Stable Stem Enabled Shannon Entropies Distinguish Non-coding RNAs From Random Backgrounds

Yingfeng Wang  
Department of Computer Science  
University of Georgia  
Athens, GA 30602, USA  
Email: ywang802@uga.edu

Amir Manzour  
Institute of Bioinformatics  
University of Georgia  
Athens, GA 30602, USA  
Email: amanzour@uga.edu

Pooya Shareghi  
Department of Computer Science  
Institute of Bioinformatics  
University of Georgia  
Athens, GA 30602, USA  
Email: pooya@uga.edu

Timothy I. Shaw  
Department of Computer Science  
Institute of Bioinformatics  
University of Georgia  
Athens, GA 30602, USA  
Email: gatech@uga.edu

Ying-Wai Li  
Center for Simulational Physics  
University of Georgia  
Athens, GA 30602, USA  
Email: ywli@uga.edu

Russell L. Malmberg  
Department of Plant Biology  
University of Georgia  
Athens, GA 30602, USA  
Email: russell@plantbio.uga.edu

Liming Cai  
Department of Computer Science  
University of Georgia  
Athens, GA 30602, USA  
Email: cai@cs.uga.edu

Abstract—The computational identification of RNAs in genomic sequences requires the identification of signals of RNA sequences. Shannon base pairing entropy is an indicator for RNA secondary structure folding certainty, in the detection of structural non-coding RNAs (ncRNAs). Under the Boltzmann ensemble of secondary structures, the probability of a base pair is estimated from its frequency across all the alternative equilibrium structures. However, such an entropy has yet to deliver the desired performance distinguishing ncRNAs from random sequences. Developing novel methods to improve the entropy measurement may result in a more effective ncRNA gene finding based on structure detection.

This paper shows that the measuring performance of base pair entropy can be significantly improved with a constrained secondary structure ensemble in which only canonical base pairs are assumed to occur, and energetically stable stems are required, in a fold. This constraint actually reduces the space of the secondary structure and may lower probabilities of base pairs unfavorable to the native fold. Indeed, base pair entropies computed with this constrained model demonstrate substantially narrowed gaps of Z-scores between ncRNAs as well as drastic increases in the Z-score for all 13 tested ncRNA sets compared to shuffled sequences.

Keywords—RNA secondary structure, Shannon entropy, Boltzmann ensemble, base pair, base pair probability, stable stem, stochastic context-free grammar, Z-score

I. INTRODUCTION

Statistical signals in primary sequences for non-coding RNA (ncRNA) genes have been evasive [7], [29], [10]. Because single strand RNA folds into a structure, the most exploitable feature for structural ncRNA gene finding has been the secondary structure [8], [37], [22]. The possibility that the folded secondary structure may lead to successful ab initio ncRNA gene prediction methods has energized leading groups to independently develop structure-based ncRNA gene finding methods [39], [28]. The core of such a program is its secondary structure prediction mechanism, for example, based on computing the minimum free energy for the query sequence under some thermodynamic energy model [34], [35], [40], [13]. The hypothesis is that the ncRNAs’ secondary structure is thermodynamically stable. Nonetheless, the stability measures have not performed as well as one might hope [26]; there is evidence that the measures may not be effective on all categories of ncRNAs [2].

A predicted secondary structure can be characterized for its fold certainty with the Shannon base pair entropy [23], [15]. The entropy \( \sum p_{i,j} \log p_{i,j} \) of base pairings between all bases \( i \) and \( j \) can be calculated based on the partition functions for the Boltzmann’s secondary structure ensemble; the probability \( p_{i,j} \) is calculated as the total of Boltzmann’s factors over all equilibrium alternative structures that contain the base pair \((i,j)\) [25]. The entropy measure has been scrutinized with real ncRNA data revealing a strong correlation between entropy and free energy [9], [3]. However, there has been mixed success in discerning structural ncRNAs from their randomly shuffled counterparts. Both measures perform impressively on precursor miRNAs but not as well on trRNAs and some rRNAs [9], [2].

The diverse results of the entropy measuring on different ncRNAs possibly suggest that the canonical RNA secondary structure ensemble has yet to capture all ncRNAs structural characteristics. For example, a Boltzmann ensemble enhanced with weighted equilibrium alternative structures has also resulted in higher accuracy in secondary structure prediction [3]. There is strong evidence that the thermodynamic energy model can improve its structure prediction accuracy by considering energy contributions in addition to those from the canonical free energy model [38], [36]. Therefore, developing ncRNA structure models that can
effectively account for critical structural characteristics may become necessary for accurate measurement of RNA fold certainty.

This paper presents a method that computes Shannon base pair entropies based on a constrained secondary structure model. The results show substantial improvements in the Z-score of base pairing Shannon entropies on 13 ncRNA datasets [9] over the Z-score of entropies computed by existing software (e.g., NUPACK [4]). The constraint in our work requires only canonical base pairs to occur in stable stems. The constrained secondary structure model is defined with a stochastic context-free grammar (SCFG) and entropies are computed with the Inside and Outside algorithms. Our results suggest that incorporating more constraints may further improve the effectiveness of the fold certainty measure, offering improved ab initio ncRNA gene finding.

II. Method AND Model

Our method to distinguish ncRNAs from random sequences is based on the measuring of the base pairing Shannon entropy [23], [15] under a new RNA secondary structure model. The building blocks of this model are stems arranged in parallel and nested patterns connected by unpaired strand segments, similar to those permitted by a standard ensemble [25], [40], [12]. The new model is constrained, however, to contain a smaller space of equilibrium alternative structures, requiring only energetically stable stems (e.g., of free energy levels under a threshold) to occur in the structures. The constraint is basically to consider the effect of energetically stable stems on tertiary folding and to remove spurious structures that may not correspond to a tertiary fold. According to the RNA folding pathway theory and the hierarchical folding model [33], [1], [24], building block helices are first stabilized by canonical base pairings before being arranged to interact with each other or with unpaired strands through tertiary motifs (non-canonical nucleotide interactions). A typical example is the multi-loop junctions in which one or more pairs of coaxially stacked helices bring three or more regions together, further stabilized by the tertiary motifs at the junctions [21], [20]. The helices involved are stable before the junction is formed or any possible nucleotide interaction modifications are made to the helical base pairs at the junction [32].

A. Energetically stable stems

A stem is the atomic, structural unit of the new secondary structure space. To identify the energy levels of stems suitable to be included in this model, we conducted a survey on the 51 sets of ncRNA seed alignments, representatives of the ncRNAs in Rfam [11], which had been used with the software Infernal [27] as benchmarks. From each ncRNA seed structural alignment, we computed the thermodynamic free energy of every instance of a stem in the alignment data using various functions of the Vienna Package [13], [12] as follows. RNAduplex was first applied to the two strands of the stem marked by the annotation to predict the optimal base pairings within the stem, then, the minimum free energy of the predicted stem structure, with overhangs removed, was computed with RNAeval. Figure 1 shows plots of the percentages of stems at free energy levels in these 51 ncRNA seed alignments.

The peaks (with relatively high percentages) on the percentage curve of Figure 1 indicate concentrations of certain types of stems at energies levels around -4.5, -3.3, and -2.4 kcal/mol. Since a G-U pair is counted weakly toward free energy contribution (by the Vienna package), we identified the peak value -4.5 kcal/mol to be the free energy of stems of three base pairs, with two G-C pairs and one A-U in the middle or two A-U pairs and one G-C in the middle. The value -3.3 kcal/mol is the free energy of stems containing exactly two G-C pairs or stems with one G-C pair followed by two A-U pairs. Values around -2.4 kcal/mol are stems containing one G-C and an A-U pair or simply four A-U pairs.

Based on this survey, we were able to identify two energy thresholds: -3.4 and -4.6 kcal/mol for semi-stable stems and stable stems respectively. Both require at least three base pairs of which at least one is G-C pair. In conducting this survey, we did not directly use the stem structures annotated in the seed alignments to compute their energies. Due to evolution, substantial structural variation may occur across species; one stem may be present in one sequence and absent in another but a structural alignment may try to align all sequences to the consensus stem, giving rise to “misalignments” which we have observed [14]. Most of such “malformed stems” mistakenly aligned to the consensus often contain bulges or internal loops and have higher free energies greater than the threshold -3.4 kcal/mol.

B. The RNA secondary structure model

In the present study, a secondary structure model is defined with a Stochastic Context Free Grammar (SCFG) [6]. Our model requires there are at least three consecutive base pairs in every stem; the constraint is described with the following seven generic production rules:
According to the principle of maximum entropy, given we be the geometric average of the six base pair probabilities.

The starting non-terminal, \( X \), can generate an unpaired nucleotide or a base pair with the first three rules. The fourth rule generates two parallel substructures. Non-terminal \( H \) is used to generate consecutive base pairs with non-terminal \( Y \) to generate the closing base pair. Essentially, generating a stem needs to recursively call production rules with the left-hand-side non-terminals \( X \), \( H \) and \( Y \) each at least once. This constraint guarantees that every stem has at least three consecutive base pairs, as required by our secondary structure model.

### C. Probability parameter calculation

There are two sets of probability parameters associated with the induced SCFG. First, we used a simple scheme of probability settings for the unpaired bases and base pairs, with a uniform 0.25 probability for every base. The probability distribution of \( \{0.25, 0.25, 0.17, 0.17, 0.08, 0.08\} \) is given to the six canonical base pairs G-C, C-G, A-U, U-A, G-U, and U-G; a probability of zero is given to all non-canonical base pairs. Alternatively, probabilities for unpaired bases and base pairs may be estimated from available RNA datasets with known secondary structures [11], as has been done in some of the previously work with SCFGs [17], [18].

Second, we computed the probabilities for the production rules of the model as follows. To allow our method to be applicable to all structural ncRNAs, we did not estimate the probabilities based on a training data set. In fact, we believe that the probability parameter setting of an SCFG for the fold certainty measure should be different from that for fold stability measure (i.e., folding). Based on the principle of maximum entropy, we developed the following approach to calculate the probabilities for the rules in our SCFG model.

Let \( p_i \) be the probability associated with the production rule \( i \), for \( i = 1, 2, \ldots, 7 \), respectively. Since the summation of probabilities of rules with the same non-terminal on the left-hand-side is required to be 1, we can establish the following equations:

\[
\begin{align*}
    p_1 + p_2 + p_3 + p_4 &= 1 \\
    p_5 + p_6 &= 1 \\
    p_7 &= 1
\end{align*}
\]

Let

\[ q_{bp} = \sqrt[6]{0.25 \times 0.25 \times 0.17 \times 0.17 \times 0.08 \times 0.08} \]

be the geometric average of the six base pair probabilities. According to the principle of maximum entropy, given we have no prior knowledge of a probability distribution, the assumption of a distribution with the maximum entropy is the best choice, since it will take the smallest risk [16]. If we apply this principle to our problem, the probability contribution from a base pair should be close to the contribution from unpaired bases. Rule probabilities can be estimated to satisfy following equations:

\[
\begin{align*}
    p_1 &= 0.499 \\
    p_2 &= 0.499 \\
    p_3 &= 0.001 \\
    p_4 &= 0.001 \\
    p_5 &= 0.103 \\
    p_6 &= 0.897 \\
    p_7 &= 1
\end{align*}
\]

### D. Computing base pairing Shannon Entropy

Based on the new RNA secondary structure model, we can compute the fold certainty of any given RNA sequence, which is defined as the Shannon entropy measured on base pairings formed by the sequence over the specified secondary structure space \( \Omega \). Specifically, let the sequence be \( x = x_1x_2\ldots x_n \) of \( n \) nucleotides. For indexes \( i < j \), the probability \( P_{i,j} \) of base pairing between bases \( x_i \) and \( x_j \) is computed with

\[
P_{i,j}(x) = \sum_{s \in \Omega} p(s, x) \delta(x)_{i,j}^s \tag{1}
\]

where \( p(s, x) \) is the probability of \( x \) being folded into to the structure \( s \) in the space \( \Omega \) and \( \delta(x)_{i,j}^s \) is a binary value indicator for the occurrence of base pair \((x_i, x_j)\) in structure \( s \). Shannon entropy of \( P_{i,j}(x) \) is computed as [23], [15]

\[
Q(x) = -\frac{1}{n} \sum_{i < j} P_{i,j}(x) \log P_{i,j}(x) \tag{2}
\]

To compute the expected frequency of the base pairing, \( P_{i,j}(x) \) with formula (1), we take advantage of the Inside and Outside algorithms developed for SCFG [6]. Given any nonterminal symbol \( S \) in the grammar, the inside probability is defined as

\[
\alpha(S, i, j, x) = \text{Prob}(S \Rightarrow^* x_i x_{i+1} \cdots x_j)
\]

i.e., the total probability for the sequence segment \( x_i x_{i+1} \cdots x_j \) to adopt alternative substructures specified by \( S \). Assume \( S_0 \) to be the initial nonterminal symbol for the SCFG model. Then \( \alpha(S_0, 1, n, x) \) is the total probability of the sequence \( x \)’s folding under the model.

The outside probability is defined as

\[
\beta(S, i, j, x) = \text{Prob}(S_0 \Rightarrow^* x_1 \cdots x_{i-1} S x_{j+1} \cdots x_n)
\]

i.e., the total probability for the whole sequence \( x_1 \cdots x_n \) to adopt all alternative substructures that allow the sequence segment from position \( i \) to position \( j \) to adopt any substructure specified by \( S \).
Figure 2. Comparisons of the averaged Z-scores of Shannon base pair entropies computed by NUPACK and TRIPLE on each of the 13 ncRNA datasets downloaded from [9]. For each ncRNA sequence, the Z-scores were computed with respect to their 100 randomly shuffled sequences. Both single and di-nucleotide shuffling methods were used.

\[ P_{i,j}(x) \text{ then can be computed as the normalized probability of the base pair } (x_i, x_j) \text{ occurring in all valid alternative secondary structures of } x: \]

\[
\sum_{S \to aRbT} Prob(S \to aRbT, a = x_i, b = x_j) \gamma(R, S, T, i, j, x) = \frac{\alpha(S_0, 1, n, x)}{\alpha(S_0, 1, n, x)} \tag{3}
\]

where

\[
\gamma(R, S, T, i, j, x) = \sum_{j < k \leq n} \alpha(R, i + 1, j - 1, x) \times \beta(S, i, k, x) \times \alpha(T, j + 1, k, x)
\]

in which variables \( S, R, T \) are for non-terminals and variable production \( S \to aRbT \) represents rules (3)~(7) which involve base pair generations. For rules where \( T \) is empty, the summation and term \( \alpha(T, j + 1, k, x) \) do not exist and \( k \) is fixed as \( j \).

The efficiency to compute \( P_{i,j}(x) \) largely depends on computing the \textbf{Inside} and \textbf{Outside} probabilities, which can be accomplished with dynamic programming and has the time complexity \( O(nm^3) \) for a model of \( m \) nonterminals and rules and sequence length \( n \).

III. TEST RESULTS

We implemented the algorithm for Shannon base pair entropy calculation into a program named TRIPLE. We tested it on ncRNA datasets and compared its performance on these ncRNAs with the performance achieved by the software NUPACK [4] developed under the Boltzmann standard secondary structure ensemble [5], [25].

A. Data preparation

We downloaded the 13 ncRNA datasets previously investigated in Table 1 of [9]. They are of diverse functions, including pre-cursor microRNAs, group I and II introns, RNase P and MRP, bacterial and eukaryotic signal recognition particle (SRP), ribosomal RNAs, small nuclear spliceosomal RNAs, riboswitches, tmRNAs, regulatory RNAs, rRNAs, telomerase RNAs, small nucleolar RNAs, and Hammerhead ribozymes.

The results from using these datasets were analyzed with 6 different types of measures, including Z-score and \( p \)-value of minimal free energy (MFE), and Shannon base pair entropy [9], in comparisons with random sequences. While the six measures correlate to varying degrees, using MFE Z-score and Shannon base pair entropy may be sufficient to cover the other measures. However, these two measures, as the respective indicators for the fold stability and fold certainty of ncRNA secondary structure, have varying performances on the 13 ncRNA datasets.

For our tests, we also generated random sequences as control data. For every ncRNA sequence, we randomly shuffled it to produce two sets of 100 random sequences each; one set was based upon single nucleotide shuffling, the other was based upon di-nucleotide shuffling. In addition, all ncRNA sequences containing nucleotides other than A, C, G, T, and U were removed for the reason that NUPACK [4] doesn’t accept sequences containing wildcard symbols.

B. Shannon entropy distribution of random sequences

Both the software NUPACK (with the pseudoknot function turned off) and our program TRIPLE computed base pair probabilities on ncRNA sequences and random sequences. In particular, for every ncRNA sequence \( x \) and its associated randomly shuffled sequence set \( S_x \), the Shannon entropies of these sequences were computed.

A Kolmogorov-Smirnov test (KS test) [19] was applied to verify the normality of the entropy distributions from all randomly shuffled sequence sets. The results show that about 99% of the sequence sets fail to reject the hypothesis that entropies are normally distributed with 95% confidence level. This indicates that we may use a Z-score to measure performance.

C. Z-score scores and comparisons

For each ncRNA, the average and standard deviation of Shannon entropies of the randomly shuffled sequences were estimated. The Z-score of the Shannon entropy \( Q(x) \) of ncRNA sequence \( x \) is defined as follows.

\[
Z(x) = \frac{\mu(Q(S_x)) - Q(x)}{\sigma(Q(S_x))} \tag{4}
\]

where \( \mu(Q(S_x)) \) and \( \sigma(Q(S_x)) \) respectively denote the average and standard deviation of the Shannon entropies of the random sequences in set \( S_x \). The Z-Score measures how well entropies may distinguish the real ncRNA sequence \( x \) from their corresponding randomly shuffled sequences in \( S_x \).

Figure 2 compares the averages of the Z-scores of Shannon base pair entropies computed by NUPACK and by TRIPLE on each of the 13 ncRNA datasets. It shows that TRIPLE significantly improved the Z-scores over NUPACK across all the 13 datasets.

To examine how the Z-scores might have been improved by TRIPLE, we designated four thresholds for Z-scores, which are 2, 1.5, 1, and 0.5. The percentages of sequences of each dataset with Z-score greater than or equal to the thresholds were computed. Table I shows details of the Z-score improvements when di-nucleotide shuffling was used. TRIPLE
performed better than NUPACK in 13, 13, 12, and 10 datasets with thresholds of 2, 1.5, 1, and 0.5, respectively. With a single nucleotide random shuffling our method also performs better than NUPACK in the majority of datasets (table not shown). In particular, TRIPLE performed better than NUPACK in 13, 12, 12, and 9 datasets with thresholds of 2, 1.5, 1, and 0.5, respectively. In summary, our method had the much better performance in most datasets.

IV. DISCUSSION

This work introduced a modified ensemble of ncRNA secondary structures. The comparisons of performances between our program TRIPLE and NUPACK (implemented based on the canonical structure ensemble) have demonstrated a significant improvement in the entropy measure for ncRNA fold certainty by our model.

We note that there is only one exceptional case observed from Table 1: SRP whose entropy Z-score performance was not improved (as much as other ncRNAs) when $Z < 1.5$. The problem might have been caused by the implementation technique rather than the methodology. Most of the tested SRP RNA sequences (Eukaryotic and archaeal 7S RNAs) are of length around 300 and contain about a dozen stems. In many of them, consecutive base pairs are broken by internal loops into small stem pieces, some having only two consecutive canonical pairs; whereas, in our SCFG implementation we simply required three consecutive base pairs as a must in a stem, possibly missing the secondary structure of many of these sequences. This issue with the SCFG can be easily fixed, e.g., by replacing the SCFG with one that better represents the constrained Boltzmann ensemble in which stems are all energetically stable.

To ensure that the performance difference between TRIPLE and NUPACK was not due to the difference in the thermodynamic energy model (Boltzmann ensemble) and the simple statistical model (SCFG) with stacking rules, we also constructed two additional SCFG models, one for the unconstrained base pairs and another requiring at least two consecutive canonical base pairs in stems. Tests on these two models over the 13 ncRNA data set resulted in entropy Z-scores (data not shown) comparable to those obtained by NUPACK but inferior to the performance of TRIPLE. We attribute the impressive performance by TRIPLE to the constraint of “triple base pairs” satisfied by real ncRNA sequences but which are hard to achieve for random sequences.

Since the entropy $Z$-score improvement by our method was not uniform across the 13 ncRNAs, one may want to look into additional other factors that might have contributed to the under-performance of certain ncRNAs. For example, the averaged GC contents are different in these 13 datasets, with SRP RNAs having 58% GC and standard deviation of 10.4%. A sequence with a high GC content is more likely to produce more spurious, alternative structures, possibly resulting in a higher base pair entropy. However, since randomly shuffled sequences would also have the same GC content, it becomes very difficult to determine if the entropies of these sequences have been considerably affected by the GC bias. Indeed, previous investigations [30] have revealed that, while the base composition of a ncRNA is related to the phylogenetic branches on which the specific ncRNA may be placed, it may not fully explain the diverse performances of structure measures on various ncRNAs. Notably it has been discovered that base compositions are distinct in different parts of rRNA secondary structure (stems, loops, bulges, and junctions) [31], suggesting that an averaged base composition may not suitably represent the global structural behavior of an ncRNA sequence.

This constraint of stable stems, implemented by TRIPLE, was intended to capture the energetic stability of helical structures in the native tertiary fold [24], [33]. Since the ultimate distinction between a ncRNA and a random sequence lies in its function (thus tertiary structure), critical tertiary characteristics may be incorporated into the structure ensemble to further improve the fold certainty measure. Actually, ncRNA sequences from the 51 datasets demonstrated certain sequential properties that may characterize tertiary interactions (see section II. A.), e.g., coaxial stacking of helices. However, to computationally model tertiary interactions, a model beyond a context-free system would be necessary.

Although this method and technique have been developed with reference to non-coding RNAs, it is possible that
protein-coding mRNAs would display similar properties, when sufficient structural information about them has been gathered.

ACKNOWLEDGMENT

This research project was supported in part by NIH BISTI grant (No:R01GM072080-01A1), NIH ARRA Administrative Supplement to this grant, and NSF IIS grant (No:0916250).

REFERENCES

[1] Batey, R.T., Rambo, R.P., and Doudna, J.A. (1999) Tertiary motifs in RNA structure and folding, Angew. Chem. Int. Ed. 38, 2326-2343.
[2] Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, Bioinformatics 20:17:2911-7.
[3] Ding, Y. and Lawrence, CE. (2003) A statistical sampling algorithm for RNA secondary structure prediction, Nucl. Acids Res. 31:24: 7280-7301.
[4] Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E., and Pierce, N.A. (2007) Thermodynamic analysis of interacting nucleic acid strands. SIAM Rev 49: 65-88.
[5] Dirks, R.M. and Pierce, N.A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots, J. Comput. Chem. 25: 1295-1304.
[6] Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. J. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge UK.
[7] Eddy, SR. (2001) Non-coding RNA genes and the modern RNA world, Nature Reviews Genetics 2:12, 919-929.
[8] Eddy, SR. (2002) Computational genomics of noncoding RNA genes, Cell, 109:2, 137-40.
[9] Freyhult, E., Gardner, P., and Moulton, V. (2005) A comparison of RNA folding measures BMC Bioinformatics, 6: 241.
[10] Griffiths-Jones, S. (2007) Annotating noncoding RNA genes, Annu. Rev. Genomics and Human Genetics 8: 27998.
[11] Griffiths-Jones, S., Moxon, S., Marshall, M, Khana, A, Eddy, SR, and Bateman, A. (2005) Rfam: Annotating Non-Coding RNAs in Complete Genomes. Nucleic Acids Research, 33: D121-D141.
[12] Ivo L. Hofacker, (2003) Vienna RNA secondary structure server, Nucleic Acids Research 31:13, 3429-3431.
[13] Hofacker, IL., Fontana, W., Stadler, PF., Bonhoeffer, LS., tacker, M, and Schuster, P. (1994) Fast folding and comparison of RNA sequence structures, Monatsh Chem 125: 167-168.
[14] Huang, Z., Mohebbi M., Malmberg, R., and Cai, L. (2010) RNAv: Non-coding RNA secondary structure variation search via graph Homomorphism, Proceedings of Computational Systems Bioinformatics 2010: 56-68.
[15] Huynen, M., Gutell, R., and Konings, D., (1997) Assessing the reliability of RNA folding using statistical mechanics. Journal of Molecular Biology 267: 1104-1112.
[16] Jaynes, E. T. (1968) "Prior Probabilities" IEEE Transactions on Systems Science and Cybernetics 4 (3): 227241.
[17] Klein RJ and Eddy SR. (2003) RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4: 44.
[18] Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars, Nucl. Acids Res. 31 (13): 3423-3428.
[19] Kolmogorov, A. (1933) Sulla determinazione empirica di una legge di distribuzione G. Inst. Ital. Attuari, 4, 83.
[20] Laing, C. and Schlick. T. (2009) Analysis of four-way junctions in RNA structures. J. Mol. Biol, doi:10.1016/j.jmb.2009.
[21] Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs, RNA 12, 83-93.
[22] Machado-Lima A, del Portillo HA, and Durham AM. (2008) Computational methods in noncoding RNA research. Journal of Mathematical Biology 56(1-2): 15-49.
[23] Mathews, DH. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10: 1178-1190.
[24] Masquida, B. and Westhof, E. (2006) A modular and hierarchical approach for all-atom RNA modeling, in RNA World, 3rd Edition, (ed. Gesteland, Cech, and Atkins). CSHL Press.
[25] McCaskill, JS. (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. Biopolymers 29: 1105-1119.
[26] Moulton, V. (2005) Tracking down noncoding RNAs, PNAS 102:7, 2269-2270.
[27] Nawrocki EP, Kolbe DL, Eddy SR. (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25:10: 1335-1337.
[28] Pedersen, J.S., Bejerano, G., Siepel, A., Rosenberg, K., Lindblad-Toh, K., Lander, E., Rogers, J., Kent, J., Miller, W., and Haussler, D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Computational Biology 2:4, e33.
[29] Schattner, P. (2003) Computational gene-finding for non-coding RNAs in Noncoding RNAs: Molecular Biology and Molecular Medicine, Barciszewski and Erdmann, ed.
[30] Schultz, E.A., Hraber, P.T., and LaBean, T.H. (1999) Estimating the contributions of selection and self-organization in secondary structure, Journal of Molecular Evolution 49, 7683.
[31] Smit, S., Yarus, M., and Knight, B. (2006) Natural selection is not required to explain universal compositional patterns in tRNA secondary structure categories. RNA 12, 1-14.
[32] D. Thirumalai, (1998) Native secondary structure formation in RNA may be a slave to tertiary folding, Proc. Natl. Acad. Sci. USA 95, 11506-11508.
[33] Tinoco, I. and Bustamante, C. (1999) How RNA folds, Journal of Molecular Biology, 293, 271-281.
[34] Turner, DH., Sugimoto, N., Kierzek, R., and Dreiker, SD. (1987) Free energy increments for hydrogen bond in nucleic acid base pairs. Journal of Am. Chem. Soc. 109: 192.
[35] Turner, DH., Sugimoto, N., and Freier, S.M. (1988) RNA structure prediction. Ann. Rev. Biophy. Biophys. Chem. 7: 167-192.
[36] Tyagi, R. and Mathews, D.H. (2007) Predicting helical coaxial stacking in RNA multibranch loops, RNA v.13(7): 939-951.
[37] Uzilov A V, Keegan JM, Mathews DH. (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 7:173.
[38] Walter, A.E., Turner, D.H., Kim, J., Matthew, H.L., Muller, P., Mathews, D.H., and Zuker, M. (1994) Coaxial stacking of helices enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proceedings of National Academy of Sciences, USA, 91, 9218-9222.
[39] Washietl, S., Hofacker, I.L., and Stadler, P.F. (2005) Fast and reliable prediction of noncoding RNAs, Proceedings of National Academy of Sciences, USA, 102:7-22454-2459.
[40] Zuker, M. and Steigler, P. (1981) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133-148.