Extensive Assessment of Metrics on RNA Secondary Structures and Relative Ensembles

Marco Barsacchi
University of Pisa
Dept. of Information Engineering
largo L. Lazzarino, 56122
Pisa, IT

Andrea Baù
University of Pisa
Dept. of Information Engineering
largo L. Lazzarino, 56122
Pisa, IT

Alessio Bechini*
University of Pisa
Dept. of Information Engineering
largo L. Lazzarino, 56122
Pisa, IT
a.bechini@ing.unipi.it

ABSTRACT

The increasing interest in non-coding RNAs (ncRNAs) and their functions pushed towards the development of analysis techniques to get as much information as possible from plain RNA sequences. Thus, being able to properly compare RNA secondary conformations is of prominent importance in any structural investigation. Several different metrics have been proposed to catch topological dissimilarities in RNA secondary structures, but so far their specific features have not been assessed yet against extensive datasets. Such a characterization is also crucial in analyzing structural ensembles, because results strictly depend on the specific distance used. The current availability of large ncRNA databases has made it possible an extensive comparison of different metrics, across both intra- and inter-ensemble structures. Correlation analysis has uncovered the relative descriptive power of such metrics, providing indications on their possible practical use in different contexts.

CCS Concepts

•**Applied computing** → **Bioinformatics;** *Molecular sequence analysis; Molecular structural biology;*

Keywords

RNA secondary structure; RNA metrics; RNA structures ensembles; base-pair distance; correlation analysis

1. INTRODUCTION

The discovery of the RNA capability to support a variety of functions in cells has changed the traditional view of RNA as a passive medium between DNA and proteins, vitalizing research around non-coding RNA (ncRNA for short). Nowadays, ncRNAs are known to be involved in a plethora of biological mechanisms, from gene regulation to protein synthesis and chromosome structures, and experimental validations are growing year after year. Recent research has shown a positive correlation between an increased proportion of non-coding versus coding RNA and an organism's developmental complexity [11]. Much of the ncRNA functions depends on its structure.

Several different representations have been proposed for RNA secondary structures; however, the most popular belong to a few different families [7]. In the *dot-racket representation* S is represented by a string of length $|x|$, where each $(i, j) \in S$ correspond to a '`(`' in position i and a '`)`' in position j. One dot '`.`' is used for an unpaired base. In *circle representation* each base is represented by a dot on the circumference of a circle of arbitrary size. Lines are drawn to connect the paired bases. A similar graphical representation keep the "base" dots along a line, with arcs connecting paired bases. *Tree representation* is mainly aimed at comparing secondary structures [10]. Multiple different tree representations can be possibly defined to account for the hierarchical arrangements of secondary features. Finally, in *mountain representation* each base pair is represented by a horizontal line over the primary sequence at a height that is dictated by its position in the sequence.

The prediction of RNA secondary structures from plain sequences has been widely investigated, and currently several approaches and tools try to approximate real functional structures. The three-dimensional RNA conformation directly relates to the function, but it can be hardly predicted in a direct way. As the RNA folding process is thought to be hierarchical [12], the secondary structure prediction represents a fundamental step towards the tertiary folding. Under the usual thermodynamic hypothesis, a native conformation corresponds to the structure with the Minimum Free Energy (MFE) [5].

The energy of a structure must be carefully evaluated, and in the present work it has been computed according to the Nearest Neighbour Thermodynamic Model (NNTM), tuned on a set of experimentally measured parameters [5]. Moreover, a broad variety of methods for predicting an ensemble

*Corresponding author

SAC'16, April 4-8, 2016, Pisa, Italy

http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

of structures admissible for a given sequence exists. Stochastic sampling based on a Boltzmann statistical model is the most common procedure [2] for ensemble sampling. Nonetheless, apart from considering the energy of each conformation, a precise characterization of members in an ensemble of RNA secondary structures asks for the identification of concise indicators that would summarize specific topological similarities (or dissimilarities) across members: such indexes may be then used as ensemble "features". The specific choice for pairwise distances across the ensemble structures influences the information content in the ensemble itself, influencing in turn the outcomes of procedures (like clustering) aimed at extracting it. However, even if base-pair distance metrics has a predominant position in the literature in comparing secondary structures, given its inherent simplicity and its low computational weight, several other metrics have been proposed.

In this work, we consider five different possible metrics (Table 1) and we computationally investigate how they behave in describing dissimilarities between structures. The study refers to a very large set of ncRNA sequences and, to the best of our knowledge, such an extensive comparison has never been done before. Correlations have been uncovered, assessing the statistical relationships between metrics in different settings.

2. METRICS FOR RNA SECONDARY STRUCTURES

An RNA strand can be represented as a string over an alphabet of nucleotides {A, C, G, U}. In an RNA sequence x, the base at position i is indicated as x_i. Bases in a strand tend to form *pairs*, and one base can either participate at most in one pair, or be unpaired. An RNA secondary structure S over x is the set of all the relative base pairs, each indicated by the two indexes $(i, j), i < j$ of the composing bases, such that: 1) the bases in the pair $(i, j) \in S$, i.e. (x_i, x_j), form either a Watson-Crick (AU,CG) or a Wobble (GU) base-pair; 2) every position i can appear at most once across all index pairs in S, 3) for every $(i, j) \in S$, $|j - i| > 3$ must hold (just for steric constraints in hairpin loops), and, if pseudoknots are not considered, 4) if $(i, j) \in S$ and $(k, l) \in S$ and $i < k$, then either $i < j < k < l$ or $i < k < l < j$ holds.

Given the importance of measuring dissimilarity between RNA secondary structures, diverse metrics for comparison have been proposed since the emergence of RNA secondary structure prediction. The base-pair (BP) distance was the first metric to be proposed and, likely, at present it is still the most used. It is plainly defined as the total number of base pairs that occur in one structure, but not in the other (Equation 1) [1].

$$d_{BP}(S_1, S_2) = \left|S_1 \bigcup S_2\right| - \left|S_1 \bigcap S_2\right| \tag{1}$$

Even f it is computationally convenient, this coarse metric does not capture much of the structural information. Several modifications have been proposed measuring maximal distance between any two base pairs in a pair of secondary structures. For each base pair $b_1(i, j) \in S_1$ and $b_2(i', j') \in S_2$ the distance between b_1 and b_2 is defined

$$d_0((i, j), (i', j')) := max\left[|i - i'|, |j - j'|\right] \tag{2}$$

Then $d_Z(S_1, S_2)$ between the two structures is defined to be the smallest d such that for every $b_1 \in S_1$ there is a base pair in S_2 within distance d_0 at most d of b_1, and (to ensure symmetry) for every base pair $b_2 \in S_2$ there is a base pair in S_1 within distance d_0 at most d of b_2 [7]. The relaxed base-pair (RBP) score has been proposed to overcome the discriminative limitation inherent to the BP score [1]. It is based upon a relaxation parameter $t \geq 0$ that indicate the degree of relaxation,

$$d_{RBP}(S_1, S2) = min\left\{m \in \mathbb{Z} \mid m \geqslant 0, \Delta_k \leq tm \text{ if } k > m\right\} \tag{3}$$

where k iterate on Δ_k that is the vector of sorted distances between each base pair of S_1 and the nearer base pair of S_2 and between each base pair of S_2 and the nearer base pair of S_1. Even if several solutions have been proposed, the choice of the correct value for t, as far as we know, still remains an open problem.

Table 1: Metrics for RNA secondary structures

Metric:	ID:	References:
BP distance	BP	[1]
Mountain distance	MD	[7]
String edit distance	SE	[3]
Tree edit distance	TE	[4, 10]
Coarse tree edit distance	HTE	[4]

Mountain metrics are based on "mountain representation" of RNA secondary structures, and are quick to compute and easy to handle from a theoretical point of view [7]. Considering a structure S of length n, let f_S be a vector of length n, where $f_S(i)$ is equal to the number of ')' minus the number of '(' found in the range $[i, n]$. Then d^p_M is defined as:

$$d^p_M(S_1, S_2) = \left(\sum_{i=1}^{n} |f_{S_1}(i) - f_{S_2}(i)|^p\right)^{\frac{1}{p}} \tag{4}$$

Tree edit distances are built upon an ordered rooted tree representation of secondary structures. In a nutshell, a tree can be transformed into another by a series of defined elementary operations (i.e, insertion, deletion, relabeling, although some metrics define case-apt edit operations), each with an associated cost. The minimum cost required to turn one tree into another defines the tree edit distance between the two. Among many possible tree distances, we have evaluated both the full structure based and the coarse grained based tree edit distances, as implemented in RNAdistance from the ViennaRNA package [4]. Furthermore, a string-edit distance can be applied both to complete and coarse-grained representations.
All the metrics chosen for cross-evaluation are listed in Table 1.

3. METHODS

First, a set of sequences has been selected; for each sequence structural ensembles have been produced and metrics evaluated on the ensembles. Then results have been analyzed by means of correlation coefficient estimation and PCA. Finally, for certain sequences, ensembles have been thoroughly studied using diverse clustering procedures.

3.1 Dataset

A set of 510,055 RNA functional sequences has been gathered from fRNAdb[1] [6], a comprehensive database of non-coding RNA (ncRNA) sequences. Among these, excessively short sequences yield a very small number of different structures, while ensembles of excessively long ones do not exhibit consistent diversity, with our current methods. So, only sequences in the range 50-250 nt have been considered, making a total of 138,903 sequences. To the best of our knowledge, this is the largest set ever used for tests of this kind. Notably, ncRNAs have been considered because, among other RNA categories, they more likely rely on structure to perform their function.

Furthermore, as non structural RNA sequence database, we have used a pool of random generated sequences in the same length range. The current sets selection have been guided by the aim of studying intra-ensemble relation between structures, i.e. studying ensemble of similar structures from the same sequence. To extend our study, however, a third set has been generated comprising structures randomly generated from random sequences of the same length, to asses how metrics behave and correlate when comparing structure that are not related. Finally, particular sequences that show clusters in ensemble, due to different *in-vivo* conformation, such as riboswitches, have been selected from previous studies [8] and an ensemble test has been executed.

3.2 Generation of Ensembles

For each sequence in the dataset an ensemble of $N = 20$ Boltzmann sampled structures has been generated with RNAsubopt from the ViennaRNA package [4]. For each metric, we have computed the $N(N-1)/2$ pairwise distances between all couples of samples. For $N = 20$ we have evaluated 190 distances per metric per structure. Reduced ensemble size are due to limited computational resources. However, in evaluating how ensemble characterisation is outlined by diverse metrics, we have generated bigger ensembles of $N = 1000$ structures, roughly corresponding to 450,000 distances per metric. This has been done for a limited number of structures for which diverse *in-vivo* clusters of structures are known.

3.3 Distance and Correlation Evaluation

Distances between structures have been evaluated using the metrics defined in Table 1 and described in Section 2. As a preliminary step, considering the ensembles for all the dataset entries, the pairwise distances have been computed, according to each target metric. Moreover, using the same method, all the pairwise distances have been computed for the random sequence dataset as well. Finally, even if our work was primarily directed at assessing intra-ensemble performances, cross-sequences analyses have been performed, on a set of random structures from random sequences described before.

3.4 Clustering Procedures

Clustering has been executed using *Partition Around Medoids* (PAM), a well-known k-medoids algorithm, which takes all the pairwise distances as input. This algorithm implements a partition clustering method, based on the definition of k *representative* objects, called *medoids*. Results have been compared with a k-means clustering using both energy of structures and bp-distance w.r.t. the MFE. Clustering consistency has been measured by means of silhouette score (Sil) [9].

3.5 Decomposition

The target metrics for this work have been thoroughly analysed by means of different approaches. In particular, PCA analysis has been used to spot out redundancy in the set of metrics, and possibly uncover what subset of metrics better describes the structural diversity in an ensemble of structures originated from a single sequence. To make comparable the results from sequences of different lengths, distances have been normalized before analysis.

4. EXPERIMENTAL RESULTS

The analysis performed in our work has been roughly divided into two phases: single sequences analysis and ensemble analysis. The methods are further described in Section 3.

4.1 Single Sequences Analysis

In the first place we have analysed how diverse RNA secondary structure metrics correlate with each other. On the matrix of data produced for each metric we have studied correlation between corresponding data series.In this phase of the study, while using structures produced in ensembles for each sequence, we have not studied ensemble distribution. Results can be compared with those generated from a pool of random sequences, as shown in the middle third of the Table 2. Moreover, in the first sets we have evaluated correlation in the set of shorter sequences, as shown between parentheses in the upper third of Table 2. Finally, on a third set of random sequences, distances have been measured between structures generated from different sequences.

4.1.1 Correlation estimation

Correlations between metrics have been evaluated on the dataset. Results are outlined in Table 2. Our results suggest a high degree of correlation between metrics. According to our results, string-edit distance and tree-edit distance on the full structure tree representation, correlate almost perfectly. Furthermore, even generating random sequences, and evaluating correlation on that random dataset, values seems not to vary (Middle part of Table 2). An important consideration is related to the fact that sequence length seems to impact correlation between metrics; in the upper third of Table 2 results related to limit our correlation analysis on sequences shorter than $100nt$ are shown between parenthesis. Correlation coefficients decrease as sequence length decreases. So far we have studied inter-ensemble distances, i.e. distances between structures extracted from the same sampled ensemble for a particular sequence. What happens, studying distances between different structures from ran-

[1] http://www.ncrna.org/frnadb/

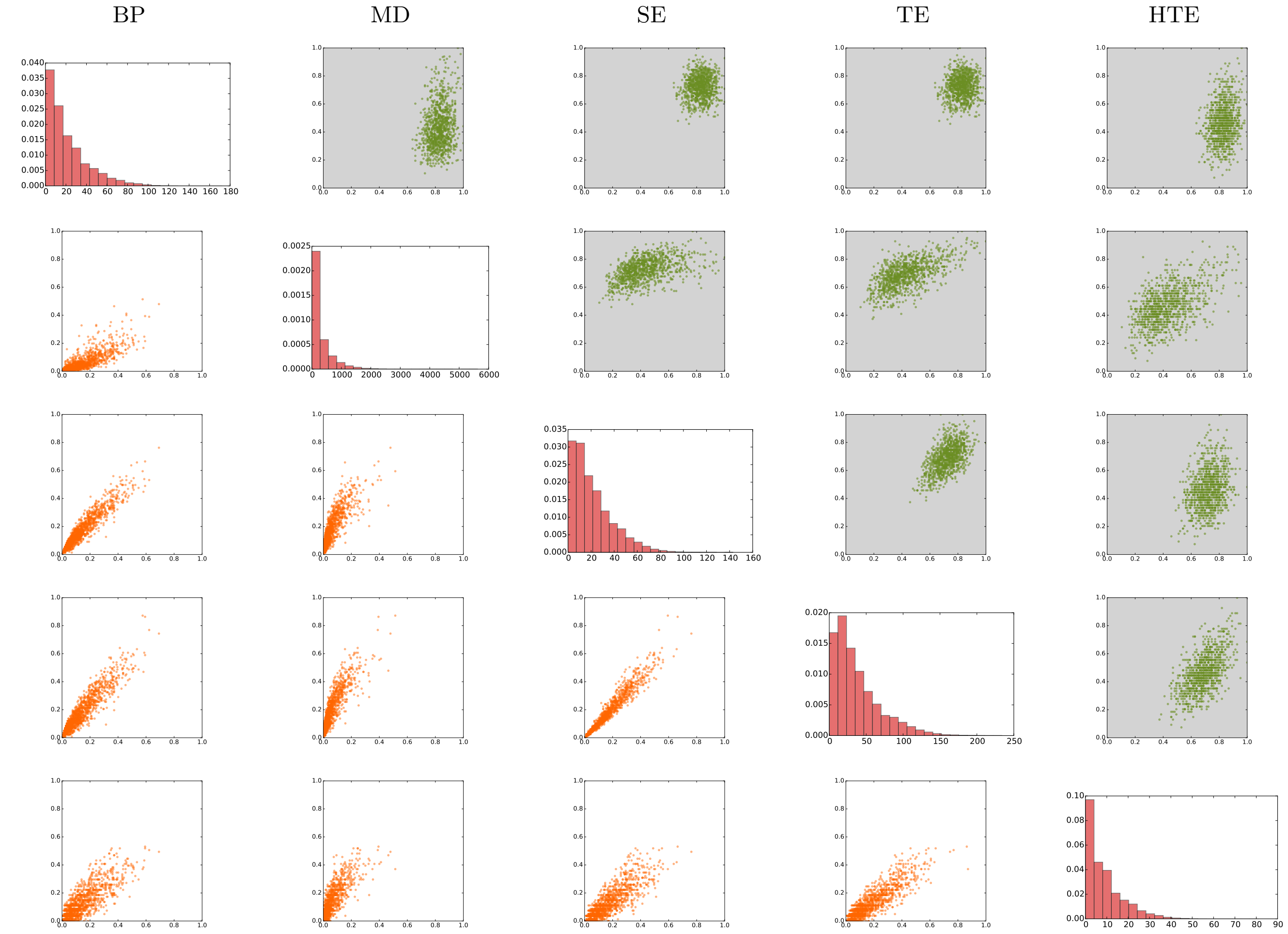

Figure 1: Outline of correlation between target metrics

In plot captions, target metrics are indicated by their IDs, as reported in Table 1. Charts placed in the lower-left triangular portion are scatter plots of pairwise distances for intra-ensemble structures (i.e. internal to the same Boltzmann-sampled ensembles), on a random subset of all the sequences. On the contrary, scatter plots in the upper-right part refer to pairwise distances for randomly-chosen inter-ensemble structures of equal length. On the diagonal, the overall distributions of inter-ensemble distances are shown.

domly generated diverse sequences, i.e. not comparing only structures from the same sequences, is that the correlation coefficients decrease significantly (Lower third of Table 2). Correlations between various metrics on our dataset have been graphically shown, as scatter plots, in Figure 1.

4.1.2 PCA

PCA has been executed on the dataset, aiming at finding redundancy in the distances analysed here. Our results point out that a great degree of redundancy exists, at least for intra-ensemble comparison. Results are outlined in Table 3. According to what we have computed, in the first case, one single component greatly explains the majority of the variance. This component seems to be made in equal part from the whole set of measures. This suggests equal importance for each measure. This result struggles to be confirmed when the degree of correlation decreases, i.e. in measuring distances between structures produced by diverse random sequences (Table 4). In the worst case, when the first component struggles to explain a little more than 50% of the variance, its composition is roughly the same as before, even if base-pair distance explains as itself roughly the 20% of the variance.

Table 3: PCA Components and Variance Explained

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Var. Expl.	90.48%	4.41%	3.56%	1.18%	0.37%

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
BP	0.44	-0.52	0.14	**-0.70**	-0.12
MD	0.42	**0.70**	0.54	-0.15	0.06
SE	0.45	-0.34	0.11	0.44	0.68
TE	**0.46**	-0.12	0.02	0.51	**-0.71**
HTE	0.43	0.33	**-0.82**	-0.14	0.10

Table 2: Correlation Between Metrics

fRNAdb Sequences [in parentheses, limited to seqs <100 nt]					
	BP	**MD**	**SE**	**TE**	**HTE**
BP	1				
MD	0.811 (0.752)	1			
SE	0.952 (0.917)	0.847 (0.835)	1		
TE	0.934 (0.880)	0.878 (0.877)	0.976 (0.963)	1	
HTE	0.829 (0.702)	0.820 (0.749)	0.859 (0.761)	0.894 (0.811)	1
Random Sequences					
BP	1				
MD	0.822	1			
SE	0.960	0.853	1		
TE	0.943	0.879	0.975	1	
HTE	0.852	0.844	0.879	0.913	1
Random Sequences, Different Structures					
BP	1				
MD	0.210	1			
SE	0.134	0.519	1		
TE	0.133	0.607	0.590	1	
HTE	0.230	0.515	0.310	0.597	1

Table 4: PCA Components and Variance Explained Random Sequences, Different Structures

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Var. Expl.	52.06%	19.56%	13.94%	8.50%	5.94%

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
BP	-0.17	**0.94**	-0.23	0.09	-0.11
MD	-0.51	0.01	0.05	**-0.85**	0.10
SE	-0.44	-0.24	**-0.73**	0.27	0.38
TE	**-0.54**	-0.20	0.06	0.23	**-0.78**
HTE	-0.47	0.08	0.63	-0.78	0.47

4.2 Ensemble characterization

In our work we have tried to outline how diverse metrics could be used to characterise ensembles of structures producing diverse results. For a limited set of sequences we have generated bigger ensembles, as described in Section 3.1. On this ensembles pairwise distances have been computed for different metrics and have been used as inputs to diverse clustering procedures.

Supplementary Fig “blabla” shows k-medoids clustering outcome on the ensemble for a relevant sequence (Notably the “thiM_TPP riboswitch” as described in [8]). All the subplots in the figure depict structure in the ensemble, for a particular riboswitch sequence, known to present *in-vivo* bistable states, in a bp-distance w.r.t. the MFE structure vs. energy of the structures. Diverse colors outline clustering results, using diverse distance matrices as input. The latter plot shows instead a k-means clustering using both energy and base pair distance w.r.t. to the mfe as features. The latter clustering shows higher performances, as measured by silhouettes. Nonetheless both the BP and MD k-medoids clustering shows quite good silhouettes even if the classic scatter plot used in representing RNA ensemble fails to show these groups.

5. CONCLUSIONS

In the presented work, different metrics on RNA secondary structures have been evaluated against a big dataset of ncRNA sequences, focusing in particular on structures from ensembles calculated for each sequence. The first relevant result is the high degree of correlation for all the studied metrics, in case intra-ensemble distances are taken into account: in this setting, each distance value is computed between different possible structures for the same sequence. This holds both for real and for random sequences. Indeed, the correlation analysis on distances for structures randomly produced from diverse randomly generated sequences, yields far less impressing results. Although apparently obvious, the fact that distance correlations for inter-ensemble structures are basically the same for real ncRNA and random sequences can be interpreted as a characteristic of the metrics themselves, instead of being related to the specific dataset used. Moreover, further analysis shows that correlation values slightly depend on the sequence length, as shorter sequences give lower values. This may be due to an effect of the sampling procedure, which on fixed-size ensembles catches more distant conformations for smaller sequences that for longer ones. The main practical results can be itemized as follows: In measuring distances between structures in the same ensemble, all the metrics offer the same contribution in explaining variance. The high correlation between metrics indicates that they all offer a similar descriptive power. Because of its simplicity, the base-pair distance (BP) can thus be conveniently used instead of more complex ones.

Conversely, in comparing structures from different sequences, the use of at least three measures seems an advisable choice. We propose the use of mountain distance (MD) combined with base-pair (BP) and string-edit (SE) distance, to assure a proper variance explanation, yet keeping low the required computational effort.

Furthermore, our results point out that, in comparing RNA structures on the described setting, the tree representation does not necessarily provide relevant benefits over the string representation. Actually, for intra-ensemble distances TE and SE have very high correlation values (~ 0.976).

Assessing the influence of using a specific metric in providing a well-described ensemble is a complex job. In our tests, we considered diverse clustering outcomes for a particular sequence, known to be present *in-vivo* in two alternate forms. On this testbed, only three metrics lead to satisfying outcomes, with two distinct clusters ($Sil > 0.6$).

Future work will necessarily address more complex energy landscapes, aiming at finding better ways to build descriptive and representative ensembles of RNA secondary structures.

6. REFERENCES

[1] P. Agius, K. P. Bennett, and M. Zuker. Comparing RNA secondary structures using a relaxed base-pair score. *RNA*, 16(5):865–878, 2010.

[2] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

[3] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, Feb. 1966.

[4] R. Lorenz, S. Bernhart, C. Honer zu Siederdissen, H. Tafer, C. Flamm, P. Stadler, and I. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[5] 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*, 288(5):911–940, 1999.

[6] T. Mituyama, K. Yamada, E. Hattori, H. Okida, Y. Ono, G. Terai, A. Yoshizawa, T. Komori, and K. Asai. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Research*, 37(suppl 1):D89–D92, 2009.

[7] V. Moulton, M. Zuker, M. Steel, R. Pointon, and D. Penny. Metrics on RNA secondary structures. *J. Computational Biology*, 7(1-2):277–92, Jan. 2004.

[8] G. Quarta, K. Sin, and T. Schlick. Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. *PLoS computational biology*, 8(2):e1002368, Jan. 2012.

[9] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov. 1987.

[10] B. A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Computer applications in the biosciences : CABIOS*, 4(3):387–393, 1988.

[11] R. Taft and J. Mattick. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology*, 5(1):P1, 2003.

[12] I. Tinoco Jr and C. Bustamante. How RNA folds. *J Mol Biol*, 293(2):271–281, 1999.