EVALUATION OF RNA SECONDARY STRUCTURE MOTIFS USING REGRESSION ANALYSIS

Mohammad Anwar
School of Information Technology and Engineering, University of Ottawa
e-mail: manwar@site.uottawa.ca

Marcel Turcotte
School of Information Technology and Engineering, University of Ottawa
e-mail: turcotte@site.uottawa.ca

Abstract

Recent experimental evidences have shown that ribonucleic acid (RNA) plays a greater role in the cell than previously thought. An ensemble of RNA sequences believed to contain signals at the structure level can be exploited to detect functional motifs common to all or a portion of those sequences. We present here a general framework for analyzing multiple RNA secondary structures. A family of related RNA structures may be analyzed using statistical regression methods.

In this work, we extend our previously developed algorithm, Seed, that allows to explore exhaustively the search space of RNA sequence and structure motifs. We introduce here several objective functions based on thermodynamic free energy and information content to discriminate native folds from the rest. We assume that the variation across the various scores can be represented by a statistical model. Regression analysis permits to assign separate weight for each score, allowing one to emphasize or compensate the variance that differs across the different scores. A statistical model can be formulated using techniques from regression analysis to obtain a template or scoring model that is able to identify putative functional regions in RNA sequences.

We show that thermodynamic based regression models are effective to associate the variation of scores obtained from different functions. The models can generally identify motifs with high measures of specificity and positive predicted value to known motifs. A good scoring method will allow to eliminate invalid motifs thereby reducing the size of the hypothesis space.

***Keywords**: Motif discovery; ribonucleic acid; secondary structure; thermodynamics; linear regression.*

1 Introduction

While the main role of DNA (*deoxyribonucleic acid*) is the storage of genetic information, RNA (*ribonucleic acid*), a versatile bipolymer, has many different biological roles to fulfill. The widely recognized functions include carrying and recognizing genetic information as messenger RNA (mRNA) and transfer RNA (tRNA). In addition to these, RNA has been implicated in several other biological events.

Recent developments have shown that RNAs also have important regulatory functions. Examples include micro-RNAs that regulate the expression of genes by binding to the 3'-untranslated regions of specific mRNAs [1]. They exercise control over those RNAs that code for proteins. The function of these regulatory elements depends on the presence of motifs that are conserved both in structure and more loosely in sequence.

Post-transcriptional regulation of gene expression often involves secondary structure elements located in the untranslated regions of mRNAs [2]. These untranslated regions share distinct structural features in RNA molecules and are correlated with their biological function. Finding these similar structural features in a group of RNA sequences believed to share the same function can provide substantial information predicting which parts of the sequence are functionally important.

To address the requirement of computational tools to analyze RNA sequences, various methods have been developed. The most popular approach to structure prediction is perhaps through energy minimization. Here, the free energy modeled as the sum of the contributions of independent cycles using nearest neighbor model [3]. Although the predictions have been on an average about 70% accurate [4], low accuracy predictions are also made. There are several reasons why free energy minimization methods can fail. Firstly, the lowest free energy conformation may not coincide with the native conformation. This can be due to experimental errors in determining the free energy parameters, or simply because there are several free energy conformations, and it can be difficult to distinguish the native conformation from the others. Secondly, not all the factors that determine the energy of a fold are understood, and computational time limitations would make it infeasible to include all of such influences. Many situations are highly approximated, especially the calculation of multiloop structures. No tertiary interactions are included in the model, no pseudoknotted structures and no interactions with the cellular environment are considered.

Seed is a data-driven exploratory tool that is designed to search a space of conserved RNA secondary structure motifs [5], [6]. It identifies all the conserved RNA structure motifs from a set of unaligned sequences. The approach used for discovering secondary structure motifs has two components. First, the search space is generated from the seed sequence using suffix arrays. Secondly, suffix arrays are used to match secondary structure elements. See [7] for an introduction to suffix arrays. The main steps of the Seed algorithm are as follows.

1. Select a seed sequence;
2. Construct the most specific motif;
3. General-to-specific search of the motif space;
4. Report the motifs.

Seed is built with user defined parameters to relax or restrict the size of the search space and accordingly the exe-

1-4244-0038-4 2006 IEEE
CCECE/CCGEI, Ottawa, May 2006

cution time. It also allows to impose a running time restriction for cases when the search space is very large. Unlike other motif finding algorithms, Seed has the ability to predict consensus secondary structures with conserved base pairs.

2 Method

The purpose of a scoring function is to distinguish the biological relevant motifs amongst an ensemble of motifs, that is, to approximate the biological meanings of the motifs in terms of a mathematical function. We expect the shape of the molecule to be primarily determined by the bases involved. As a result, understanding the factors determining the stability of an RNA structure is important in predicting the correct structure. There are several factors which contribute to the stability of a RNA structure. They are,

- The number of G-C vs A-U and G-U base pairs (stable structures contain higher energy bonds);
- Number of base pairs in stem (longer stems result in more bonds);
- The number of base pairs in a hairpin loop;
- The number of unpaired bases (unpaired bases decrease the stability).

An effort to circumvent the limitations of the nearest neighbor model is to have a linear combination of the free energy of multiple input sequences, when folded into a common structure [8], [9], [10]. As the number of input sequences increases, it is unlikely that the common structure(s) will fold into a bad minimum free energy. Although the results of using more sequences are promising, the prohibitive time/space complexity of these approaches restricts their use to few sequences (< 4) that are less than hundred nucleotides long. Considering the above contributing factors and limitations to the model, we introduce several functions combining the free energy of all or some of the matches of a given motif. These functions are: *TSum*, *TBest* and *TWorst*.

Seed is a framework for finding conserved RNA motifs in a set of unaligned sequences. It produces an ensemble of structures that are present in the majority of the input sequences. Here, we study several scoring schemes and evaluate their potential to discriminate native folds from the rest.

Each motif matches at least $min\_support \times k$ sequences, and up to k sequences; where *min_support* is a user defined parameter. Also, certain motifs will occur more than once in any sequence. Thus, there are several possible approaches to calculate the free energy score for a given motif and set of matches. For a given motif, *TSum* is defined as the sum of the free energy of all the occurrences in all the matching sequences. *TBest* is the sum of the lowest free energy match in each sequence. Finally, *TWorst* is the sum of the highest free energy match from each sequence. In particular, given a set of sequences and matches, the functions are calculated as

$$\begin{aligned} TSum &= \sum_i \sum_j m_{i,j} \\ TBest &= \sum_i \min_j m_{i,j} \\ TWorst &= \sum_i \max_j m_{i,j} \end{aligned}$$

where $m_{i,j}$ is the free energy of the jth match in the ith sequence.

Information content has often been used in the context of sequence pattern discovery. Accordingly, we include a function that consists of the sum of the information content contributions from unpaired and paired regions. Shannon uncertainty (H) was calculated for each loop position and was subtracted from the maximum uncertainty possible, to give the information content (in bits). $H = -\sum P_i \log_2 P_i$ summed over each base pair (i = A, U/T, G, C), where the observed nucleotide frequencies of each base i from the input sequences is used to estimate P_i. A nucleotide in a stem is base paired to its pairing partner, which increases the information content relative to an unpaired nucleotide in a loop. There are sixteen possible two nucleotide pairs giving a maximum uncertainty of 4 bits. Out of these pairs, there are four Watson-Crick base pairs (A-T, T-A, G-C, C-G). For a Watson-Crick base pair, the uncertainty is 2 bits making the information content 4 bits - 2 bits = 2 bits. If G-U (and U-G) base pair are considered, the uncertainty is reduced to 2.8 bits; the information content of this pairing is 4 bits - 2.8 bits = 1.2 bits. The resulting loop and stem information contents were added to calculate the total information content.

In [11], the correlation between the normalized energy (energy per base pair) and % G-C was studied. It was observed that they are negatively correlated. This is expected due to the fact that G-C base pairs have lower free energy than the other base pairings. We include this distinguishing feature with the above defined functions.

Correlation coefficient was studied in [6] to measure the usefulness of these scoring functions. Strong correlation was observed between the normalized scores and Positive Predicted Value (PPV)/sensitivity across the datasets used. The value of r ranged from -0.783 to -0.95. For some instances, it was even lower (better). However, no particular function outperformed the other; the highest correlation coefficient was different across each dataset. It was also pointed out that the linear combination of free energy score and information content outperformed either of the two scores alone.

Since the free energy scores were particularly effective to separate high PPV/sensitivity motifs from the rest, we assume that variations in the scoring functions can be represented by a statistical model. We used the most classical field of statistical analysis, regression, to analyze a set of RNA secondary structure motifs. A statistical model can be constructed using techniques from regression analysis to obtain a template or model that is able to discriminate native folds from the rest.

In order to remove the undue influence of the scores with large outlying numerical values, we normalize the variants by dividing the scores with the number of sequences matched and number of matches (for *TSum*). In addition, we re-normalize the scores by length and number of base pair present in the motif. Multiple regression models were built to show the relationship between the set of scores obtained above and different performance measures (see below). The measures were used as the response parameter for building the models.

2.1 Building Base Models Using Selection Criterion

Having obtained all the explanatory variables, we define the maximum model, that is, the model containing all the explanatory variables that could possibly be present in the final model. Including all the explanatory variables may introduce the risk of collinearity, that is, two or more variables being linearly dependent. One of the common purported remedy of collinearity is variable selection. Since the number of variables in our case is not large, we opted for a best subset regression.

The number of possible models built for each data set with different variants is large. A statistical selection criterion approach is taken to determine which model is better than the rest. Although many different selection criteria have been suggested through time, we used measures of coefficient of determination (R^2) and Mallows C_m criterion to pick the best model for each dataset used.

2.2 Performance Measures

2.2.1 Matthews Correlation Coefficient, PPV

We call **references**, the secondary structures that were obtained from the tRNA compilation by Sprinzl and the Comparative RNA Web Site. We define as **true positives** (TP) the base pairs that are occurring in both structures, reference and predicted, **false positives** (FP), the base pairs that are occurring in the predicted structure but not in the reference one, and **false negatives** (FN), the base pairs that are occurring in the reference structure but not in the predicted one. Offsets were not allowed.

The **positive predictive value** (sometimes called PPV) is defined as the fraction of the predicted base pairs that are also present in the reference structure, $TP/(TP+FP)$. The **sensitivity** is defined as the fraction of the base pairs from the reference structure that are correctly predicted, $TP/(TP+FN)$. Finally, we also measured the **Matthews Correlation Coefficient**, as defined by Gorodkin, Stricklin and Stormo [12]:

$$\sqrt{\frac{TP}{(TP+FN)} \times \frac{TP}{(TP+FP)}}$$

When a given motif matched the input sequence more than once, the performance indexes of the match having the highest PPV are reported.

2.2.2 Distance

The measures described above quantify the agreement between the predicted and reference structures with respect to the position of the base pair in the sequence under consideration. Although this is a refined measurement, it does not consider similar structures placed at an offset greater or less than the length of the structure.

In order to have a more general way of comparison, we consider to measure the dissimilarity between the secondary structures. A simple measure would be the base pair distance, that is, the number of base pairs that have to be opened or closed to transform one structure into the other. However this comparison restricts both the structures to be of same length. Another measure to compare structures is by encoding the secondary structure by ordered trees, followed by the computation of edit distance. This approach is a part of the Vienna RNA package [13]. The edit distance defines a metric of the number of insertion, deletion and replacements of nodes in the tree.

2.3 Data

All the 3' UTR entries containing the keyword histone as well as an HSL3 feature were extracted from UTRdb release 19 [14]. A total of 28 sequences was obtained. The length of the sequences varies from 51 to $1,955$ nucleotides, with an average length of 701 nucleotides. The dataset consists of the following entries: 3HSA054868, 3HSA041812, 3HSA027954, 3HSA034695, 3HSA079397, 3HSA082131, 3HSA047510, 3HSA083260, 3HSA083338, 3HSA083659, 3HSA048427, 3HSA049188, 3HSA084501, 3HSA086570, 3HSA086915, 3HSA087013, 3HSA089561, 3HSA058723, 3HSA058724, 3MMU017942, 3MMU040716, 3MMU043604, 3MMU045939, 3MMU-046704, 3MMU004991, 3MMU004994, 3MMU004995 and 3DRE-005245.

All the mammalian 5' UTR entries containing the keyword ferritin and a valid IRE motif were extracted from UTRdb release 19 [14]. A total of 14 sequences was obtained. The length of the sequences varies from 58 to 2,188 nucleotides, with an average length of 378 nucleotides. The dataset consists of the following entries: 5DLE000003, 5HSA021933, 5HSA033035, 5HSA060296, 5HSA072191, 5HSA073036, 5HSA079314, 5MMU018600, 5MMU-025452, 5MMU027798, 5MMU032372, 5RNO004780, 5RNO005974 and 5RNO007816.

A tRNA dataset was assembled using a subset of the sequences from Masoumi and Turcotte [10]. Seven sequences having approximately the same length were used. These are generally challenging sequences for traditional approaches, such as MFOLD [15]. The following entries were extracted from the compilation by Sprinzl *et al.* [16], [17]: RD0260, RD0500, RD1140, RD2640, RE2140, RE6781 and RF6320.

TABLE I

Runtime statistics

Experiment	#Sequences	#Motifs	#Matches
HSL3	27	357	1,945,328
IRE	13	110	167,076
tRNA	7	5,518	3,407,012
5S	7	365,505	152,741,463

Similarly, a 5S dataset was assembled using a subset of the sequences from Masoumi and Turcotte [10]. Seven sequences having approximately the same length were used. These also are generally challenging sequences for traditional approaches, such as MFOLD. The following entries were extracted from the Comparative RNA Web Site [18], [19], [20]: V00336, X02627, X04585, M24839, X67579, AJ251080 and M25591.

2.4 Evaluation of Models

With the base model determined for the corresponding dataset, we evaluate the model on the basis of its performance on remaining dataset. The standard approach to evaluating regression model performance is through additive, residual based loss function. One test procedure is by calculating the standard error of estimate. However, since we want the scoring function to rank the motifs in decreasing order of the measure used, we used a rank based evaluation as proposed by [21].

Rank based evaluation offer advantages over residue based evaluation by being robust to outliers and providing insights about the local performance of the model. They also provide statistics that are similar to the commonly used correlation coefficients in data analysis. Value of $\hat{\tau}$ and $\hat{\rho}$ range from -1 to +1, where +1 corresponds to a perfect model performance and -1 to making all possible errors.

3 Results

Different parameters were used in Seed to generate suitable stem-loop structures for each dataset. The search space contained motifs of varying degree of PPV and sensitivity. Table I shows the number of motifs and match made by Seed for each data set.

3.1 Base Models

As discussed earlier, the selection strategy for choosing the best model was due to best subset regression. In this we identify the best-fitting models that can be constructed with the specified predictors variables. We restrict our results to models having the dependent variable as MCC due to space limitations.

Analysis of the R^2 and C_m statistic yielded in the following models shown in Table II.

TABLE II

Regression Models

$$\begin{aligned}
Y_{MCC}^{HSL3} &= 0.09 + 203\%GC + 12.1 N_2\,TInfo - 125 N_{21}\,TBest \\
&\quad -163 N_{21}\,TWorst + 230 N_{21}\,TSum \\
Y_{MCC}^{IRE} &= -110 + 308\%GC + 8.23\,NTInfo + 0.93 N_{11}\,TBest \\
&\quad -0.71 N_{11}\,TWorst - 3.42 N_{11}\,TSum \\
Y_{MCC}^{tRNA} &= 25.5 - 74.9\%GC + 0.809\,TInfo - 1.28 N_{11}\,TBest \\
&\quad +1.33 N_{11}\,TWorst - 2.62 N_{11}\,TSum \\
Y_{MCC}^{5S} &= 14.3 + 22.1\%GC + 0.212\,TInfo - 2.66 N_{11}\,TBest \\
&\quad +2.74 N_{11}\,TWorst - 1.45 N_{11}\,TSum
\end{aligned}$$

3.2 Visualization of Ranking Performance

Visualization of ranking performance can often provide additional insights about the model performance. Figure 1 shows the percentage of correct ranked pairs as a function of rank as described in [21]. The ranking statistics are listed in Table IV. The area under the curve corresponds to $\tilde{\rho}$, where $\tilde{\rho}$ is the re-normalized version of $\hat{\rho}$ ($\tilde{\rho} = 1/2 + \hat{\rho}/2$), and the area above the curve represents the sum of incorrectly ranked pairs.

3.3 Identification of Native Folds

Tables III show the sensitivity, PPV, MCC of best motifs predicted by the different models built on MCC. We can see that the PPV/sensitivity of the predicted structures of HSL3 data by different models is often 100%. The $HSL3_{MCC}$ model (regression model built on HSL3 data set) performed well on IRE, tRNA and itself. For IRE data set, the top ranked motif had a sensitivity of 92.7% and a PPV of 100%. There is a slight drop in the performance on tRNA data set resulting in a PPV of 88.2% and sensitivity 73.7%. For 5S data set, the model identifies a motif with PPV/sensitivity of 41.2/18.1%. IRE_{MCC} performed well on all the data sets. The top motif picked from 5S has an PPV and sensitivity of 72.5% and 30.1% respectively. For the remaining motifs, this model has the same performance as $HSL3_{MCC}$.

To study the performance of models built on higher complexity data set, we built models on 7 tRNA and 7 5S sequences. We can see that $tRNA_{MCC}$ performs well on all the data sets (Table III). The PPV of the top tRNA motif increased to 96% as compared to 88.2% ranked by IRE_{MCC}. However, the sensitivity drops to 71.2%. Excluding the results on the base data set, $tRNA_{MCC}$ has the highest average sensitivity (74.26%) over the other models. Figure 2 illustrates these result for two of the four experiments. Finally, models built on 5S data set appear to perform well on all the 4 data sets. It is interesting to see that all the models on 5S have the highest average PPV over other models. This suggests that models built on complex structures and large number of examples (motif) are capable to detect motifs of higher PPV.

TABLE III
Performance of MCC models

	Dataset	MOTIF#	AVGSEN	AVGPPV	AVGMCC
HSL3	*HSL3*	261	100	100	100
	IRE	85	92.7	100	96.28
	tRNA	569	73.7	88.2	80.62
	5S	38708	18.1	41.2	27.3
IRE	*HSL3*	261	100	100	100
	IRE	85	92.7	100	96.28
	tRNA	569	73.7	88.2	80.62
	5S	39929	30.1	72.5	46.71
tRNA	*HSL3*	261	100	100	100
	IRE	85	92.7	100	96.28
	tRNA	4984	71.2	96	82.67
	5S	39929	30.1	72.5	46.71
5S	*HSL3*	356	83.3	100	91.26
	IRE	109	58.9	100	76.74
	tRNA	4968	76.2	100	87.29
	5S	39929	30.1	72.5	46.71

TABLE IV
Ranking statistics : AVGMCC

	HSL3		IRE		tRNA		5S	
	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$
HSL3	0.802	0.907	0.612	0.719	0.607	0.794	0.381	0.535
IRE	0.760	0.885	0.614	0.705	0.608	0.799	0.410	0.571
tRNA	0.793	0.905	0.577	0.675	0.718	0.887	0.551	0.736
5S	0.779	0.888	0.511	0.596	0.684	0.862	0.575	0.761

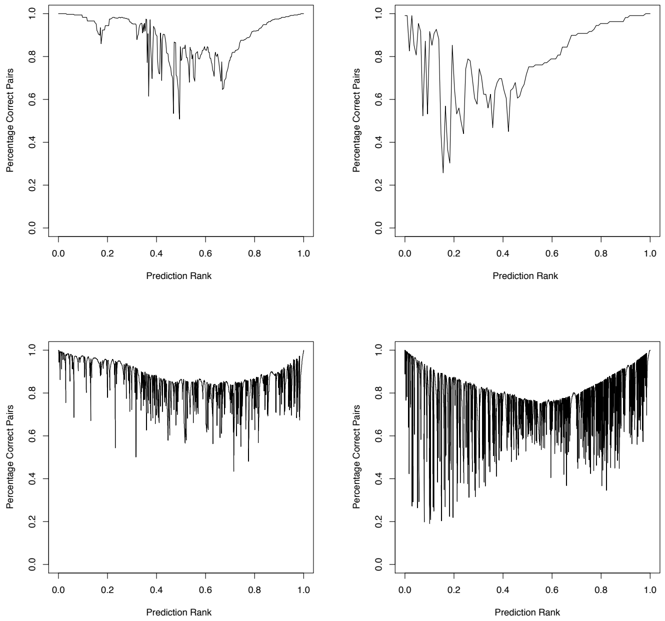

Figure 1. Ranking Performance. Starting from top left in clockwise direction: Performance of $tRNA_{MCC}$ regression model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA data sets

4 Discussion and Conclusion

In this work, we presented a scoring method for the software system Seed. The purpose was to distinguish the biologically relevant RNA secondary structure motifs from the rest. We evaluated our method on four different datasets having varying range of complexity. Two datasets we constructed consisted of selected members of UTRdb database where the coding region are flanked by two untranslated regions (5'UTR and 3'UTR). Others were assembled using a subset of sequences from [10].

We found that the method was able to identify high PPV motifs. We also found that the performance of models based on the PPV and MCC were better than Distance; MCC being the best among the three. The presence of outliers had a considerable effect on the residual based measures and made them unreliable and inappropriate for model comparison. On the other hand, the ranking based evaluation were much more stable and robust to outliers, allowing a confident model comparison. We were able to detect greater performance by models that were build using more number of examples. The performance of models build on less complex structures generally showed inconsistent results on predicting complex structures.

With the visualization of ranking performance, we were able to see that certain examples were out of place. Analyzing the common characteristics of these examples could lead to conclusions about the shortcomings of the models and possible ways to improve its ranking performance. It is likely that some moderate refinements to regression models would improve its PPV without altering its efficiency, for example, taking into account insertions/deletions (having a penalizing factor) and addition of other variants of thermodynamic scores.

Seed produces an enormous amount of consensus secondary structures that are structural, partially or fully instantiated (conserved pairs). Results indicate that a structure with high degree of specificity (information content) and low energy scores for all functions *TSum*, *TBest* and *TWorst* is deemed significantly stable. Regression based models were effective to associate the variation of scores obtained from different functions and the different performance measures.

We have introduced a scoring function formulation, implemented on the software Seed, designed to pick the best prediction(s) of RNA secondary structure motifs. The advantage of scoring functions is that they give us an intuitive mean by which to compare different possible configurations of motif structure and locations. This general approach of using regression models built on multiple functions to obtain different estimates of an RNA secondary structure can be useful models for motif discovery.

Acknowledgments

This work has been supported by the Natural Sciences and Engineering Research Council of Canada as well the Canadian Fund for Innovation.

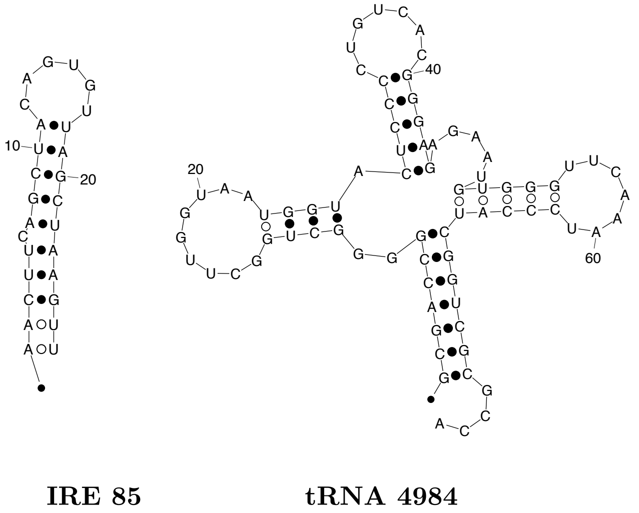

Figure 2. Secondary structure diagrams for the top ranked motifs according to $tRNA_{MCC}$ regression for the IRE dataset (left) and the tRNA dataset (right). Correctly predicted base pairs are shown with filled circles while missing base pairs (false negative predictions) are shown with open circles.

References

[1] D. P. Bartel, "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function," *Cell*, vol. 116, pp. 281–297, 2004.

[2] F. Mignoe, C. Gissi, S. Liuni, and G. Pesole, "Untranslated Regions of mRNAs," *Genome Biology*, vol. 3, no. 3, pp. 0004.1–0004.10, 2003.

[3] P. N. Borer, B. Dengler, I. Tinoco, Jr., and O. C. Uhlenbeck, "Stability of Ribonucleic Acid Double-Stranded Helices," *J. Mol. Biol.*, vol. 86, pp. 843–853, 1974.

[4] K. J. Doshi, J. J. Cannone, C. W. Cobaugh, and R. R. Gutell, "Evaluation of the Suitability of Free-Energy Minimization Using Nearest-Neighbor Energy Parameters for RNA Secondary Structure Prediction," *BMC Bioinformatics*, vol. 5, pp. 105, 2004.

[5] T. Nguyen and M. Turcotte, "Exploring the Space of RNA Secondary Structure Motifs Using Suffix Arrays," *6th International Symposium on Computational Biology and Genome Informatics (CBGI 2005)*, pp. 1291–1298, 2005.

[6] M. Anwar, T. Nguyen, and M. Turcotte, "Identification of Consensus RNA Secondary Structures Using Suffix Arrays," unpublished, 2006.

[7] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch, "Replacing Suffix Trees with Enhanced Suffix Arrays," *Journal of Discrete Algorithms*, vol. 2, no. 1, pp. 53–86, 2004.

[8] D. H. Mathews and D. H. Turner, "Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences," *J. Mol. Biol.*, vol. 317, pp. 191–203, 2002.

[9] B. Masoumi and M. Turcotte, "Simultaneous Alignment and Structure Prediction of RNAs: Are Three Input Sequences Better than Two?," in *2005 International Conference on Computational Science (ICCS 2005)*, S. V. Sunderam et al., Ed., Atlanta, USA, May 22-25 2005, Lecture Notes in Computer Science 3515, pp. 936–943.

[10] B. Masoumi and M. Turcotte, "Simultaneous Alignment and Structure Prediction of Three RNA Sequences," *International Journal of Bioinformatics Research and Applications*, vol. 1, no. 2, pp. 230–245, 2005.

[11] P. P Gardner E. Freyhult and V. Moulton, "A Comparison of RNA Folding Measures," *BMC Bioinformatics*, vol. 6, no. 4, 2005.

[12] J Gorodkin, S. L. Stricklin, and G. D. Stormo, "Discovering Common Stem-Loop Motifs in Unaligned RNA Sequences," *Nucl. Acids Res.*, vol. 29, no. 10, pp. 2135–2144, 2001.

[13] I. L. Hofacker, "Vienna RNA Secondary Structure Server," *Nucl. Acids Res.*, vol. 31, pp. 3429–3431, 2003.

[14] G. Pesole et al., "UTRdb and UTRsite: Specialized Databases of Sequences and Functional Elements of 5' and 3' Untranslated Regions of Eukaryotic mRNAs. Update 2002," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 335–340, 2002.

[15] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure," *J. Mol. Biol.*, vol. 288, pp. 911–940, 1999.

[16] M. Sprinzl and K. S. Vassilenko, "Compilation of tRNA Sequences and Sequences of tRNA Genes," *Nucl. Acids Res.*, vol. 33, no. suppl_1, pp. D139–140, 2005.

[17] M. Sprinzl and K. S. Vassilenko, "Compilation of tRNA Sequences and Sequences of tRNA Genes," http://www.uni-bayreuth.de/departments/biochemie-/trna, September 2004.

[18] R. R. Gutell, "Comparative RNA Web Site," http://www.rna.icmb.utexas.edu, July 2004.

[19] J. J. Cannone et al., "The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs," *BMC Bioinformatics*, vol. 3, no. 2, 2002.

[20] J. J. Cannone et al., "The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs: Corrections," *BMC Bioinformatics*, vol. 3, no. 15, 2002.

[21] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-Based Evaluation of Regression Models," in *The Fifth IEEE International Conference on Data Mining (ICDM '05)*, Houston, Texas, 2005, pp. 370–377.