Published in final edited form as:

J Mol Biol. 2011 October 21; 413(2): 473–483. doi:10.1016/j.jmb.2011.08.033.

Statistical Potentials for Hairpin and Internal Loops Improve the Accuracy of the Predicted RNA Structure

David P. Gardner^a, Pengyu Ren^b, Stuart Ozer^c, and Robin R. Gutell^{a,1}

^aCenter for Computational Biology and Bioinformatics, Section of Integrative Biology in the School of Biological Sciences, and The Institute for Cellular and Molecular Biology, University of Texas at Austin, 2401 Speedway, Austin, TX 78712, USA

^bThe Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas 78712-1062

^cMicrosoft Corporation, 1 Microsoft Way, Redmond, WA 98052

Abstract

RNA is directly associated with a growing number of functions within the cell. The accurate prediction of different RNAs higher-order structure from their nucleic acid sequences will provide insight into their functions and molecular mechanics. We have been determining statistical potentials for a collection of structural elements that is larger than the number of structural elements determined with experimentally determined energy values. The experimentally derived free-energies and the statistical potentials for canonical base pair stacks are analogous, demonstrating that statistical potentials derived from comparative data can be used as an alternative energetic parameter. A new computational infrastructure - RNA Comparative Analysis Database (rCAD) - that utilizes a relational database was developed to manipulate and analyze very large sequence alignments and secondary structure datasets. Using rCAD, a richer set of energetic parameters for RNA fundamental structural elements including hairpin and internal loops was determined. A new version of RNAfold was developed to utilize these statistical potentials. Overall, these new statistical potentials for hairpin and internal loops integrated into the new version of RNAfold demonstrated significant improvements in the prediction accuracy of RNA secondary structure.

Keywords

statistical potentials; RNA folding; comparative analysis; RNA structure; accuracy of the predicted RNA structure

Introduction

“The comparative approach indicates far more than the mere existence of a secondary structural element; it ultimately provides the detailed rules for constructing the functional form of each helix. Such rules are a transformation of the detailed physical relationships of a helix and perhaps even reflection of its detailed energetics as well. (One might envision a future time when comparative sequencing provides energetic measurements too subtle for physical chemical measurements to determine).”¹

¹ Corresponding Author: robin.gutell@mail.utexas.edu; Fax: (512) 471-3878; Tel: (512) 232-4814.

Supplementary Data: Supplementary data associated with this article is available online at <http://www.rna.cccb.utexas.edu/SIM/4C/>.

The RNA sequences and their structures that we observe today are the last record of their biological ancestry. The snapshots of these RNA structures are the result of their evolution from a simpler structure and organization to their more sophisticated and complex state. Traditional experimental manipulation of biological systems expands our understanding of this system. These laboratory experiments are designed to test or expand upon a hypothesis, which are based in part on the underlying principles of RNA structure and a predicted or experimentally determined higher-order structure. In contrast, Mother Nature's experiments during the evolution of RNA are derived from an apparent random collection of mutations and other changes to the biological systems. The molecules and cells that survive these mutations reveal the characteristics of the RNA that maintain the integrity of their structure and function. Thus, the task for comparative analysis is complementary to hypothesis driven experimentation. Experimentalists prove, disprove, or determined more details for their hypothesis while comparative analysis attempts to decipher the principles that are the boundary conditions for the collections of biological data that have survived their evolutionary process.

The first stage of comparative analysis is the collection of a phylogenetically diverse set of RNA sequences and structures, followed by the comparative and covariation analysis of these linear strings of the four nucleotides in RNA, Adenine (A), Guanine (G), Cytosine (C), and Uracil (U), to identify a secondary structure that is similar for each of the RNA sequences that are in the same RNA family. For each of these RNA families, such as tRNA, 16S ribosomal (r)RNA, etc., many different sequences fold into the same higher-order structure. Encrypted in these relationships between sequence and higher-order structure models are the fundamental rules that govern the multiple levels of RNA structure, starting with the formation of the smaller structural elements such as the base pair and base stacking, continuing to larger structural elements that are composed of different types and arrangements of these base pairs and base stacks, and culminating in the formation of significantly larger higher order structure that have the capacity to dynamically catalyze chemical reactions and change their higher-order structure. To facilitate the RNA's function, these fundamental rules for RNA structure are also directly associated with the folding of an RNAs primary structure into its secondary, tertiary and quaternary structure.

Comparative analysis is composed of multiple dimensions of information. New technology is providing us with significant amounts of data for each of the dimensions of RNA: 1) nucleotide sequences for organisms that span the entire phylogenetic tree of life, 2) the accurate prediction of the secondary structures that are similar for each of the sequences in a single RNA family, 3) analysis of the high-resolution crystal structures and the comparative structure models are revealing different RNA structural motifs and elements that are the basic building blocks of a complete RNA structure and 4) the historical record of these evolving RNAs provides insight into their evolutionary dynamics and phylogenetic relationships.

In contrast to comparative analysis, physical biochemists usually use different experimental methods to solve simplified model systems that are less complex than the structure of the entire RNA. In particular, many laboratories have been obtaining free-energy values for different structural elements. Approximately 66% of many RNA structures are composed of a set of base pairs that form a regular helix^{2; 3}. The energetic values for consecutive base pairs have been studied for more than 25 years, initially focusing on canonical (ie. G:C, A:U, and G:U) and later non-canonical base pairs^{4; 5; 6; 7}. The energetic values for other types of structural elements, including helices with dangling ends⁸, hairpin⁹, internal^{10; 11} and multi-stem¹² loops, co-axial stacking¹³ and other structural motifs, e.g. the UAA/GAN motif¹⁴, have also been determined.

The most widely used program (and its derivatives) to predict an RNA secondary structure with the minimal free-energy from a single nucleic acid sequence is Mfold¹⁵. Early studies revealed that the accuracy of the predicted structures is dependent in part on the free-energy values for different structural motifs and the length of the RNA molecule¹⁶. As more free-energy values were determined for consecutive base pairs and new RNA structural motifs, the prediction accuracies increased. For example, the identification of the GNRA, UUCG, and CUUG hairpin tetraloops^{17; 18} and the subsequent determination of their extra stable free-energy value^{19; 20} resulted in an improvement in the prediction accuracy¹⁶. Subsequent studies showed the prediction accuracy is dependent on the phylogenetic group of the RNA molecule and the distance separating the nucleotides that are base paired, i.e. simple distance²¹. An analysis of a significantly larger dataset substantiated these earlier studies²² while providing a more detailed assessment of the factors that affect the prediction accuracy. For example, base pairs with a smaller simple distance occur significantly more frequently than base pairs with larger simple distances and the prediction accuracy of individual base pairs decreases exponentially as their simple distance increases²².

Thus a larger number of free-energy values for a variety of structural elements are required to accurately and routinely predict the secondary structure for an RNA molecule. Carl Woese's remarkable foresight in 1983 that comparative analysis can be used to determine RNA energetic measurements of higher-order structural elements was not appreciated at that time. However, this approach has been used in the prediction of protein structure^{23; 24; 25; 26; 27; 28; 29}, suggesting that Woese's idea could have the potential to reveal free-energy values for RNA that are not easily discernable with experimental methods. Within the past few years, statistical potentials determined with comparative analysis^{30; 31} for a few RNA structural elements were similar to the free energy values determined with experimental methods. The replacement of base pair stacking energetic parameters with statistical potentials generated from an analysis of RNA crystal structures showed similar prediction accuracies³⁰. These results emphasize that comparative data can be used to create similar energy values for some structural elements.

Previously, we determined statistical potentials for canonical base pair stacks that occur within a regular helix. While the statistical potentials for canonical base pair stacks resulted in a very minimal improvement in the accuracy of the predicted secondary structure, a larger improvement was observed when statistical potentials were determined for the nucleotides immediately flanking the ends of the helix and in small internal loops (1×1, 1×2, 2×2)³¹ and used in place of the equivalent experimentally determined energetic parameters.

Statistical learning procedures are another form of a knowledge-based approach for improving energetic parameters. Methods using stochastic context-free grammars (SCFGs) showed prediction accuracies³² near those of RNAstructure³³ and Mfold¹⁵. CONTRAfold³⁴ is based upon conditional log-linear models (CLLMs) which are an extension of SCFGs³⁴. The energetic parameters used by CONTRAfold were selected to maximize the conditional likelihood of the structures within the sequences analyzed. Andronescu et al.³⁵ utilized Constraint Generation and Boltzmann Likelihood methods to estimate their energetic parameters used by the program MultiFold.

Our confidence in Woese's 1983 statement influenced the development of our RNA Comparative Database (rCAD) (manuscript in preparation). The objective of this paper is to utilize rCAD to determine a richer set of energetic parameters from our comparative analysis of RNA sequences and their structures. We have developed new statistical potentials for hairpin and internal loops, but not base pair stacks and multi-stem loops. A modified version of RNAfold^{36; 37} was developed to utilize this new set of statistical potentials. The objective

of this paper was to quantify the effect that our new statistical potentials had on the accuracy of the predicted secondary structure model.

Results and Discussion

Hairpin Loop Comparative/Potential Ratio

To determine the likelihood that a structural element will occur in the correct structure, a ratio of the number of occurrences of that element in the comparative structure model divided by the number of potential occurrences of that element in the same RNA molecular class was determined (see methods). An example of the C/P ratio for tetraloop hairpin loops in bacterial 16S rRNA is shown in Figure 1. A few of the highlights are: 1) five of the tetraloop hairpin loops with any closing canonical base pairs have a C/P value greater than 0.5; 2) the closing base pair of these hairpin loops can alter the C/P values. For example, the C:G closing base pair usually increases the C/P values significantly for the twenty tetraloops shown in Figure 1.

The different closing base pair's affect on the C/P value for tetraloops are available at the CRW Site (<http://www.rna.ccbb.utexas.edu/SAE/2D/index.php>). Also available are the C/P ratios for hairpin loops of length 3-5 and for all of the molecular classes used in this study. The other structural statistics at the CRW Site, i.e. nucleotide, base pairs, internal and multi-stem loops all reveal significant biases in the frequencies of the sequences and their lengths. This general concept is used to create the statistical potentials.

Hairpin Loop Statistical Potentials

Hairpin loop statistical potentials were created and tested using eq. (2) and (4) (see methods). The sixteen RNA molecular classes (see methods) included in the creation of our statistical potentials were the bacterial and eukaryotic 5S rRNA, bacterial and eukaryotic 16S rRNA, bacterial 23S rRNA, tRNA³⁸, bacterial RNase P class A³⁹, bacterial signal recognition particle (SRP)⁴⁰, U1 spliceosomal RNA⁴¹, hepatitis C virus internal ribosome entry site (HCV IRES)⁴², ykok leader⁴³, TPP⁴⁴ and SAM⁴⁵ riboswitches, iron response element (IRE)⁴⁶, human immunodeficiency virus type 1 dimerisation initiation site (HIV DIS)⁴⁷ and UnaL2 Line 3' element⁴⁸. The first flanking (closing) canonical base pair is included when generating our comparative and potentials counts and statistical potentials.

For hairpin loops of length four, the values of m and b in eq. (2) (see methods) with the best accuracy were 2.25 and 0.8, respectively. For the restricted range of 0 to 2 for $-\ln(C/P)$ (see methods), the statistical potentials of hairpin loops of length four will vary from 5.3 to 0.8 kcal/mol with 5.3 kcal/mol set as the default value. Hairpin loops of different sizes will have different m and b values (Supplemental Data, see Excel file HPComparison). Statistical potentials were generated for 908 hairpin loops plus default values.

The approach used to determine the statistical potentials for hairpin loops is illustrated with a comparison with recent experimentally derived tetraloop free-energy values⁴⁹. For the 1536 possible combinations (256 hairpin loops $\times$ 6 base pairs), 1225 (80%) had an absolute difference less than 0.5 kcal/mol and 1243 (81%) had an absolute difference less than 1.0 kcal/mol. 191 (12%) combinations had an absolute differences between 1.025 and 2.0 kcal/mol and 102 (7%) with differences between 2.075 and 3.1 kcal/mol (Supplemental Data, see Excel file HPComparison). The 14 tetraloop closing base pair combinations with the largest absolute difference all had smaller kcal/mol values and thus are more energetically stable. However, the majority of the combinations (232 out of 311) with absolute difference greater than 0.5 kcal/mol had experimentally derived energetic values smaller (i.e. more stable) than the derived statistical potential.

For tri-loops, the experimentally derived free-energy values were taken from Thulasi et al. 2010⁵⁰. Only six out of the 384 (.2%) tri-loop combinations had an absolute difference of less than 1.0 kcal/mol between the experimentally derived free-energies and statistical potentials. Most of the tri-loops (369 out of 384) (94%) had absolute differences between 1.0 and 2.0 kcal/mol. The absolute difference for the other 23 combinations ranged from 2.028 to 2.61 kcal/mol (Supplemental Data, see Excel file HPCcomparison). For the penta-loops comparison, the energetic parameters from TURNER04^{6; 51} were used. Of the 6144 possible penta-loops combinations, 3354 (55%) had an absolute difference of 0.5 kcal/mol or less and 4674 (76%) had an absolute difference less than 1.0 kcal/mol. 1146 (19%) had an absolute difference between 1.02 and 2.0 kcal/mol, 287 (5%) between 2.068 and 3.0 kcal/mol and 36 (.6%) between 3.1 and 4.0 kcal/mol. The remaining pentaloop has an absolute difference of 4.408 kcal/mol (Supplemental Data, see Excel file HPCcomparison). Statistical potentials have been created for hairpin loops for all observed lengths in the molecular classes studied with comparative methods.

Internal Loop Statistical Potentials

Internal loop statistical potentials were created using eq. (2) and (4). The same sixteen RNA molecular classes used in the generation of the hairpin loop statistical potentials were used for the internal loops. Both base pairs flanking an internal loop are included in the generation of statistical potentials for internal loops. For 1×1 internal loops, the values of m and b in eq. (2) (see methods) with the best accuracy were 2.5 and -1.0, respectively. For the restricted range of 0 to 2 for $-\ln(C/P)$ (see methods), the statistical potentials of 1×1 internal loops will vary from 4.0 to -1.0 kcal/mol with 4.0 kcal/mol set as the default value. Internal loops of different sizes will have different m and b values (Supplemental Data, see Excel file ILComparison). Statistical potentials were generated for 1368 internal loop plus default values.

The approach used to determine the statistical potentials for internal loops is illustrated with 1×1 internal loops. For these internal loops, the absolute differences between the statistical potentials and the Turner 2004⁶ experimentally derived energetic parameters were usually large. There are 360 possible 1×1 internal loops – 6 base pairs × 6 base pairs × 10 internal loops. Only 57 out of the 360 (16%) had an absolute difference of less than 1.0 kcal/mol and just 10 (3%) had absolute differences between 1.0 and 2.0 kcal/mol. 130 (36%) had absolute differences between 2.0 and 3.0 kcal/mol, 111 (30%) had an absolute difference between 3.0 and 4.0 kcal/mol. The 30 1×1 internal loops with the largest difference between experimentally derived free-energies and statistical potentials all had a G-G internal loop. The values for the experimentally derived free-energies and statistical potentials for all 360 1×1 and all 9216 2×2 internal loops are in the supplemental data (Excel file ILComparison). Statistical potentials have been created for internal loops for any length observed on the 5' and 3' sides of the loop in those molecular classes studied with comparative methods.

Evaluation of Hairpin Loop Statistical Potentials

The prediction of an RNA structure is evaluated with the statistical potentials for hairpin loops. In previous versions of RNAfold, the only hairpin loops with specific free-energy values were tri- and tetraloops. Free-energy values for longer hairpin loops were calculated using the length of the hairpin loop and the composition of the first and last nucleotides of the hairpin loop and the flanking (closing) base pair. To determine if statistical potentials generated with eq. (2) and (4) would improve the accuracy of RNA secondary structure prediction, the program RNAfold^{36; 37} was modified to accept detailed statistical potentials for hairpin loops of any length. When testing the hairpin loop statistical potentials, the experimentally derived energetic parameters (TURNER99) for base pair stacks and internal and multi-stem loops were used.

Similar to previous studies^{21; 31}, sensitivity has been used to gauge prediction accuracy. Sensitivity is defined as the number of canonical base pairs in the predicted minimal free-energy structure present in the comparative model divided by the total number of comparative canonical base pairs. Differences in prediction accuracy are defined as (sensitivity using statistical potentials)-(sensitivity using other energetic parameters and/or folding programs). If a program returns suboptimal structures, only the optimal structure is used in our analysis.

Results in the Supplemental Data (see supplemental.pdf, pages 1-4) reveal that the statistical potentials for hairpin loops improved the prediction of the RNA structure.

Evaluation of Internal Loop Statistical Potentials

To utilize the new internal loop statistical potentials, the functionality of RNAfold was again extended to accept a wider range of energetic parameters. The original version of RNAfold had specific free-energy values for internal loops of lengths 1×1, 1×2, 2×2 and 2×3. For larger internal loops, the calculation of the experimentally derived free-energy values was based on the number of nucleotides in the internal loop plus the composition of the ends of the internal loop and both flanking base pairs. The modified RNAfold accepts specific free-energy values for internal loops of any size. When testing hairpin loop statistical potentials, the experimentally-derived energetic parameters (TURNER99) for base pair stacks and hairpin and multi-stem loops are used.

Results in the Supplemental Data (see supplemental.pdf, pages 1-4) reveal that the statistical potentials for the internal loops improved the prediction of the RNA structure.

Combining Statistical Potentials and Comparison with Other Programs

The prediction accuracy using the combination of hairpin and internal loop molecule-independent statistical potentials for all sixteen RNA molecular classes were compared with the results from four other RNA folding programs – RNAfold³⁶ (TURNER99), RNAstructure³³ using just TURNER04 and using TURNER04 plus the newer tri- and tetraloop thermodynamic parameters^{49; 50}, CONTRAfold³⁴ and MultiFold (BL* parameter set)³⁵. RNAfold and RNAstructure utilize experimentally derived energetic parameters while CONTRAfold and MultiFold use parameters derived with statistical learning. When testing the hairpin and internal loop statistical potentials with RNAfold, the experimentally-derived energetic parameters (TURNER99) for base pair stacks and multi-stem loops are used.

Overall, the combined molecule-independent statistical potentials outperformed the other four programs (Figs. 2a and 2b). On average, over the sixteen RNA molecular classes, our statistical potentials scored 15% higher than RNAfold (TURNER99), 14% for RNAstructure (TURNER04), 14% higher for RNAstructure (TURNER04 Plus), 12% for CONTRAfold and 13% for MultiFold. Our statistical potentials outperformed all four programs for all sixteen RNA molecular classes with the exception of the ykok leader RNA where RNAfold (TURNER99) matched our score and RNase P A where CONTRAfold scored 3% higher. The difference in accuracy between our statistical potentials and the competing program with the best results for a given molecule ranged from -3% (RNase P A) to 15% (UnaL2 Line 3' element) (Figs. 2a and 2b). On average, our statistical potentials outperformed the program with the best results for a given RNA molecule by 7% (Supplemental Data, see Excel file Accuracies.xlsx). Standard deviation results for each program on each molecule are contained in the Supplemental Data (see supplemental.pdf, pages 5-6).

Two methods were used to evaluate the cross-validation of the statistical potentials. The first utilized the same method used for MultiFold³⁵. The results in the Supplemental Data reveal

that the accuracies of the predicted RNA secondary structures are very similar between the training and testing on the full set of sequences and on an 80/20% split (see supplemental.pdf, pages 7-8). The second method tested our statistical potentials and the four other RNA folding programs against nine control RNA molecular classes (see methods) that were not used in the generation of the statistical potentials. On average, over these nine RNA molecular classes, our statistical potentials essentially equaled the performances of the four other RNA folding programs (Supplemental Data, see supplemental.pdf, pages 9-14).

Given that our approach utilizes comparative data for generating the statistical potentials, it is not surprising that they perform only on par with the other RNA folding programs over the control RNA molecular classes. The nine RNA molecular classes in our test set must have some structural elements that are not present and/or absent in the original 16 classes. This indicates that increasing the set of RNA molecular classes used to generate the statistical potentials is necessary before the statistical potentials will have higher accuracies for a larger number of molecular classes. During the course of these studies we observed improvements in the accuracies for a larger number of molecular classes as the training set included more RNA families.

RNA Folding Website

RNA sequences can be folded on our modified RNAfold program that contains our new statistical potentials. The URL is: <http://www.rna.ccbb.utexas.edu/SAE/2E/Folding2D/>. The C# code and the new statistical potentials will also be made available at this website.

Summary

The focus of this study was to improve the energetic parameters for hairpin and internal loops. Previously, the base pair stack statistical potentials created with comparative data, on average, only slightly improved the prediction accuracy demonstrating that statistical potentials can generate analogous energetic parameters³¹. This minor improvement in the accuracy from the base pair stack statistical potentials was not as much as we anticipated. However our previous analysis did reveal that flanking nucleotides of the hairpin and internal loops did have a more pronounced improvement, suggesting that a richer set of statistical potentials for the loop regions of the secondary structure could have a larger enhancement in the accurate prediction.

The new comparative analysis system in development in the Gutell lab, rCAD (manuscript in preparation), was used to determine this collection of statistical potentials that represents more of the structural elements present in RNA molecules. This new set of energetic parameters used a new structural statistic – the comparative/potential ratio. The RNAfold program was modified to utilize our larger set of statistical potentials since it originally had more limited hairpin and internal loop energetic parameters.

This modified RNAfold program and our new hairpin and internal loop statistical potentials demonstrated significant increases in the prediction accuracy of RNA secondary structure. Over sixteen RNA molecular classes, the statistical potentials always outperformed the four existing RNA folding programs with the exception of two RNA molecules where our accuracies were equal to or slightly worse than one other program. On average, the improvements ranged from 12% to 15% compared to the competing four programs. Our program predicted the RNA secondary structure better in 78 of the 80 comparisons. When our program was not included in these comparisons, RNAfold (TURNER99) and RNAstructure (TURNER99+) outperformed the other programs in 19 out of 64 comparisons, 20 out of 64 for RNAstructure (TURNER04), 39 out of 64 for MultiFold and 45 out of 64 for CONTRAfold. Our statistical potentials also were approximately the same

as the performances of the other four programs when tested over the nine additional control RNA molecular classes that were not used in the generation of the statistical potentials.

Our intention with this work was to determine if this generalized approach would improve the prediction of RNA secondary structure beyond current approaches. Given that this approach did significantly increase prediction accuracy in the sixteen training RNA molecular classes, in the future we will extend and improve upon our generalized approach with a variety of approaches.

We will add more RNA molecular classes when generating the statistical potentials. We also will aim to identify the most essential structural elements and components that will produce the highest accuracy of the predicted RNA structure. This should help identify general structural families and reduce the number of needed energetic parameters. We will also investigate extending the statistical potentials and folding program to utilize non-nearest neighbor effects.

Methods

Comparative and Potential Secondary Structural Elements

A potential secondary structural element, such as a hairpin loop, internal loop or a helix, is defined as the set of nucleotides that forms the motif. This potential structural element may or may not occur in the comparative secondary structure of the RNA molecule, while every comparative structural element is a potential structural element. Our objective is to generate a statistical potential from the ratio of comparative and potential structural elements.

Potential hairpin loops are a set of consecutive nucleotides of a specific length that are flanked by two or more canonical base pairs in the RNA sequence (Fig. 3a). The determination of a potential internal loop initiates with a comparative helix. The nucleotides flanking the 5' and 3' ends of this helix that contain at least two potential canonical base pairs are identified (Fig. 3b-c). The nucleotides between the comparative and potential helices are defined as a potential internal loop.

Creation of Statistical Potentials

A basic assumption in the creation of the statistical potentials is:

$$-\ln(C/P) \sim \text{free-energy} \quad (1)$$

where C is the frequency of a structural element appearing in the comparative structure and P is the potential frequency of the structural element. Every comparative structure is considered to be a potential structure as well; C/P will have values in the range between 0 and 1. A typical statistical potential utilizes $-\ln(C)$ with C normalized with the frequency of individual nucleotides. The formula proposed here can be considered as normalized by the potential to form a structure element. A statistical potential is determined with the equation:

$$-m(\ln(C/P)) + b = SP \quad (2)$$

where SP is a statistical potential and m and b are global parameters that will be selected to optimize the overall accuracy of the folding program. For the vast majority of structural elements, the comparative count will be 0 or the C/P ratio too low and the default value will be used. Restricting the range of values for $-\ln(C/P)$ between 0 and 2 provides the best

prediction accuracies; this restricts C/P values to a minimum of .01. If a structural element has no potential structures or the C/P value is less than .01, the C/P value is set to .01. The default value for a structural element is set to:

$$-m * 2 + b = \text{default} \quad (3)$$

Molecule-independent Statistical Potential

Initially, a set of statistical potentials will be generated for each type of RNA molecular class analyzed (e.g. 16S rRNA – bacteria). The statistical potentials for each molecule-specific set will not have detailed values for all possible structural elements. Our ultimate goal is to create one set of statistical potentials that are applicable for all types of RNAs. To create a molecule-independent set of statistical potentials, each molecule-dependent set is treated as a member of a Boltzmann distribution. For every secondary structural element, the molecule-independent statistical potential is a Boltzmann weighted sum of statistical potentials from each molecule i :

$$SP_{\text{molecule-ind}} = \frac{\sum_{i \in I} \exp(-SP_i / k_b T) SP_i}{\sum_{i \in I} \exp(-SP_i / k_b T)} \quad (4)$$

CRW Site

The Gutell lab's CRW Site³⁸ <http://www.rna.ccbb.utexas.edu/DAT/3C/Structure/index.php> has a diverse collection of secondary structure models predicted from comparative analysis for different phylogenetic groups of the 5S, 16S, and 23S rRNAs, tRNAs for different amino acids, and group I and II introns. The number of secondary diagrams currently available is 1092, while the number of sequences with only basepair information is 54525. The accuracy of these secondary structure models is extremely high; approximately 97% of the base pairs in the ribosomal RNAs structures predicted with comparative methods are present in the high resolution crystal structure⁵².

RNA Comparative Analysis Database

All sequence and comparative structure information is stored in the RNA Comparative Analysis Database (rCAD). rCAD at the time this manuscript was submitted contains 293,039 aligned RNA sequences and their comparative structure information. This data is utilized to determine the number of structural elements in the comparative structures. rCAD also contains structural statistics (comparative and potential counts) on nearly 500,000 different internal loops and almost 2.3 million different hairpin loops.

RNA Molecular Classes

The RNA molecule sequences and structures initially studied for their comparative and potential counts of structural elements and used in the generation of the statistical potentials were aligned and created by the Gutell lab and are available at <http://www.rna.ccbb.utexas.edu/DAT/3C>. They include sequences from the bacterial and eukaryotic phylogenetic groups and from 5S, 16S and 23S rRNA and tRNA.

Additional RNA sequences and structures were obtained from the RFam web site⁵³. These included bacterial RNase P class A, bacterial signal recognition particle (SRP), U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader,

TPP and SAM riboswitches, iron response element (IRE), human immunodeficiency virus type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element. All of these sequences and structures were taken from their respective RFam full alignments.

For the training and initial testing of the statistical potentials, sequences with a similarity of greater than 97% were removed to minimize the folding of duplicate RNA sequences. Also, only complete or nearly complete sequences were analyzed. The total number of RNA sequences analyzed for testing RNA secondary structure accuracy for each molecular class is: 1094 bacterial and 258 eukaryotic 16S rRNA, 65 bacterial 23S rRNA, 230 bacterial and 310 eukaryotic 5S rRNA, 2112 tRNA, 274 RNase P class A, 937 U1 spliceosomal RNA, 1049 bacterial SRP, 550 HCV IRES, 188 ykok leader, 726 TPP and 589 SAM riboswitches, 371 IRE, 136 HIV DIS and 572 UnaL2 Line 3' element. The number of sequences and their average length are available in supplemental data (Supplemental Data, see supplemental.pdf).

For the additional testing of control RNA molecules, seven sets of RNA sequences and structures were obtained from the RFam web site. These are the RNase P B³⁹, Hammerhead III ribozyme⁵⁴, purine riboswitch⁵⁵, hepatitis delta virus (HDV) ribozyme⁵⁶, HIV ribosomal frameshift signal⁵⁷, GEMM cis-regulatory element⁵⁸ and the R2 RNA element⁵⁹. All of these sequences are taken from their respective RFam seed alignment. Two sets of RNA sequences and structures are from the Gutell lab - mitochondrial and archaeal 16S rRNA³⁸.

The total number of RNA sequences for each of the nine classes is: 366 RNase P B, 84 Hammerhead III ribozymes, 133 purine riboswitches, 33 HDV ribozymes, 145 HIV ribosomal frameshift signal, 162 GEMM cis-regulatory element and 15 R2 RNA element. 128 and 143 RNA sequences were tested for mitochondrial and archaeal 16S rRNA, respectively. The number of sequences and their average length are available in supplemental data (Supplemental Data, see supplemental.pdf).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This paper is dedicated to Dr. Carl Woese for his intuition that comparative analysis could reveal “energetic measurements too subtle for physical chemical measurements to determine”, and to our erstwhile colleague Dr. Jim Gray whose pioneering work on transaction control enables database systems to be the foundation for Jim's vision of the ‘Fourth Paradigm’, following experimental, theoretical, and computer science. Jim appreciated that the overwhelming amount of multiple dimensions of information was not strictly a computer science problem, but instead a collaborative effort between computer scientists and (in this case) molecular biologists. The authors are also most grateful to Yuxing Li, Jamie Cannone, Ame Wongsu, and Yanan Jiang for help establishing the RNA folding website. Grants from the Robert A. Welch Foundation (grant numbers F-1691 (PR) and F-1427 (RG)), National Institutes of Health (grant numbers R01 GM0796686 (PR) and R01 GM067317 (RG) and GM085337 (RG)), and Microsoft Research TCI/ER grant (RG) were essential for this project to come to fruition. The authors appreciated the constructive comments from the reviewers and editor.

References

1. Woese CR, Gutell R, Gupta R, Noller HF. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol Rev.* 1983; 47:621–69. [PubMed: 6363901]
2. Gutell RR, Weiser B, Woese CR, Noller HF. Comparative anatomy of 16-S-like ribosomal RNA. *Prog Nucleic Acid Res Mol Biol.* 1985; 32:155–216. [PubMed: 3911275]
3. Gutell RR, Cannone JJ, Shang Z, Du Y, Serra MJ. A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol.* 2000; 304:335–54. [PubMed: 11090278]

4. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci USA*. 1986; 83:9373–7. [PubMed: 2432595]
5. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*. 1999; 288:911–40. [PubMed: 10329189]
6. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*. 2010; 38:D280–2. [PubMed: 19880381]
7. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem*. 1998; 37:14719–35. [PubMed: 9778347]
8. Liu JD, Zhao L, Xia T. The dynamic structural basis of differential enhancement of conformational stability by 5'- and 3'-dangling ends in RNA. *Biochemistry*. 2008; 47:5962–75. [PubMed: 18457418]
9. Antao VP, Tinoco I Jr. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res*. 1992; 20:819–24. [PubMed: 1371866]
10. Schroeder SJ, Burkard ME, Turner DH. The energetics of small internal loops in RNA. *Biopolymers*. 1999; 52:157–67. [PubMed: 11295748]
11. Walter AE, Wu M, Turner DH. The stability and structure of tandem GA mismatches in RNA depend on closing base pairs. *Biochem*. 1994; 33:11349–54. [PubMed: 7537087]
12. Diamond JM, Turner DH, Mathews DH. Thermodynamics of three-way multibranch loops in RNA. *Biochem*. 2001; 40:6971–81. [PubMed: 11389613]
13. Walter AE, Turner DH. Sequence dependence of stability for coaxial stacking of RNA helices with Watson-Crick base paired interfaces. *Biochem*. 1994; 33:12715–9. [PubMed: 7522562]
14. Shankar N, Kennedy SD, Chen G, Krugh TR, Turner DH. The NMR Structure of an Internal Loop from 23S Ribosomal RNA Differs from Its Structure in Crystals of 50S Ribosomal Subunits. *Biochem*. 2006; 45:11776–11789. [PubMed: 17002278]
15. Zuker M. On finding all suboptimal foldings of an RNA molecule. *Science*. 1989; 244:48–52. [PubMed: 2468181]
16. Jaeger JA, Turner DH, Zuker M. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA*. 1989; 86:7706–10. [PubMed: 2479010]
17. Woese CR, Winker S, Gutell RR. Architecture of ribosomal RNA: constraints on the sequence of “tetra-loops”. *Proc Natl Acad Sci USA*. 1990; 87:8467–71. [PubMed: 2236056]
18. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol*. 1990; 216:585–610. [PubMed: 2258934]
19. Tuerk C, Gauss P, Thermes C, Groebe DR, Gayle M, Guild N, Stormo G, d'Aubenton-Carafa Y, Uhlenbeck OC, Tinoco I Jr, et al. CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc Natl Acad Sci USA*. 1988; 85:1364–8. [PubMed: 2449689]
20. Antao VP, Lai SY, Tinoco I Jr. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res*. 1991; 19:5901–5. [PubMed: 1719483]
21. Konings DA, Gutell RR. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*. 1995; 1:559–74. [PubMed: 7489516]
22. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*. 2004; 5:105. [PubMed: 15296519]
23. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*. 1976; 9:945–950. [PubMed: 1004017]
24. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*. 2005; 15:285–289. [PubMed: 15939584]
25. Floudas CA, Fung HK, McAllister SR, Monnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*. 2006; 61:966–988.

26. Kryshchuk A, Venclovas C, Fidelis K, Moulton J. Progress over the first decade of CASP experiments. *Proteins*. 2005; 61:225–236. [PubMed: 16187365]
27. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci*. 2006; 15:2507–2524. [PubMed: 17075131]
28. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci USA*. 2007; 104:3177–82. [PubMed: 17360625]
29. Xu BS, Yang YD, Liang HJ, Zhou YQ. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins-Structure Function and Bioinformatics*. 2009; 76:718–730.
30. Dima RI, Hyeon C, Thirumalai D. Extracting stacking interaction parameters for RNA from the data set of native structures. *J Mol Biol*. 2005; 347:53–69. [PubMed: 15733917]
31. Wu JC, Gardner DP, Ozer S, Gutell RR, Ren P. Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. *J Mol Biol*. 2009; 391:769–83. [PubMed: 19540243]
32. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*. 2004; 5:71. [PubMed: 15180907]
33. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*. 2010; 11:129. [PubMed: 20230624]
34. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. 2006; 22:e90–8. [PubMed: 16873527]
35. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. *RNA*. 2010; 16:2304–18. [PubMed: 20940338]
36. Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res*. 2003; 31:3429–31. [PubMed: 12824340]
37. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatsh Chem*. 1994; 125:167–188.
38. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*. 2002; 3:2. [PubMed: 11869452]
39. Brown JW. The Ribonuclease P Database. *Nucleic Acids Res*. 1999; 27:314. [PubMed: 9847214]
40. Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C, Samuelsson T. SRPDB: Signal Recognition Particle Database. *Nucleic Acids Res*. 2003; 31:363–364. [PubMed: 12520023]
41. Kretzner L, Krol A, Rosbash M. *Saccharomyces cerevisiae* U1 small nuclear-RNA secondary structure contains both universal and yeast-specific domains. *Proc Natl Acad Sci USA*. 1990; 87:851–855. [PubMed: 2405391]
42. Gallego J, Varani G. The hepatitis C virus internal ribosome-entry site: A new target for antiviral research. *Biochemical Society Transactions*. 2002; 30:140–145. [PubMed: 12023841]
43. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA*. 2004; 101:6421–6. [PubMed: 15096624]
44. Miranda-Rios J, Navarro M, Soberon M. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc Natl Acad Sci USA*. 2001; 98:9736–9741. [PubMed: 11470904]
45. Grundy FJ, Henkin TM. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Molecular Microbiology*. 1998; 30:737–749. [PubMed: 10094622]
46. Hentze MW, Kuhn LC. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl Acad Sci USA*. 1996; 93:8175–8182. [PubMed: 8710843]
47. McBride MS, Panganiban AT. The human immunodeficiency virus type 1 encapsidation site is a multipartite RNA element composed of functional hairpin structures. *Journal of Virology*. 1996; 70:2963–2973. [PubMed: 8627772]

48. Baba S, Kajikawa M, Okada N, Kawai G. Solution structure of an RNA stem-loop derived from the 3' conserved region of eel LINE UnaL2. *RNA*. 2004; 10:1380–1387. [PubMed: 15273327]
49. Sheehy JP, Davis AR, Znosko BM. Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA*. 2010; 16:417–29. [PubMed: 20047989]
50. Thulasi P, Pandya LK, Znosko BM. Thermodynamic Characterization of RNA Triloops. *Biochemistry*. 2010; 49:9058–9062. [PubMed: 20843054]
51. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA*. 2004; 101:7287–92. [PubMed: 15123812]
52. Gutell RR, Lee JC, Cannone JJ. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*. 2002; 12:301–10. [PubMed: 12127448]
53. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. Rfam: updates to the RNA families database. *Nucleic Acids Res*. 2009; 37:D136–40. [PubMed: 18953034]
54. Murray JB, Terwey DP, Maloney L, Karpeisky A, Usman N, Beigelman L, Scott WG. The structural basis of hammerhead ribozyme self-cleavage. *Cell*. 1998; 92:665–673. [PubMed: 9506521]
55. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*. 2003; 113:577–86. [PubMed: 12787499]
56. Chen PJ, Kalpana G, Goldberg J, Mason W, Werner B, Gerin J, Taylor J. Structure and replication of the genome of the hepatitis delta-virus. *Proc Natl Acad Sci USA*. 1986; 83:8774–8778. [PubMed: 2430299]
57. Biswas P, Jiang X, Pacchia AL, Dougherty JP, Peltz SW. The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy. *Journal of Virology*. 2004; 78:2082–2087. [PubMed: 14747573]
58. Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH, Breaker RR. Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science*. 2008; 321:411–3. [PubMed: 18635805]
59. Ruschak AM, Mathews DH, Bibillo A, Spinelli SL, Childs JL, Eickbush TH, Turner DH. Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA*. 2004; 10:978–987. [PubMed: 15146081]

Figure 1.

The ranked order of the twenty tetraloop hairpin loops (with any closing canonical base pair) with the highest comparative/potential ratios (red bars) are shown along the x-axis. The C/P ratio for each of these tetraloop hairpin loops is shown on the y-axis. The ratios for tetraloop hairpin loops flanked by any canonical base pair are shown as red bars while the tetraloop hairpin loops flanked by a CG base pair are shown as blue bars. The values are for bacterial 16S rRNA.

Figure 2. RNA secondary structure prediction accuracies for four RNA folding programs: RNAfold, RNAstructure (TURNER04 & TURNER04 plus newer tri- and tetraloop thermodynamic parameters), CONTRAfold, MultiFold and RNAfold using statistical potentials. Results for sixteen RNA molecular classes are divided into a) bacterial 5S rRNA, eukaryotic 5S rRNA, bacterial 16S rRNA, bacterial 23S rRNA, tRNA, eukaryotic 16S rRNA, RNase P A and bacterial SRP; b) U1 spliceosomal RNA, hepatitis C virus internal ribosome entry site (HCV IRES), ykok leader, TPP and SAM riboswitches, iron response element (IRE), HIV type 1 dimerisation initiation site (HIV DIS) and UnaL2 Line 3' element.

Figure 3.

a) Nucleotides in the tetraloop hairpin loops that occur in the comparative structure for a modified *Escherichia coli* 16S rRNA secondary structure between positions 118 and 241 are colored green. For this figure the *E.coli* sequence was changed at a few positions to create better examples of potential base pairings that form hairpin loops. Potential tetraloop hairpin loop, as defined by four nucleotides that are closed by two or more canonical base pairs, are colored red. The base pairs flanking the tetraloop hairpin loops are circled and connected with a red line. Nucleotides that are base paired in the comparative structure are connected with a thick black line. b) Nucleotides in the internal loop that occur in our modified *Escherichia coli* comparative secondary structure between positions 139 and 184 are colored green; b&c) Nucleotides in potential internal loops are colored red and the nucleotides that form a set of base pairs within the potential helix in the internal loop are circled and connected with a red line. Nucleotides that are base paired in the comparative structure are connected with a thick black line.