1378–1386 *Nucleic Acids Research, 2009, Vol. 37, No. 5* *Published online 7 January 2009*
doi:10.1093/nar/gkn987

RNA structure prediction from evolutionary patterns of nucleotide composition

S. Smit[1,2,*], R. Knight[3] and J. Heringa[1]

[1]Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit, 1081 HV Amsterdam, [2]Centre for Medical Systems Biology, Niels Bohrweg 2, 2300 RA Leiden, The Netherlands and [3]Department of Chemistry and Biochemistry, University of Colorado, Boulder CO 80309, USA

Received September 11, 2008; Revised and Accepted November 21, 2008

ABSTRACT

Structural elements in RNA molecules have a distinct nucleotide composition, which changes gradually over evolutionary time. We discovered certain features of these compositional patterns that are shared between all RNA families. Based on this information, we developed a structure prediction method that evaluates candidate structures for a set of homologous RNAs on their ability to reproduce the patterns exhibited by biological structures. The method is named SPuNC for 'Structure Prediction using Nucleotide Composition'. In a performance test on a diverse set of RNA families we demonstrate that the SPuNC algorithm succeeds in selecting the most realistic structures in an ensemble. The average accuracy of top-scoring structures is significantly higher than the average accuracy of all ensemble members (improvements of more than 20% observed). In addition, a consensus structure that includes the most reliable base pairs gleaned from a set of top-scoring structures is generally more accurate than a consensus derived from the full structural ensemble. Our method achieves better accuracy than existing methods on several RNA families, including novel riboswitches and ribozymes. The results clearly show that nucleotide composition can be used to reveal the quality of RNA structures and thus the presented technique should be added to the set of prediction tools.

INTRODUCTION

The discovery of many noncoding RNAs, involved in catalysis and gene regulation, has intensified the attention for RNA research (1). To understand the structural, functional and mechanistic properties of these molecules, structure prediction is an essential tool. Efforts aimed at predicting RNA structures from sequence information started over three decades ago (2), and the field is still under rapid development as evidenced by the many prediction methods developed in recent years [reviewed in (3–9)]. Differences between methods arise from the type of input (single versus multiple sequences, aligned versus unaligned sequences), their prediction target (base pair types and structural topology), and the kinds of evidence they use for the prediction (e.g. covariance, thermodynamics or experimental data). Free energy minimization methods are extensively developed, but their accuracy is limited due to several factors including our incomplete knowledge of the RNA folding process (5,6,8). Comparative sequence methods are in general more accurate than single-sequence minimum free energy methods. When many (hundreds or thousands) of sequences are available, very accurate structural models can be derived [over 97% in the ribosomal RNA (10)]. However, when a limited number of sequences is available, as is often the case, accuracy typically ranges from 30%–90%, depending on the number, length and similarity of the sequences in the alignment (5).

Here we present a novel approach to RNA structure prediction for a set of homologous sequences, exploiting an information source thus far unused for this purpose: evolutionary patterns of nucleotide composition. We assess the quality of candidate structures in an ensemble using generic compositional patterns exhibited by biological structures. Our method fits in with two prediction strategies that are being explored in recent years: mining the information provided by an ensemble of (suboptimal) RNA structures and combining multiple information sources in a single prediction method. The former strategy has led to efficient algorithms for generating all suboptimal structures within a certain energy range (11) and for statistical sampling from the complete Bolzmann ensemble (12). The second strategy emerged because the use of a single source often does not lead to sufficient

*To whom correspondence should be addressed. Tel: +31 020 5983714; Fax: +31 020 5987653; Email: S.Smit@few.vu.nl

accuracy of the predictions when a limited number of sequences is available. The more powerful methods at this moment combine evidence from various sources. Examples are RNAalifold (13), RNAstructure (14), and Bayesfold (15), using covariation, thermodynamics, and experimental data. In addition, important advances in structure prediction are made due to the use of novel information sources, such as abstract shapes (16) as implemented in the RNAshapes package (17,18) or 'nucleotide cyclic motifs' (19) as implemented in the MC-Fold and MC-Sym pipeline (20).

To complement current structure prediction methods, we describe how nucleotide composition can aid structure prediction. Different classes of structural RNAs are known to have certain compositional biases (21) and within an RNA family the composition of structural elements changes in consistent ways throughout the evolution (22). Building on these observations, we found certain characteristics of these patterns that are shared between all RNA families. The key idea of our method is to use these characteristic patterns of nucleotide composition, known to be exhibited by biological sequences and structures, to distinguish between realistic and unrealistic foldings. The method is built around a scoring function that captures the universal properties. It is used to evaluate many structures in an ensemble. Structures more similar to the true structure will display the expected trends and will receive a more positive evaluation.

In this work we address the following questions. Can the patterns of nucleotide composition be used to identify the most realistic structures in an ensemble? Is a consensus structure calculated from top-scoring structures more accurate than the consensus from the whole ensemble? How do predictions using the nucleotide composition as the only source compare to predictions from existing structure prediction methods? First we will introduce the concept of nucleotide composition and the generic features. Then we will explain how this information is used for structure prediction. The research questions will be addressed in a performance test involving a number of different RNA families, ranging from the hammerhead ribozyme to the large subunit of the ribosomal RNA.

MATERIALS AND METHODS

We have developed a method for RNA structure prediction using characteristic patterns of nucleotide composition as observed in biological structures. The core of the method is a scoring function that describes these patterns. The prediction method is coined SPuNC for 'Structure Prediction using Nucleotide Composition'. The SPuNC algorithm consists of the following steps: (1) given a multiple sequence alignment, generate an ensemble of candidate structures, (2) score all structures in the ensemble with a scoring function, and (3) return top-scoring structure(s) or consensus structure. In this section we will describe the compositional patterns and their generic features, the construction and application of the scoring function, the ensemble generation, and the performance measurements of the algorithm. A web interface to the method, the source code, and the datasets used in this study are available at www.ibi.vu.nl/programs/spuncwww.

Nucleotide composition

The combination of an alignment and a structure that the sequences fold into can be described in terms of nucleotide composition, which can be calculated for different parts of the RNA structure. We distinguish four structural elements in RNA structures (Figure 1A and B). The category 'stem' contains all paired residues, the category 'loop' contains all unpaired residues that connect the upstream and downstream half of a helix, 'bulge' contains all unpaired regions that connect exactly two helices, and 'other' contains all other unpaired residues, including multi-helix

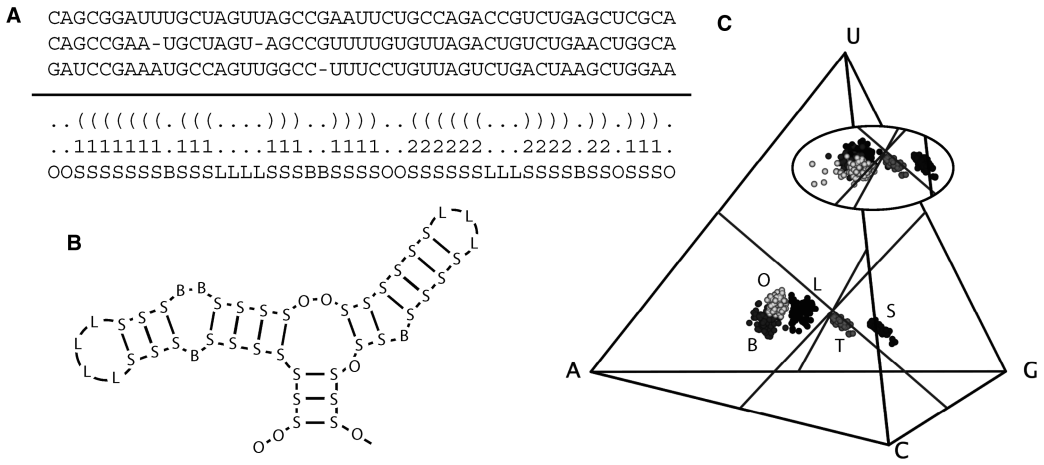

Figure 1. RNA structure and nucleotide composition. (**A**) Alignment of three RNA sequences with a reference structure (both vienna structure and pairing mask) and the corresponding structural classification. (**B**) Secondary structure diagram indicating four different structural categories: S = stem, L = loop, B = bulge, O = other. (**C**) RNA composition space, containing the nucleotide composition of 80 SSU rRNA sequences decomposed under the *E. coli* reference structure [source: Comparative RNA Web (23)]. The space contains five distributions: S = stem, L = loop, B = bulge, O = other, T = total sequence. The inset shows the compositional patterns generated by an incorrect structure, containing more scatter and overlap in the loops and bulges and increased scatter and lower GC content in the stems.

junctions, ends and pseudoknotted regions. This classification scheme is chosen over the more coarse-grained paired/unpaired scheme and the more detailed six-way classification (22), because it distinguishes between three important unpaired categories and it can handle pseudoknots unlike the six-way classification (in other words, it can be applied to any collection of base pairs in which each base has at most one pairing partner).

The nucleotide composition of a structural element, which includes all residues in the sequence classified as such, can be calculated as the fraction of each of the four bases U, C, A and G in this element (degenerate bases, gaps, and other unknown characters are ignored in this calculation). A vector of these four fractions (for example U = 0.1, C = 0.2, G = 0.3 and A = 0.4) can be plotted in composition space, also known as the RNA simplex (Figure 1 C).

Composition can be measured in three directions. Traditionally nucleotide composition is described as GC content which is the fraction of G + C. Similarly, one can calculate the fraction of U + C and the fraction of U + G. These three fractions together give a full description of the composition of a set of residues, and form three orthogonal axes in composition space. Coordinates along these axes can be used to perform calculations on the data.

Nucleotide composition changes over evolutionary time. A single sequence in combination with its structure corresponds to five dots in composition space: one for each structural element (stem, loop, bulge and other) and one for the composition of the full sequence, which is structure independent. When this data is plotted for multiple sequences that fold into the same structure, composition space contains five distributions showing the variation in composition across the sequences (Figure 1 C). In this study, an RNA structure is described by these five distributions.

Generic patterns

For the purpose of structure prediction we are interested in compositional features that are shared by all RNA families. Therefore, building on the general observations made on RNA composition (21), we analyzed the evolutionary patterns of nucleotide composition in many RNA families. Even though the exact location and variation of the compositional distributions within the RNA simplex are family-specific, we identified several generic features, extending the observations made in several ribosomal RNA families (22).

From the multi-family survey it became apparent that the distribution of the stems has a very specific shape and location in composition space. It is an axis-like distribution along Chargaff's axis (GC axis) with large variation in GC content, and little variation in UG and UC content, describing wobble and nonstandard base pairs respectively. The distribution is often biased towards the G-U edge, caused by G-U wobble pairs, and towards the A-G edge, because base pairs involving As and Gs are more common than base pairs involving Us and Cs. The unpaired regions form very tight and distinct distributions. In many families there is a clear separation between loops and bulges and unpaired regions have a low GC content in comparison to the stems.

The quantification of these trends was not straightforward. Since the exact location and variation differs between families, absolute values to describe these properties are not applicable to all RNAs. The critical step towards describing the universal properties was the realization that real biological structures exhibit particular compositional patterns relative to other (incorrect) candidate structures for the input alignment. Thus, the underlying idea is that when an alignment is correctly decomposed according to the true biological structure, we will observe the characteristic tight distributions, but an incorrect structure will produce scatter in all distributions and overlap between the unpaired regions (this difference between 'right' and 'wrong' is visualized in Figure 1 C and its inset). Based on this idea, we designed a scoring function that can be used to evaluate candidate structures as to their ability to reproduce the expected patterns.

Scoring function design

Decomposing an alignment according to a given structure results in five compositional distributions in composition space as described above. These distributions can be quantitatively summarized by 'compositional properties', such as the mean value or variation along one of the compositional axes, e.g. the standard deviation of the stems along the GC axis. Differences between RNA structures will lead to differences in the compositional patterns, which will be observable by differences in the values of the compositional properties.

We systematically compared real biological structures with incorrect candidate structures on 31 compositional properties in three well-studied RNA families: phenylalanyl-tRNA, bacterial 5S rRNA and 16S rRNA from gamma-proteobacteria. We tested both the mean and SD along each axis in each of four structural elements (24 properties) and seven composite properties, combining multiple elements and/or axes. For each property we calculated the distribution (mean and SD) of values in several ensembles of candidate structures, the Z-score of the real structure (with respect to this distribution), and similarly the Z-scores of the 10 most accurate structures in each ensemble. Finally, we selected five properties that showed consistently low (< -0.5) or high ($> +0.5$) Z-scores with little variation across all families (listed in Table 1 left side, the reference Z-scores and weights will be explained below).

The selected compositional properties and their Z-scores observed in the training data are combined into a single scoring function. This function, when applied to a novel alignment with unknown reference structure and an ensemble of candidate structures, assigns a score to each ensemble member indicating its ability to reproduce the expected compositional patterns. Specifically, the score is the weighted root mean square deviation (RMSD) from the reference Z-scores. We experimented with two kinds of reference Z-scores (trained and extreme) and weighted versus unweighted compositional properties,

Table 1. Compositional properties

Compositional property			Reference Z[d]		Weights[e]	
Metric[a]	Structure[b]	Axis[c]	Trained	Extreme	Weighted	Unweighted
SD	S	UC	−1.270	−2	1/0.674	1
SD	S	UG	−0.985	−2	1/0.535	1
MEAN	S	UA	−0.686	−1	1/0.694	1
SD	LB	UX	−0.519	−1	1/0.806	1
MEAN	LBO	UA	+0.814	+2	1/0.719	1

[a]Metric: Standard deviation (SD) or mean (MEAN).
[b]Structural element: stem (S), loop (L), bulge (B), other (O).
[c]Compositional axis: UX indicates the average over all three axes.
[d]Reference Z-score: We experimented both with trained Z-scores as observed in the training dataset and with exaggerations of these (extreme Z-scores).
[e]Compositional properties in the scoring function can be weighted or not. Weights are one divided by the SD of the Z-scores observed in the training set.

as specified in Table 1. Trained Z-scores are the average Z-scores observed in the training set. Extreme Z-scores are rough extrapolations of the observed Z-scores, indicating that the Z-score should be very low (−2), low (−1) or very high (+2). The weights that we tested are defined as one over the SD of the observed Z-scores in the training set and basically indicate how reliable the reference Z-score is: properties showing high variation in the training set receive a lower weight. Equation 1 specifies the exact scoring function.

$$Score_m = \sqrt{\frac{1}{N}\sum_{p=1}^{N} w_p (z_m - z_{\text{ref}})^2} \qquad \mathbf{1}$$

Thus, for a given sequence alignment and ensemble of candidate structures, we first calculate the distribution of values for each property p (five in this case) in the ensemble. Subsequently, for each ensemble member m and for each property p, we calculate the Z-score of the member (z_m). We then sum the squared distance of z_m to the reference Z-score for the property (z_{ref}) weighted by w_p over all properties. Finally, we take the mean and the square root of this sum.

Scoring function application

The scoring function assigns an RMSD score to each member of an ensemble of structures. Realistic structures will receive a low score and more unrealistic structures a higher score. The scoring function thus roughly arranges the structures in the ensemble from more accurate to less accurate. Note that, a perfect correlation between RMSD score and accuracy is not necessary and would be impossible to achieve: optimizing the properties and parameters for a specific family would improve the correlation, but would diminish the generality of the scoring function. For this reason, reporting the single best-scoring structure is not informative, so we report the average accuracy of a percentage of top-scoring structures, determined by a cutoff (top-cutoff, usually 5% or 10%). In addition,

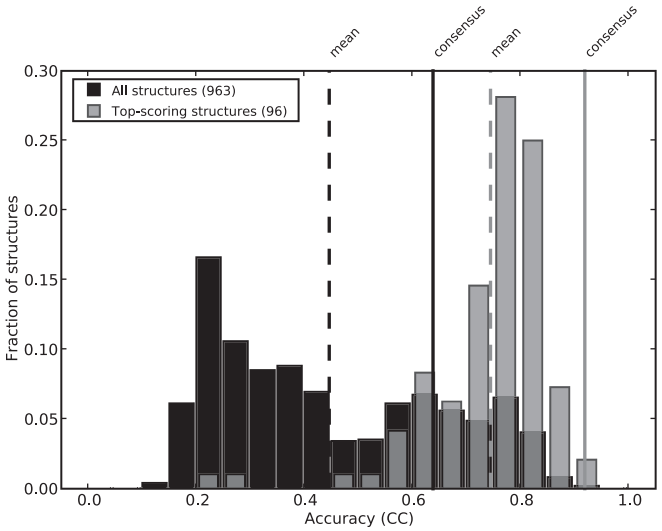

Figure 2. Application of the scoring function. Data shown is for an alignment of 20 5S rRNA sequences and an ensemble of 963 unique structures. The scoring function used the extreme Z-scores and weighted properties. Consensus structures were calculated with a top-cutoff of 10% and a bp-cutoff of 0.4. The accuracy of the structures is expressed by the correlation coefficient (CC). In both sets, the mean accuracy is indicated with a dotted line and the accuracy of the consensus structure with a solid line. The height of a bar corresponds to the fraction of structures in the corresponding accuracy range, and all bars within one set add up to 1.0.

we report the accuracy of a consensus structure calculated from the top-scoring set (see Consensus calculation section). This structure contains the most reliable base pairs from the given set of top-scoring ensemble members.

In Figure 2 we visualize the application of the scoring function using 5S rRNA, one of the families in the training set, as a specific example. Given an alignment of 20 sequences, we first sample a thousand structures from the Boltzmann ensemble (see Ensemble generation section) and keep the unique structures (963 structures). Next, we calculate an RMSD score for each structure in the ensemble. For both the full ensemble and the 10% highest-scoring structures (96 structures), Figure 2 shows the accuracy distribution of all structures in the set, the mean accuracy of the structures, and the accuracy of the consensus structure calculated from the set (see Prediction accuracy section below for details on the reported accuracies). The clear shift in distribution and the improved mean and consensus accuracy in the top-scoring set relative to the full ensemble demonstrates the ability of the scoring function to select the most realistic structures from the ensemble. Since the 5S rRNA data was used for training the algorithm, this example merely shows that the scoring function can memorize learned values. The generalization of the scoring function will be presented in the Results section.

Prediction accuracy

The accuracy of a predicted structure with respect to the reference structure can be determined both at the base-pair level and at the level of the structural classification.

The base-pair similarity involves three numbers: true positives (TP) are base pairs that occur both in the prediction and in the reference, false positives (FP) are predicted base pairs that are not in the reference structure, and false negatives (FN) are base pairs in the true structure that were not predicted. These values can be combined into three accuracy metrics (24,5,7). Sensitivity (SEN), defined as TP/(TP + FN), is the fraction of the true structure that is predicted correctly. The positive predictive value (PPV), defined as TP/(TP + FP), reports what fraction of the predicted base pairs is correct. The correlation coefficient (CC), an approximation of Matthew's correlation coefficient, combines these two values in $\sqrt{(SEN \times PPV)}$ (25). The classification similarity (CS) expresses the topological similarity between two structures. It is defined as the fraction of positions in the classification strings with identical classification. The classification similarity is relevant in this study, because the patterns of nucleotide composition are calculated from the structural classification, and a single base-pair change may have a large effect on the overall topology of the structure.

Consensus calculation

Selecting the most reliable base pairs, such as by calculating a centroid structure, or eliminating unreliable base pairs (based on their base pair probabilities) is a proven concept in RNA structure prediction (26,27). We select the most reliable base pairs by calculating a consensus structure given a set of structures. The first step is to list the frequency of occurrence of each base pair in the set of structures. The consensus structure contains all base pairs that occur with a certain frequency or higher as determined by the 'base-pair cutoff' (bp-cutoff). Eligible base pairs are added to the consensus structure one by one from high to low frequency if both the 5′ and 3′ position are not in the consensus yet. In this way, the consensus structure is free of conflicts, i.e. each base interacts with at most one other base, but might include pseudoknots. At a very high (strict) bp-cutoff only very reliable base pairs will be included in the consensus, leading to a high PPV, but a lower SEN. In contrast, a more relaxed cutoff will result in a higher SEN, but comes with the risk of including false predictions (lower PPV). By testing the consensus accuracy at different cutoff values (0 to 1, steps of 0.05) in several ensembles, we determined that a bp-cutoff of 0.4 results in the best accuracy (CC). The consensus structures in this study are therefore calculated at this cutoff value.

Ensemble generation

Given an RNA alignment, the first step in the procedure is to create an ensemble of candidate structures to be evaluated. For this task, we used the program RNAsubopt (11) from the Vienna RNA package (28) with the -p option to sample suboptimal structures proportional to their Boltzmann weights (12). For each alignment we randomly sampled a thousand structures and removed duplicates from the set. The true structure is not necessarily present in this set. The arbitrary limit of a thousand structures is chosen to limit the CPU time spent on sampling for long sequences. However, the SPuNC web interface provides three different sampling methods and gives the user control over the sample size and removal of duplicate structures.

The prediction power of our method is limited by those RNA structures present in the ensemble. Ideally the structures in the ensemble are distributed over the full range of accuracy up to 100%. For shorter sequences the current sampling methods suffice to generate an ensemble with these characteristics. However, for longer sequences (such as RNase P, and SSU/LSU rRNA) these methods produce insufficient coverage of the spectrum. For these cases we generated, in addition to the original ensemble, an extended ensemble using knowledge of the reference structure. Specifically, we constrained RNAsubopt with constraints fixing between 5% and 95% of either the unpaired or paired positions in the true structure. The extended ensemble is a merge of 500 constrained and 500 unconstrained structures (unique structures only). The purpose of applying our method to the extended ensemble is to show its potential if such extended ensembles could be generated without knowledge of the true structure.

Method performance

The input for our method is an alignment of related RNA sequences that presumably fold into the same structure. We tested the method on 15 different alignments from multiple sources (29,23,30,31), covering a wide range of RNA families (Table 2). The data are divided into three sets. The training set (T) contains three families, which are also used to derive the rules. The validation set (V) contains four families with 'new' data. The datasets in the benchmark group (B) are used for performance comparison of the algorithm. These sets were taken directly from the BRaliBase I benchmark study (5), although the sequences were realigned using MUSCLE (32) and the sequence corresponding to the reference structure was added in case no perfect match was found. The reference

Table 2. Alignments and reference structures

Set	RNA family	Seq	Len	Bps	PK
T	tRNA-PHE	20	77	21	no
T	5S rRNA	20	122	37	no
T	16S rRNA	20	1546	478	yes
V	Hammerhead rz.	8	51	14	no
V	Purine rs.	12	77	23	yes
V	TPP rs.	12	102	20	no
V	glmS rz.	13	172	52	yes
B	tRNA-PHE (H)	11	73	20	no
B	tRNA-PHE (M)	11	74	20	no
B	RNase P (H)	9	385	122	yes
B	RNase P (M)	11	431	122	yes
B	SSU rRNA (H)	11	1551	478	yes
B	SSU rRNA (M)	11	1598	478	yes
B	LSU rRNA (H)	12	2940	869	yes
B	LSU rRNA (M)	12	3197	869	yes

Set, T = training, V = validation, B = benchmark; Seq, number of sequences in the alignment; Len, length of the alignment; Bps, number of base pairs in the reference structure; PK, whether reference structure contains pseudoknots.

structures were either consensus structures that came with the alignments or base pair selections, using RNAview (33) and MC-Annotate (34), from experimentally determined structures downloaded from the RCSB Protein Data Bank (35). Pseudoknots were included in the reference, whereas noncanonical base pairs were excluded.

RESULTS

Here, we report the accuracy of our method on the 15 alignments specified in Table 2. The data are divided into a training, a validation and a benchmark group. The results are calculated using the five compositional properties specified in Table 1. In Table 3 we report the accuracy of the results obtained using extreme reference *Z*-scores and weighted compositional properties.

Scoring function identifies realistic structures

The scoring function succeeds in identifying the most realistic structures from an ensemble. This becomes immediately clear from comparing the mean accuracy of all structures in the ensemble with the mean accuracy of the set of top-scoring structures (Table 3 accuracy column 1 and 3). In the training set the average increase in accuracy is 20.7% (even 25.9% when including the extended ensembles). The method does not only perform well on the training group, as might be expected, but also generalizes to RNA families not used to derive the rules. In the validation dataset the average increase in accuracy is 13.9% with an increase above 20% for the hammerhead ribozyme and the TPP riboswitch. The results are least convincing for the benchmark set, where the average increase is only 8% (or 13.7% when using the extended ensembles). The first reason is that our method does not perform well on the high similarity tRNA dataset, because there is no enough diversity among the sequences to produce any compositional patterns. The second reason for the lower performance is the quality of the ensembles for the large molecules (RNase P, SSU rRNA, LSU rRNA). In the original ensembles, not using any information from the true structure, the most accurate structures have a CC of 0.736, 0.667 and 0.548 respectively. Possibly the sample size needs to be increased for molecules of this size, because a sample of a thousand structures covers a relatively smaller part of structure space than an equally sized sample for a shorter molecule. The method performs much better on the extended ensembles where the accuracy of the most accurate structure is above 0.9.

Improved consensus from top-scoring structures

It has been shown that the centroid structure of a set of suboptimal structures (defined as the structure with the shortest total distance to all structures in the set) is more accurate than the minimum free energy (MFE) prediction (26). Our results confirm this finding: the ensemble consensus (similar to a centroid structure) is more accurate than the MFE prediction in all datasets (Table 3 accuracy column 2 and 5). Moreover, we address the question whether the consensus structure calculated from the top-scoring structures ('top consensus') is more accurate than the consensus structure calculated from the full ensemble ('ensemble consensus'). As described above, the average accuracy of top-scoring structures is higher than the average accuracy of all structures in the ensemble, and thus

Table 3. Prediction accuracy (CC) for 15 alignments using extreme reference *Z*-scores and weighted properties in the scoring function, a top-cutoff of 0.1, and a bp-cutoff of 0.4

Alignment	Ensemble size	Accuracy (CC)					
		Ensemble mean	Ensemble consensus	Top mean	Top consensus	RNA fold	RNA alifold
tRNA-PHE	311	0.579	0.976	0.815	1.000	0.748	1.000
5S rRNA	963	0.446	0.638	0.744	0.919	0.521	0.839
16S rRNA	908 (1000)	0.379 (0.492)	0.555 (0.696)	0.465 (0.737)	0.577 (0.865)	0.388	0.465
Training: mean accuracy		0.468 (0.506)	0.723 (0.770)	0.675 (0.765)	0.832 (0.928)	0.553	0.768
Hammerhead rz.	157	0.573	0.886	0.773	0.889	0.809	1.000
Purine rs.	336	0.736	0.861	0.813	0.909	0.841	0.861
TPP rs.	795	0.500	0.700	0.725	0.886	0.505	0.868
glmS rz.	857	0.553	0.697	0.610	0.823	0.579	0.745
Validation: mean accuracy		0.591	0.786	0.730	0.877	0.683	0.869
tRNA-PHE (H)	576	0.477	0.592	0.518	0.462	0.390	0.950
tRNA-PHE (M)	547	0.573	0.837	0.865	0.976	0.611	0.976
RNase P (H)	999 (1000)	0.500 (0.576)	0.645 (0.751)	0.572 (0.609)	0.647 (0.700)	0.420	0.698
RNase P (M)	990 (1000)	0.385 (0.476)	0.595 (0.666)	0.464 (0.499)	0.503 (0.672)	0.351	0.617
SSU rRNA (H)	990 (1000)	0.448 (0.606)	0.622 (0.824)	0.514 (0.749)	0.617 (0.900)	0.474	0.650
SSU rRNA (M)	990 (1000)	0.412 (0.574)	0.680 (0.838)	0.461 (0.689)	0.621 (0.869)	0.415	0.808
LSU rRNA (H)	996 (1000)	0.401 (0.559)	0.605 (0.809)	0.403 (0.752)	0.579 (0.891)	0.426	0.687
LSU rRNA (M)	996 (1000)	0.353 (0.519)	0.601 (0.798)	0.393 (0.776)	0.644 (0.902)	0.386	0.798
Benchmark: mean accuracy		0.444 (0.545)	0.647 (0.764)	0.524 (0.682)	0.631 (0.796)	0.434	0.773

If we applied the method to an extended ensemble, using the true structure, in addition to the normal ensemble generated by RNAsubopt, the results are displayed between parentheses. Average values between brackets use the results for extended ensembles where possible.

we expect the consensus from this set to be also more accurate than the consensus from the full set.

For the majority of datasets in our test the consensus from top-scoring structures identified by the scoring function is significantly more accurate than the ensemble consensus (Compare accuracy column 2 and 4 in Table 3). The average increase is 10.9% (15.8%) and 9.1% in the training and validation set respectively. The improvement is most pronounced in datasets with shorter sequences and a diverse ensemble, such as for 5S rRNA, the TPP riboswitch and the glmS ribozyme. For tRNA-PHE, the hammerhead ribozyme, and the purine riboswitch there are fewer structures in the ensemble which are all highly accurate. For these datasets the top consensus has about the same accuracy as the ensemble consensus, because little or no improvement over the ensemble consensus was possible in the first place. Again we obtain mixed results on the benchmark dataset. Performance is poor on the high-similarity tRNA alignment, because of reasons outlined above. For the three large molecules (RNase P, SSU rRNA and LSU rRNA) we note that the ensemble consensus is more accurate in case of the original ensemble generated by RNAsubopt, and that the top consensus is better in case the extended ensemble is used. The consensus calculation seems to be mainly hampered by the composition of the ensemble, since the scoring function rejects the least accurate structures for both ensembles and changing the bp-cutoff did not affect the results. This underlines the fact that our method would clearly benefit from improved ensemble-generation methods for these large molecules.

Accuracy compared to other methods

It is important to compare the performance of a novel structure prediction method to that of existing approaches. A full benchmark study is outside the scope of this work, but we provide two ways to put the accuracy of our method in perspective. First, we applied our method to the datasets provided by the BRaliBase I benchmark study (5), and thus the accuracies reported here can be compared to those in the original benchmark. However, this comparison can be approximate at best, because the sequences are realigned using MUSCLE (32), we use the pseudoknotted structure as reference for all datasets, we do not adjust the reference structure to the alignment via the consistency criterion, and we do not distinguish inconsistent, contradicting and compatible false positives. To provide a more direct framework in addition to this approximate comparison, we applied both RNAfold (28,36,37) and RNAalifold (13) to all alignments in our data collection (Table 3 last two columns). RNAfold predicts the minimum free, energy structure for a single sequence. This type of method achieves in general lower accuracies than multiple-sequence comparative approaches, and we thus expect our method to make more accurate predictions than RNAfold. In contrast, RNAalifold is a top-of-the-line prediction method for multiple-sequence alignments combining thermodynamics and covariation. RNAalifold outperformed most other methods tested in the BRaliBase I benchmark and is thus a good target for our method.

The most remarkable result is that the consensus from top-scoring structures is more accurate than the RNAalifold prediction for several families: the 5S rRNA, the glmS ribozyme and both riboswitches. Our method matches the results of RNAalifold for the tRNA alignment in the training set and the medium-similarity tRNA alignment in the benchmark set. On the hammerhead ribozyme our method achieves good accuracy (0.889), but not as good as RNAalifold (1.0). For the longer sequences our method is hampered by the insufficient coverage in the ensemble, as indicated above. We do demonstrate however that our method has the ability to enhance structure prediction for long molecules when diverse enough ensembles can be generated, but that is of limited practical value at the moment. It emphasizes the need to create more diverse ensembles that include structures of higher accuracy.

Additional results

Table 3 contains the results using the extreme reference Z-scores and weighted addition in the scoring function, because they produced the most accurate predictions. We provide the results for the other three method settings through the SPuNC website. The performance of the method was actually quite consistent across the different settings. The accuracy difference between trained and extreme references was slightly larger than between weighted and unweighted addition. The positive predictive value (PPV) is generally higher than the sensitivity (SEN). The results based on classification similarity (CS) are in agreement with those based on correlation coefficient (CC). Further experiments are necessary to find the optimal Z-score references and weights in the scoring function.

For completeness we also calculated the correlation between the RMSD score and the accuracy (specifically CC) for each dataset, even though the overall correlation does not influence the prediction results as long as low-scoring structures are highly accurate. The results are available through the web interface. In general the observed correlations between score and accuracy (CC) are relatively weak but highly significant (strongest correlation coefficient observed is -0.821, $r^2 = 0.67$).

DISCUSSION

The structure-prediction method introduced in this paper uses evolutionary patterns of nucleotide composition to assess the quality of candidate structures for a multiple-sequence alignment. At the heart of our method lies a scoring function that assigns a score to each ensemble member, reflecting its ability to produce the patterns observed in biological structures. The performance test of our method showed that the scoring function is able to identify the most realistic structures in an ensemble. In addition, the prediction accuracy was further improved by calculating a consensus structure from the top-scoring ensemble members. Our method performed well on a

wide variety of RNA families, including several novel types of RNA such as riboswitches and ribozymes.

The SPuNC algorithm takes full advantage of the topology of the molecule by exploiting signal from distinct unpaired regions in addition to that from paired regions. Also beneficial is the insensitivity to the exact alignment at each sequence position as long as the residues are classified in the correct structural element. The negative side effect of this feature is that the method can not distinguish between two structures that differ in their base pairs but result in the same classification string, but no practical consequences hereof were observed. Our method has the potential to incorporate pseudoknots and noncanonical base pairs, although adjustments to the ensemble-generation methods and the scoring function would be necessary to fully benefit from these features.

The current scoring function was designed using the compositional patterns as the only source of evolutionary signal. The purpose was to demonstrate that simple, biologically acceptable rules—stems vary mostly in GC content and are GC-rich, unpaired regions are GC-poor and have rather consistent composition over evolutionary time—can be used for structure prediction. Within this framework, the five selected compositional properties have been shown to result in accurate predictions, in some cases even outperforming existing prediction methods. However, further research is necessary to optimize the points of reference and weights in the scoring function.

Rather than capitalizing on this information source in an isolated fashion, future efforts should aim at combining compositional patterns with other sources of information. A promising direction is the incorporation of the scoring function, along with other observations, in a Bayesian framework such as BayesFold (15). The Bayesian inference approaches are becoming more feasible with advances in computational power and already have been shown to be valuable in RNA structure prediction (9). The scores based on the patterns of nucleotide composition could potentially function as prior probabilities on the ensemble members or as a set of observations used to update the posterior probabilities.

Along with optimizations in the scoring function, our method will benefit from improved ensemble-generation methods. The statistical sampling of RNA secondary structures is a major improvement over minimum free energy and near-optimal structure predictions in terms of the coverage of the energy spectrum, but further improvements can be made (9). As becomes clear from our study, a lack of accurate structures in the ensemble, as observed for long sequences such the RNAs in both ribosomal subunits, limits the predictive power of our method. A possible strategy to calculate improved ensembles is to collect all viable base pair regions (with high probability) and to combine them into new structures (which might include pseudoknots in addition). Another option to improve the performance of our method on longer sequences, which does not rely on novel methods for ensemble generation, would be to apply a genetic algorithm: start with a random sample of structures, score each ensemble member and select the most realistic structures (low RMSD scores), let them 'reproduce' into a new ensemble by constrained folding, repeat, and progress towards the true structure in this manner.

In conclusion, evolutionary patterns of nucleotide composition should be added to the toolbox for structure prediction. The presented method reaches encouraging accuracies on a wide variety of RNA families. Especially the good result on the riboswitches and ribozymes support the value of this method, because many more of these novel RNAs await discovery and structural characterization.

ACKNOWLEDGEMENTS

We are grateful to Sébastien Lemieux for providing a standalone implementation of the MC-Annotate algorithm. We also want to thank Dave Mathews and the members of the Heringa Lab and Knight Lab for their feedback on our structure prediction method.

FUNDING

This work is supported by the Centre for Medical Systems Biology (CMSB). The CMSB is supported by a Genomics Centre of Excellence grant from the Netherlands Genomics Initiative, which is funded by the Dutch government.

Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
2. Tinoco, I. Jr., Uhlenbeck, O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
3. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
4. Major,F. and Griffey,R. (2001) Computational methods for RNA structure determination. *Curr. Opin. Struct. Biol.*, **11**, 282–286.
5. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
6. Mathews,D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–532.
7. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.
8. Reeder,J., Hochsmann,M., Rehmsmeier,M., Voss,B. and Giegerich,R. (2006) Beyond Mfold: recent advances in RNA bioinformatics. *J. Biotechnol.*, **124**, 41–55.
9. Ding,Y. (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA*, **12**, 323–331.
10. Gutell,R.R., Lee,J.C. and Cannone,J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301–10.
11. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–65.
12. Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.
13. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
14. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification

constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci.*, **101**, 7287–7292.
15. Knight,R. Birmingham,A. and Yarus,M. (2004) BayesFold: rational 2° folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. *RNA*, **10**, 1323–36.
16. Giegerich,R. Voss,B. and Rehmsmeier,M. (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.
17. Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.
18. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
19. Lemieux,S. and Major,F. (2006) Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res.*, **34**, 2340–2346.
20. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
21. Schultes,E., Hraber,P.T. and LaBean,T.H. (1997) Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA*, **3**, 792–806.
22. Smit,S., Yarus,M. and Knight,R. (2006) Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, **12**, 1–14.
23. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R. D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
24. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
25. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
26. Ding,Y., Chan,C.Y. and Lawrence,C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.
27. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
28. Hofacker,I.L., Fontana,W., Stadler,P.F. Bonhoeffer,L.S., Tacker,M. and Schuster.P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
29. Szymanski,M., Barciszewska,M.Z., Erdmann,V.A. and Barciszewski,J. (2002) 5S Ribosomal RNA Database. *Nucleic Acids Res.*, **30**, 176–178.
30. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**(Database issue):D139–D140.
31. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**(Database issue):D121–D124.
32. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
33. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
34. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
35. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
36. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–48.
37. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–19.