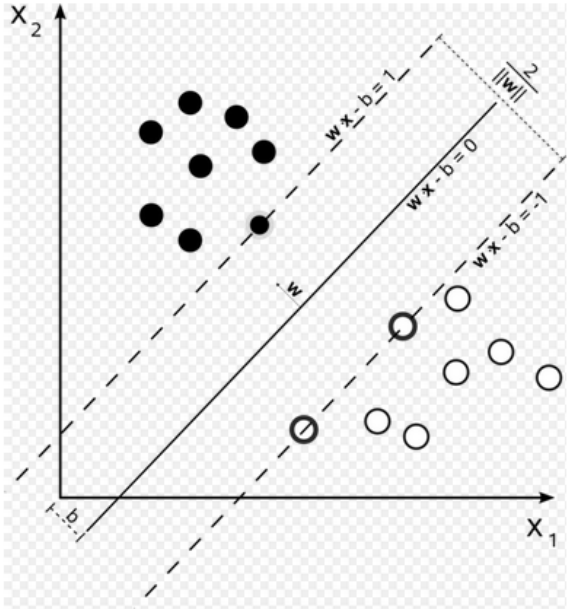

Figure 6: An example of SVM

secondary structure. Four features will be extracted to map the data point into a high dimensional space. Due to the fact that values of different features have quite different scales we adopt two kinds of scaling methods: normalization and standardization. For the former one, we use the following equation:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x is the value of the feature we want to scale. This effectively maps all the values of the feature into the range of [0,1]. For the latter, the following equation will be applied:

$$x_{new} = \frac{x - \mu}{\sigma}$$

where μ is the mean of all the values of x and σ is the standard deviation. This will convert the variable into a new variable of zero mean and unit variance.

Both methods will be used in our experiments to do data scaling. We will then choose the one which gives us better performance in classification result.

3.3 Consensus shape derivation

Major steps of consensus shape derivation are shown in Figure 7 and explained below.

1. Apply a structure prediction tool on each sequence in a group of homologous sequences.

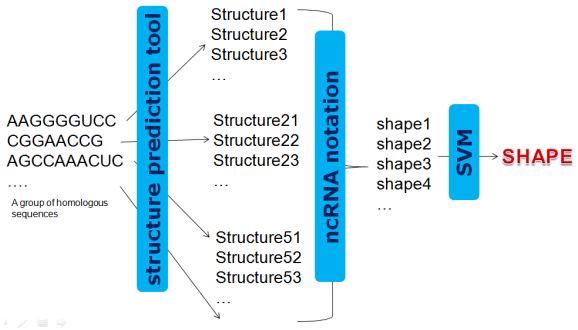

Figure 7: Consensus shape derivation pipeline

Mfold, the structure prediction tool based on minimum free energy, is applied on each sequence to obtain predicted structures which have free energy with a range of 5% from the MFE structure.

2. Represent each ncRNA sequence and its predicted structures using our annotation. Each sequence and its structures will be transformed into a simple notation using defined production rules and dynamic programming algorithm.

3. Transform ncRNA notation to an abstract shape space. Ignoring all single stranded regions and the length of stems, an abstract shape provides a rough-cut description of ncRNA secondary structures. However, when user are interested in structures with fundamental difference, information gained from abstract shape can be used to fulfill the requirement as well.

4. Each abstract shape will be represented as a data point in the classification model. The true labels of the datasets are assigned by comparing extracted abstract shapes with those of reference structures from Rfam database.

5. Create SVM classification models. The whole dataset is divided into two parts, training set and testing set. 70 percent of the dataset is used as training set and the rest is used as testing set. We will build an SVM model on the training set and it will be used for label prediction on test set.

6. Predict labels of abstract shapes on the testing set. The abstract shapes which are predicted as

positive are assumed as correct shapes, while negative abstract shapes are assumed as incorrect shapes. In case there are more than one correct shapes for the same Balibase dataset, the shape which has lowest average free energy will be selected as the shape representation for the dataset.

3.4 Performance metrics

In order to precisely assess the predictive power of the classification method, we use some typical measurements, which have been extensively used in the field of bioinformatics. The measurements used in our study include sensitivity, specificity and accuracy. Sensitivity is defined as:

$$\frac{\text{number of correctly predicted positive shapes}}{\text{total number of positive shapes}}$$

Specificity is defined as:

$$\frac{\text{correctly predicted negative shapes}}{\text{the total number of negative shapes}}$$

Accuracy is defined as:

$$\frac{\text{the number of correctly predicted datasets}}{\text{the total number of datasets}}$$

4 Experiment and result

4.1 Dataset

Our datasets in this project are from BAliBASE (Bahr, 2001). BAliBASE is an ncRNA benchmark database. It contains many datasets with different number of sequences, 2 to 15 sequences in each dataset. Each dataset corresponds to ncRAN sequences belonging to the same family. Previous works showed that sufficient number of sequences is essential to get a correct consensus structure. So, we select datasets which contain 15 sequences as the input for our experiment. The reference shape which corresponds to each dataset is derived from Rfam database. The reference shape is used as a gold standard in label assigning process and performance evaluation step.

4.2 Structure prediction

There are various tools to predict the secondary structures of ncRNA sequences. Basically, they search for structures with the minimum free energy (MFE). The structure prediction tool we use is Mfold (Zuker, 1981). Mfold has been widely used to predict ncRNA structures. It is based on dynamic programming and has a complexity of $O(n^3)$, where n is the sequence length. Mfold is based on the principle that the most stable secondary structure is the one with the minimal free energy. Mfold's parameters used in the experiment are listed here. Folding temperature is 25 degree Celsius. Minimum free energy range is 5%. And in order to avoid the explosive number of predicted structures, the number of predicted structures is limited to 50 structures per sequence.

4.3 Classification

A library for support vector machines (LIBSVM) implemented by Chih-Jen Lin (LIBSVM, 2010) is used to create classification model and perform classification task. We propose four features which combine thermodynamic and sequence contribution information. The four features are listed below.

1. Frequency of predicted structure: the number of predicted structures which can be derived into this abstract shape divide by the number of all predicted structures.
2. Average free energy of predicted structures.
3. Average of minimum free energy of predicted structures.
4. The number of sequences which have this abstract shape

In order to scale the feature values we perform both normalization and standardization on the datasets as we described in methodology part. Our experimental results show that standardization works better in this experiment. So we will scale the datasets using standardization method. The transformed values from standardization are also called z-scores.

To understand the importance of various features to the classification model, we performed an feature selection analysis in which we removed various sets of features from the model and assessed the change

Table 1: Performance using different features

features	specificity	sensitivity	accuracy
1 and 4	48.61	99.47	18.02
2 and 4	50.56	99.43	18.55
four features	61.23	99.25	66.76

in prediction result. As shown in Table 1, the performance of the classifier degrades as features are removed from the model. This indicates all the features indicate useful information of the data points, which are the predicted shapes. The model which combined all the four features achieved the best performance. Thus, the final classification model is created using all four features with the linear kernel function, which shows better performance than the other kernel functions.

4.4 Comparison with RNAcast

RNAcast was chosen as a state-of-the-art representative of RNA consensus structure predictors. RNAcast algorithm starts with taking a group of homologous sequences as an input. Then it uses RNAsubopt to predict structures of each input sequence within an MFE threshold. RNAcast identifies all shapes that occur in the predicted structure list. After that, the scoring function is used to obtain a sorted list of all common shapes. The first shape of this list is returned as the consensus shape. The three scoring functions proposed by RNAcast are listed below.

1. Rank sum score: each shape representation contributes with its individual rank in the sorted shape of its sequence rank

2. Sum of energies: sum of energies of predicted structures

3. Sum of probabilities: the sum of probabilities coming from the partition function MacCaskill (1990).

From the experiment, RNAcast found that $rank_2$ performs best, followed by $rank_3$.The rank sum score gives the worse performance. So, RNAcast used $rank_2$ in consensus shape derivation process.

Table 2: Comparison between RNAcast and our approach

method	RNACast	our approach
structure prediction tool	RNAsubopt	Mfold
MFE range	10 kcal/mol	5% above MFE
can handle IUPAC character	no	yes
number of input sequence	10	no upper bound

The method used by RNAcast is similar to our approach, except for the ncRNA representation and shape selection steps. RNAcast uses a tree structure to represent ncRNA structures, while we use CFG and transform an ncRNA structure into a one-dimensional string. For the shape selection step, RNAcast uses a scoring function, which considers only one feature at a time, to rank shapes. In our approach, we integrate four feature vectors and use SVM classifier to identify the consensus shape. Our method improves the accuracy clearly as shown in Table 1. We achieved 66.75% accuracy, while RNAcast achieved 46.52% accuracy.

4.5 Analysis

The improvement of accuracy we have achieved is due to several differences between RNAcast and our approach. The differences are shown in Table 2 and explained below.

1. Energy range: The energy range used by Mfold is set using percentage as its scale. The default value is 5%. While energy range of RNAcast is fixed by the number of energy per molecule, the default range is 10 kcal/mol. Changing energy range can expand or limit the structure search space, and hence the choice of the energy range is critical. For example, if the true shape is missing from the shape space, this shape can be neither a common shape nor a consensus shape. Thus, predicted consensus shape must be wrong in this case. Energy of structures varies depending on sequence length and folding. Therefore, Mfold used in our approach provides a more reasonable and flexible energy

range by using percentage, instead of energy range using a fixed number units higher than the MFE in RNAcast.

2. IUPAC handling: IUPAC is degenerate base symbols in biochemistry. It represents a position on a RNA sequence that can have multiple possible nucleotides. RNAcast removes every non $\{a, c, g, u, A, C, G, U\}$ character from the sequence data. This may lead to wrong structure prediction in case that many degenerate bases appear in an input sequence. The consequence will be the wrong prediction of the consensus shape.

3. The number of input sequences: there are no upper bound for the number of sequences to be considered in our approach. In terms of efficiency, more input sequences provide higher accuracy of consensus shape prediction. However, in RNAcast's approach when the number of input sequences is too large the energy range may need to be extended in order to incorporate more true structures. However, this can also include many outliers. Then the downstream analysis will be greatly affected. And if we limit the number of predicted structures the true structure may be missing from the search space. So it is not practical for RNAcast to only allow no more than 10 sequences.

5 Conclusion and future work

In this project we propose a simple ncRNA notation using context free grammar. Compared to previous work which uses complex data structure our secondary structure annotation is much simpler. At the same time it enables us to efficiently derive the abstract shape for each structure. And then we extract four features for each abstract shape which are used to classify true shapes for each dataset. SVM is used for classification. We build a classification model based on our training set and predict true shapes in our test set. Then we select an abstract shape for each dataset and compare it to the true label. We achieve a good trade-off between classification sensitivity and specificity. We compare our method with RNAcast. In terms of prediction accuracy for BAliBase datasets, we find that our accuracy is much higher. Our contribution in this project lies in two aspects. First of all we propose a simple and efficient notation to represent ncRNA secondary structures. Compared to current representation of ncRNA secondary structures such as complex tree structures, our method is much more efficient and easier to process for downstream analysis. Also it is easier to extract features for each shape using this simple string notation. Secondly, we convert the problem of shape prediction into a classification problem by extracting useful features from our structure prediction results.

As we can see from the experiment results, our classification sensitivity and specificity are not high enough to make confident consensus shape prediction. Also accuracy remains to be improved. To achieve these goals we can focus our future work on the following aspects. First of all, we need to filter some predicted structures by Mfold. In our current experimental settings, we have 50 predicted structures for each sequence. From our classification results, we find that this limitation is still too loose to filter enough noise. This noise may harm downstream feature extraction and classification. To do so we may apply a threshold to remove those structures with too high free energy. These structures may not have enough evidence to be the real structures. So their contributions should not be added to candidate abstract shapes. Also we can filter abstract shapes which do not have enough supporting structures. In the classification part, SVM does not give satisfactory performance for this specific problem. So we may consider other classification tools such as logistic regression model or Adaboost to classify abstract shapes. These methods may help to improve performance of our abstract shape prediction.

References

Mathews, D.H. and Disney, M.D. and Childs, J.L. and Schroeder, S.J. and Zuker, M. 2004. *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*. Proc. Natl. Acad. Sci. (2004) 101: 7287-7292.

Jens Reeder, Robert Giegerich 2005. *Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction*. Bioinformatics. 2005 Sep 1;21(17):3516-23. Epub 2005 Jul 14.

Robert Giegerich, Bjorn VoB, and Marc Rehmsmeier

2004. *Abstract shapes of RNA*. Nucl. Acids Res. (2004) 32 (16): 4843-4851.

Ivo L. Hofacker , Walter Fontana , Peter F. Stadler , L. Sebastian Bonhoeffer , Manfred Tacker , Peter Schuster 1994. *Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package)*. Monatshefte Fur Chemie (Chemical Monthly) 125: 167-188 (1994)

Anne Bahr, Julie D. Thompson, J.-C. Thierry, and Olivier Poch 2001. *BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations*. Nucleic Acids Res. 2001 January 1; 29(1): 323-326

Michael Zuker and Patrick Stiegler. 1981. *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* . Nucl. Acids Res., 9(1):133-148, 1981

Chih-Chung Chang and Chih-Jen Lin 2010. *http://www.csie.ntu.edu.tw/ cjlin/libsvm/.*

MacCaskill, J.S. 1990. *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*. Biopolymers, Vol. 29, pp.1105-1119

Wikipedia, Support Vector Machine *http://en.wikipedia.org/wiki/Support_vector_machine*.