Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Predicting RNA consensus stems through unsupervised clustering of unaligned sequences

E. Rogers [1], and C. Heitsch [2,*]

[1] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, 30332, USA and
[2] School of Mathematics, Georgia Institute of Technology, Atlanta, 30332, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Improvements in secondary structure prediction accuracy for a single RNA sequence, notably through Boltzmann sampling, have not been realized for multiple homologous ones; consensus structure prediction remains a significant challenge in computational biology. To close this gap, two insights are critical. One is finding the right balance between improvements in precision versus recall. Another is resolving conflicting base pairing signals through an appropriate level of structural granularity. Together, these can achieve very high accuracy predictions of native structural elements for an RNA family.
Results: `ConsensusStems` leverages RNA profiling and noise-sensitive clustering to extract common base pairing regions from Boltzmann samples for related sequences. This focus on more general structural element prediction is very successful; the cluster centroids output recover the native stems in 7 of 11 Rfam families tested, with median sequence lengths up to 300 nucleotides and structural complexity up to 10 stems. Overall, the (avg, std) centroid accuracies are: precision = (0.96, 0.08), recall = (0.95, 0.10), and approximate Mathews correlation coefficient = (0.95, 0.08). Thus, `ConsensusStems` is a important contribution to advancing RNA folding prediction.
Availability: A demonstration webserver is online at `rnaconsensus.math.gatech.edu`. Code can be downloaded for general use via `https://github.com/gtfold/ConsensusStems`.
Contact: heitsch@math.gatech.edu

1 Introduction

Accurate prediction of the common native structure for homologous RNA sequences is an open problem in computational biology [1, 2, 3, 4, 5, 6]. Given current interest in 'noncoding' RNA's role in gene splicing, editing, and regulation, this challenge has taken on new urgency in recent years. In particular, since experimental determination of 3D conformations is still time-consuming and labor-intensive, function is most often inferred from computational predictions of RNA secondary structures. Thus, improved prediction of the noncrossing, canonical base pairings common to related RNA sequences is essential to providing new functional insights.

Although sequence alignment is a typical starting point for consensus structure prediction, such methods face the difficulty that RNA pairings (i.e. complementarity of two positions i and j) are much more strongly conserved than individual nucleotide identity. In contrast, aligning RNA secondary structures rather than primary sequences can produce a more complete and accurate consensus prediction.

It is not obvious, though, which structural elements to align; minimum free energy (MFE) structures [7, 8], base pair probabilities [9, 10], sampled helices [11], or even all stable helices [12, 13, 14, 15, 16] have been tried. The key is striking the right balance; too little information, and the recall limitations of the original false negative predictions cannot be overcome. Too much, and the resulting precision is dominated by false positive predictions. Our novel `ConsensusStems` approach achieves a good balance by leveraging the predictive power of Boltzmann sampling [17] filtered through the denoising achieved by RNA profiling [18].

Stochastic sampling from the Boltzmann ensemble for a single RNA sequence is state-of-the-art in secondary structure prediction since it efficiently provides the most comprehensive folding information [19, 20]. Hence, by starting the `ConsensusStems` pipeline with Boltzmann

sampling, false negatives are minimized to the maximum extent possible under the current nearest neighbor thermodynamic model (NNTM).

False positives are then filtered by RNA profiling which extracts the structural 'signal,' i.e. the set of high frequency maximal helices known as features, from each noisy sample. This set of features is a robust signature of each Boltzmann ensemble that is easily compared between related sequences to highlight similarities and differences [21].

Such comparisons are the core of `ConsensusStems`' methodology since the NNTM prediction for each sequence is only partially correct on average [22]. However, these partial signals almost always include complementary information. Thus, false negatives are minimized by consolidating the individual sets of features over the entire family. This recovers the native helices, albeit with considerable noise.

To improve precision as well as recall, the false positives are filtered by a noise-sensitive clustering algorithm [23]. The resulting clusters are the structural 'alignment' produced by `ConsensusStems`. Specifically, each cluster is denoted by a representative centroid; this is the consensus stem prediction that those 5' and 3' segments interact exclusively with each other. Each output cluster also has an associated list of supporting sequence/feature pairs, consolidated into a sequence-specific stem. These are the regions of the individual sequences understood to be structurally 'aligned' with the consensus stem.

This focus on predicting regions of interaction, i.e. the forest rather than the trees, is critical to `ConsensusStems`' success. NNTM optimization often predicts very similar helices, presenting a conflicting base pairing signal that is challenging to resolve accurately. However, at a lower level of structural granularity, competitors tranform into allies, sending a clear true positive signal for a native pairing region — which could well a better reflection of physical reality given the stochasticity of biological systems. As will be shown, our stem abstraction clarifies base pairing patterns and increases prediction accuracy of consensus structure for multiple sequences.

By minimizing false positives as well as negatives, `ConsensusStems` achieves remarkable accuracy at this level of granularity. Tests on a diverse set of 11 different RNA families demonstrate that this new method predicts the native consensus stems with an average accuracy over 95%. It is 100% accurate 66% of the time, and the remaining four families were 89%, 89%, 86%, and 75% accurate. Hence, this approach is a major step forward in resolving the consensus structure prediction problem.

2 Approach

Our goal is predicting common regions of structural interation, called consensus stems. Results will show that while the signal at the base pair level is usually messy, viewing the same data at a lower granularity yields clear and accurate predictions.

More precisely, we generalize the standard (i, j, k) notation for helices, which denotes k consecutive base pairs closed by (i, j), to the stem notation (i, j, k, l). Stems have an extra coordinate l, since the length of the 5' region (k) may not be equal to the length of the 3' region (l). The stem coordinates (i, j, k, l) thus denote that the regions $[i, \ldots, i + k - 1]$ and $[j - l + 1, \ldots, j]$ interact exclusively with high probability.

Accurate prediction of native consensus stems requires dealing with false positives (or FP, the non-native predictions) and false negatives (or FN, the native elements not predicted). `ConsensusStems` does this in two rounds: (1) by using Boltzmann sampling to deal with FN, and profiling to deal with FP; and (2) by using multiple, normalized sequences in the same family to deal with FN, and clustering to denoise the composite family data to deal with FP.

Proof-of-principle for this approach is given in this section by analyzing a set of tRNA sequences. Although the native cloverleaf is well-known, minimum free energy (MFE) prediction accuracies for an individual sequence can range from a high of 100% to a low of 0% [18]. Hence, a consensus method which starts with a single MFE structure per sequence is likely to have a prohibitive number of FN. Hence, current best practice is to stochastically sample a set of predicted structures (typically of size 1000) from the Boltzmann ensemble for that sequence [17].

The Boltzmann sample, however, contains many more structures than the native, and necessitates post-processing to deal with FP. RNA profiling has been demonstrated to extract the key structural information from a sample [18], yielding a clear, concise — although not necessarily fully correct — signal from the noisy ensemble. High level patterns are readily seen because of the key use of abstraction, which improve computational accuracies significantly [22].

Helix	Triplet	Frequency
1	1 71 7	1000
2	24 44 7	834
3	21 47 2	760
4	46 66 7	544
5	15 64 4	340
6	50 60 4	193

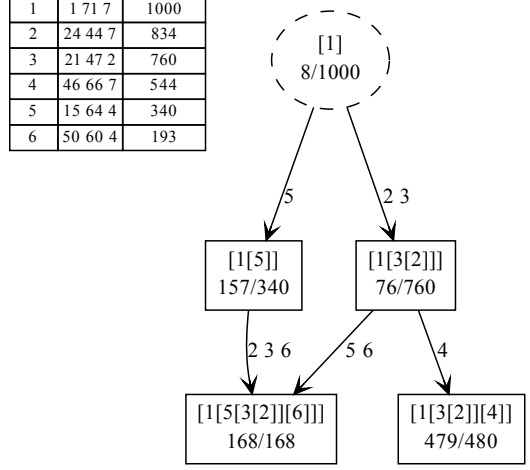


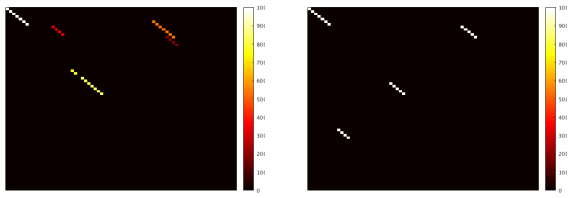
Fig. 1. Profiling output for $T.brucei$. Maximal helices are listed in descending frequency with (i, j, k) triplet and corresponding index number. Profiling uses a maximum average entropy threshold to truncate the distribution, returning only the most common helices as the selected 'features.' Each node in the graph gives a profile, i.e. a maximal combination of features, with brackets indicating nesting relationships. The ratio gives the number of sampled structures with exactly that profile (numerator) over the number with at least those features. Nodes are related as a Hasse diagram under the partial ordering of set inclusion, with edges labeled by the difference. For this sequence, the most frequent profile is $[1[3[2]][4]$ which was sampled 479 out of 1000 times and is nearly the native structure. The FP is feature #3 at $(21, 42, 2)$ with estimated probability of 76.0%. The FN of $(10, 24, 4)$ is the 11th most frequent helix with a probability of only 5.9%.

2.1 Profiling *Trypanosoma brucei* lysine tRNA

We use a representative tRNA sequence *Trypanosoma brucei* (Accession Z11880.1/124-195) from the Rfam seed alignment [24] to see that profiling successfully extracts enough native signal from a sample to be our starting point.

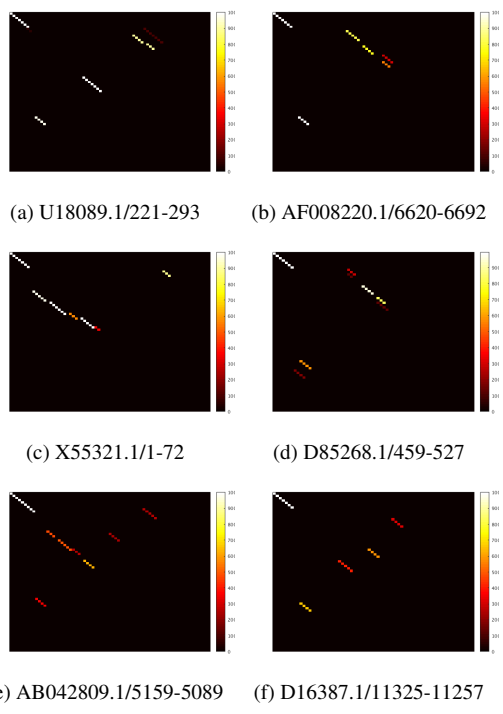
Figure 1 shows the list of features, i.e. high frequency maximal helices, and summary profile graph of *T.brucei*. By consolidating substructures with high similarity and truncating the long, noisy tail of the frequency distribution, profiling produces a clear, concise, and stable structural signal for this ensemble.

As seen in Figure 2, this signal contains a significant portion of the native cloverleaf; 3 of 4 TP helices are high frequency substructures in the *T.brucei* ensemble. However, without additional information, it is not possible to distinguish the true from false positives among the 6 features output by profiling. Likewise, although the FN helix is present in the whole sample with almost 6% frequency, there is no reason to identify



(a) Z11880.1/124-195 (b) Native

Fig. 2. Two dotplots for $T.brucei$: the 6 features extracted by profiling from the Boltzmann sample (left) and the native secondary structure (right). A base pair between positions i and j corresponds to a box with coordinates (i, j) in the (x, y) plane, with $i < j$. On the left, the colors correspond to a frequency heatmap from red/least to white/most. It is clear that the native structural signal is partially present in this ensemble, albeit noisy and incomplete.



(a) U18089.1/221-293 (b) AF008220.1/6620-6692

(c) X55321.1/1-72 (d) D85268.1/459-527

(e) AB042809.1/5159-5089 (f) D16387.1/11325-11257

Fig. 3. Heatmaps for the features of six tRNA sequences. Each square (i, j) corresponds to the base pair (i, j), with the frequency of the base pair (as measured by frequency of the maximal helix to which it belongs) reflected in the color, from the highest frequency (white) to the lowest (red). While not all the sequences have the native cloverleaf structure in the features (see Figures 2), all have at least some native helices as a feature.

this particular helix from the 44 others that are truncated from the full distribution.

This demonstrates that, although the complete native structure is seldom present with high probability in a single Boltzmann ensemble, there typically exists a significant amount of *partial* information. Moreover, as shown below, different sequences capture different parts of the native structure among their features. Hence, a consensus structure can be recovered by agglomerating the helix signal from homologous samples to produce a composite signal with (very) high accuracy.

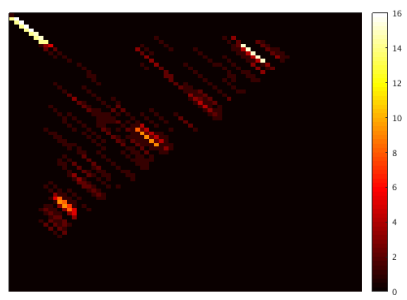


Fig. 4. Representation of the 2-D normalized grid for tRNA to which all high frequency helices are mapped. Each helix (i, j, k) is mapped to its corresponding grid points, augmenting frequency counts for cells $(i, j), \ldots, (i + k - 1, j - k + 1)$. The frequency of each cell is represented by color, with black indicating zero counts, through red up to white, the highest count. Note the general shape of the tRNA cloverleaf structure, a closing stem encompassing three stacks (see Figure 2), is present though somewhat blurry.

2.2 Leveraging information from homologous sequences

We now illustrate that (1) profiling extracts part of the native structure from a sequence reasonably consistently; and (2) the partial signals from a large enough set of homologous sequences are complementary. Hence, the common native structure can be recovered by amalgamating individual sequences' features.

Figure 3 shows six additional tRNA sequences with a typical range of features. Although it is possible that the native structure is fully predicted, as for *U18089.1*, most have only a partial signal. However, the collective signal is sufficient to recover the consensus stems, as seen in Figure 4.

Figure 4 is a composite heatmap of 30 randomly chosen tRNA sequences (including those from Figure 3). The features for each sequence are are overlayed on top of each other after normalizing for differing lengths. Figure 4 clearly contains the complete tRNA structure.

The consensus helices of Figure 2 are clearly seen in Figure 4, albeit with a signficant amount of noise. Thus, recovering the native tRNA stems at this point is matter of filtering out the FP 'noise'; this insight is the governing principle of `ConsensusStems` in using clustering to automate the extraction of the native signal. This correspondence between the consensus stems and the strong composite ones is found in general for RNA families, as the next section shows.

2.3 Examining Rfam families

We demonstrate that tRNA is not unique by considering eleven families from Rfam [24] that span the range of sequence lengths. While many of these sequences have been used as test sequences in the literature (Table 1), no standard test set has been developed to benchmark consensus structure methods. These eleven families were selected to span the range of sequence lengths available from the set of Rfam families with known structure. Sequences from the families were randomly chosen from the seed alignment, sampled and then profiled to obtain their features.

Figure 5 demonstrates that while a percentage of native helices in each family are low frequency to non-existent, the majority are strongly present in the sample as features. This continues our findings that a significant though partial native signal exists within the features recovered by profiling for each sequence.

Furthermore, we show that the composite signal of multiple Boltzmann samples recovers the native structure with noise for all families tested. Hence, our strategy can be summed in two parts: consolidate the signal

Family	Lengths		Sequence		Structure	
	Med	Range	Ident	Num	Helices	Stems
tRNA+	72.5	22	46	30	4	4
THF	98	21	62	25	5	3
TPP*+	105	89	56	29	6	5
5S*+	117	18	60	29	11	3
FMN	138	76	72	28	5	5
U1	162	18	65	25	12	5
ykoK+	168	25	61	26	13	5
glmS+	173.5	70	60	18	6	4
IRES cripavirus	199.5	36	53	24	8	4
IRES HCV*	243.5	185	86	24	20	10
metazoa SRP	298	27	70	23	18	7

Table 1. Information for 11 test families, including average length, average family pairwise sequence identity in percent, number of seed sequences analyzed, and number of helices and stems in Rfam's secondary structure. Sequences from each family were randomly chosen, and each family was chosen to span the range of lengths available from the set of Rfam families with structures. An asterisk indicates families included in the MASTR data set [25], a popular benchmark limited to shorter sequences. A plus sign indicates a family used by RNAscf [15], a method also working with helices.

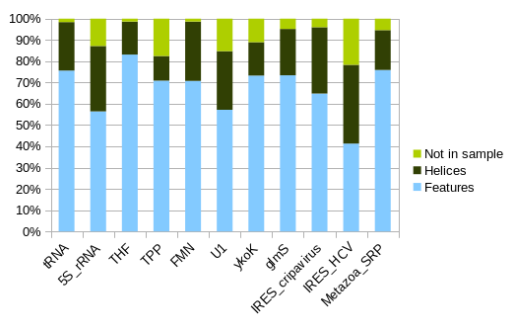


Fig. 5. Data for 11 Rfam families, indicating within a family the number of native helices with multiplicity for which it is a high frequency feature (light blue), a low frequency helix (dark green), or not present in the sample at all (light green). For k sequences in a family whose native structure has n native helices, the total number of native helices categorized is nk. For most of the families, the majority of native helices are high frequency features. Only a fraction are not present in the sample at all.

by agglomerating individual Boltzmann samples to address FNs, and filter the signal through clustering to address FPs.

3 Methods

Figure 6 illustrates the general steps of `ConsensusStems`, which starts with Boltzmann samples for a set of related sequences, and ends with a list of clusters composed of sequence/feature(s) pairs. Each cluster is characterized by a centroid stem, the generalized (i, j, k, l) coordinates defining pairing regions, which is the final prediction by `ConsensusStems`.

1. *Generate* a Boltzmann sample for each sequence in the family.
2. *Profile* each sample to get the sequence specific features.
3. *Cluster* all the feature to get the potential consensus stems.
4. *Refine* each cluster by adding in features from missing sequences.
5. *Validate* each cluster by assessing overall base pairing support for the region across sequences.

6.a. If there are new clusters found, *resample* the structures with a constraints file generated from the clusters; go back to step 2.

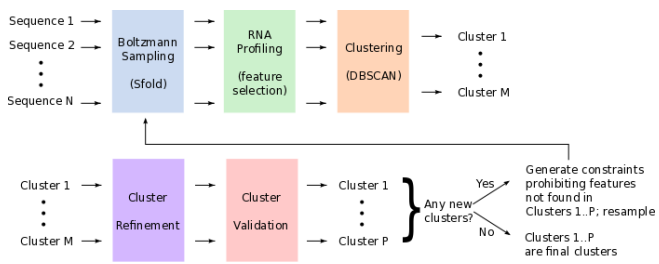


Fig. 6. The RNA `ConsensusStems` method. Each sequence in a family is sampled and profiled, yielding a set of features that are then normalized and clustered. The initial clusters are then refined by searching for potential additions from missing sequences. Finally, they are validated by assessing each sequence's possible base pairings in the region of interest. If any new clusters are identified, then the final clusters are used to make a constraints file that feeds back into Boltzmann sampling.

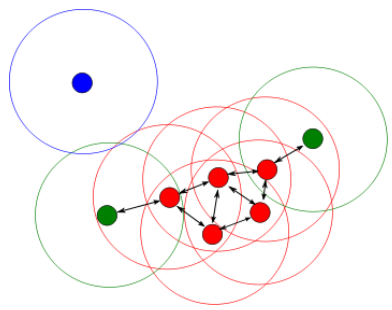


Fig. 7. Schematic of the clustering method DBSCAN. The radius ϵ is denoted by the circles, whose colors correspond to the point on which it is centered. Each red point has $P = 4$ points within its radius (including itself) and is a core point. All points in a circle's radius are reachable from the core point, and belong to the same cluster as the core point. Each of the green points are reachable from a red core point, and hence are part of its cluster, but are not themselves core points. The blue point is neither a core point nor reachable from a core point; it is considered noise and not part of a cluster.

b. If there are not any new clusters found, *terminate* the procedure.

3.1 Step one: generate a Boltzmann sample

`Sfold` 2.2 was used with default settings to generate a standard Boltzmann sample of a thousand structures for each sequence in a family. Although various programs exist that implement Boltzman sampling, `Sfold` was used because of its option to sample with constraints, which option will be used later in `ConsensusStems`.

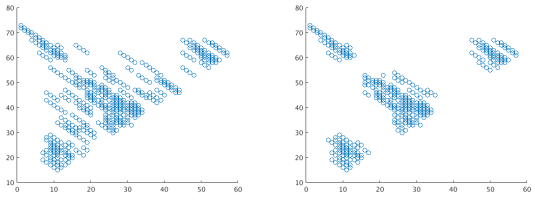
3.2 Step two: profile the samples

The output of RNA profiling gives a list of all the features with its (i, j, k) coordinates (see Figure 1), which denote a set of consecutive base pairs $(i, j), (i + 1, j - 1), \ldots, (i + k - 1, j - k + 1)$.

3.3 Step three: cluster the features

While many clustering methods exist [26], one is necessary to filter out potential 'noise' in order to recover the native signal (see Figure 4).

We chose to use the clustering method `DBSCAN` (Density-based Spatial Clustering of Applications with Noise) with its inherent concept of noise, one of the most commonly used and cited clustering algorithms [23, 27]. This algorithm classifies data points as a *core point*, *reachable point*, or as *noise* (Figure 7), using only two parameters: a radius ϵ and a minimum number of points P.



(a) All tRNA features (b) After initial clustering

Fig. 8. Figure on the left is the normalized space of all the features of tRNA. The figure on the right represents all the features found to be in a cluster after initial clustering of step two; unsupported features have been filtered out as noise.

- A *core point* has P points (including itself) within a radius of ϵ, and is part of a cluster.
- A *reachable point* is within ϵ of a core point, and is part of the same cluster as the core point.
- A point is *noise* if it is neither a core point nor a reachable point, and is not assigned a cluster.

The steps for denoising the signal and producing a clustering of all the features is thus:

1. Cluster the 1D distribution of sequence lengths to find the radius ϵ
2. Construct normalized 2D space of all features
3. Cluster the 2D distribution of base pairs in all the features
4. If majority of points are noise, adjust P and run step 3 again

3.3.1 Cluster sequence lengths to determine ϵ

For families with very similar lengths, homologous helices are located close to each other in the normalized space. Hence, the radius ϵ should be small to avoid including FP in the cluster. Conversely, for families with a wide range of lengths, a more generous radius ϵ should be used to avoid excluding FN from a cluster too narrowly defined.

We run a 1D DBSCAN on the distributions of lengths using initial $P = \frac{N}{4}$, with N being the total number of sequences in a RNA family. We begin with $\epsilon = 0$, to allow for the case that all sequences are of the same length. If the majority of the lengths are then classified as 'noise', we increment ϵ by 1 and run DBSCAN again with the new parameters. We stop as soon as we have found the lowest ϵ which produces a clustering of lengths in which the majority are either core or reachable points.

3.3.2 Construct the normalized 2D space

Many sequences in the same family have differing lengths, often by a significant amount due to insertions or deletions in the sequence. Hence, all feature coordinates are normalized to the median length in order to embed them onto a common clustering space. Given a sequence S with length d belonging to a family of median length n, the coordinates of the features of S are multiplied by $\frac{n}{d}$ and rounded up.

In an nxn grid, the (i, j) square is associated with the base pair (i, j). Each square (i, j) has an associated frequency indicating the number of features containing the (i, j) base pair; computationally, the (i, j, k) feature helix causes the frequency of squares $(i, j), \ldots, (i + k - 1, j - k + 1)$ to each be augmented. This grid will be the target of the clustering method. The frequency of all features of tRNA is emphasized in Figure 4, while the location is emphasized in Figure 8.

3.3.3 Cluster the 2D features

Before running DBSCAN to cluster the points, the distance metric needs to be considered in light of the biology of insertions/deletions (called 'indels'). The distance between two coordinates can be considered the number of indels necessary to shift one into the other. Hence, the Manhattan distance metric is used: a point (i', j') is considered within a radius ϵ of (i, j) if $|i - i'| + |j - j'| \leq \epsilon$.

Since an indel (relative to the median) of d nucleotides can occur in the sequence before i, this will cause an offset of d relative to both i and j, with an overall distance of $2d$ from the original (i, j) point. To allow for this, we set a new $\epsilon' = 2\epsilon$ for the 2-D clustering.

3.3.4 Adjust parameters if needed

Clustering is run with parameters ϵ' and P on the 2D grid. If the majority of points on the 2D grid is labeled as 'noise', clustering is run again with ϵ' held constant and P decremented by 1. This cycle repeats until a $P' \leq P$ is found such that the majority of the points are either core or reachable points. At this point, the initial clustering of points is complete (Figure 8), with each cluster being a list of sequence/feature(s) pairs. This list can be considered an implicit alignment, with the features of each sequence in the cluster structurally aligned with each other.

3.4 Step four: Cluster Refinement

Each cluster is refined by searching all the missing sequences for potential cluster members. The initial clustering could miss these features due to large differences in sequence lengths that even normalization does not fully address. Hence, in order to refine the cluster and fill in the gaps, each cluster:

1. Finds all the sequences not present in the cluster
2. Determines coordinates of each missing sequence's search window
3. Identifies all features that fit the parameters of the search window
4. Calculates probabilities of implied indel positions of found features
5. Adds feature to cluster if probability is over a threshold

3.4.1 Find missing sequences

Each cluster is associated with a list of sequence/feature pairs. All sequences not represented in the cluster are the missing sequences, which are examined one by one.

3.4.2 Determine search window

Each missing sequence is scanned for features that could plausibly be structurally aligned with the others in the cluster. The idea of plausibility is rooted in the observation that the relative length of a sequence can be roughly correlated with the relative displacement of its stem. Namely, a longer sequence tends to have 'later' stem coordinates than those of shorter sequences in the alignment, and vice versa for shorter sequences. This insight helps us to define a window for each sequence in the unnormalized, original coordinates of the sequence, in which the native stem is expected to be located.

More specifically, the location of each sequence's search window is based on the sequence's expected number of indels. This can be calculated from the displacement of the length of the sequence relative to the cluster.

To obtain a reference point for a cluster, its centroid is calculated as the median coordinates of all constituent stems (i, j, k, l) in the cluster. We also calculate the median length d_c of all sequences included in the cluster.

Given a missing sequence S_m of length d_m, any potentially homologous helix of S_m is likely to be located an offset of $\delta d = d_m - d_c$ away from the centroid stem. This window ranges from an offset of $1.25\delta d$ to $-0.25\delta d$, for reasons explained below.

As an example, consider a cluster of median length $d_c = 75$ and centroid $(10, 30, 5, 6)$, and a missing sequence S_m of length $d_m = 80$. S_m can be expected to have at least $\delta d = d_m - d_c = 5$ insertions

not present in sequences of length d_c. Since up to $\delta d = 5$ insertions could occur before the centroid, a window of up to δd after the centroid needs to be searched. Actual data demonstrates that a few indels often occur in the opposite direction (i.e. deletions in our example). To allow for this, up to 25% of δd (rounded up) is 'budgeted' in the window in the opposite direction. The window offset needs to be increased from δd to $1.25\delta d$, in order to balance out the $0.25\delta d$ in the opposite direction. For our example, this means that to find potential homologous features to the original centroid pairing region $[10, 14]$ with $[25, 30]$, we would look for those pairing the regions $[10 - 0.25\delta d, 14 + 1.25\delta d]$ with $[25 - 0.25\delta d, 30 + 1.25\delta d]$.

3.4.3 Identifying potential features

For every missing sequence, its Boltzmann sample is scanned for features forming pairings between the two regions of the determined window.

3.4.4 Calculate probabilities

The search window may identify many FP candidates that still need to be filtered out, especially if the difference in sequence lengths is large. Hence, additional filtering is done based on the candidate feature's implied locations of indels that enable its structural alignment to the cluster.

Features are only accepted into a cluster if the implied number of indels falls within acceptable boundaries of plausibility, e.g. if the ith coordinate is shifted by m indels compared to the centroid, we expect the jth coordinate to be shifted by at least m indels.

More precisely, the occurrances of indels can be modeled as a Poisson process, assuming the rate of indels as independent and identically distributed for simplicity. (We assume that if the centroid coordinate at i_c and its putative homolog in a sequence is at position i_m, then there have been $m = i_c - i_m$ indels. There can be many more, of course, as long as the net displacement equals m, but likelihood of this is far less.) In the spirit of the simple gap penalty, a uniform rate of indels is assumed for simplicity, given by $k = \frac{\delta d}{d_m}$, where δd is the absolute difference between a sequence's length d_m and the median length of the cluster d_c. Given this rate , the Poisson probability of observing $\lambda = m$ indels over p nucleotides is calculated:

$$P(\lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

The Poisson probability is then normalized over the largest probability, where $\lambda = pk$, which is the rate of indels k times the length of interest p.

Poisson probabilities are calculated for three intervals: the number of indels implied before the i_m nucleotide, between i_m and j_m, and after j_m to the end. Because a centroid stem can be composed of multiple helices, not all the interval probabilities will necessarily be favorable. Thus, we look probabilities for all three intervals, and eliminate any candidates which have poor scores of less than 50% for all of them. Any remaining candidate that is also a feature is then included in the cluster.

3.5 Step five: Cluster validation

Each cluster passes through a final validation step; clusters that have broad support across sequences are validated, while clusters having support in only a few sequences are deleted. Support from each sequence is determined by the number of base pairs in the MFE structure of the search window. This represents the best, most energetically favorable scenario. If the number of MFE base pairs in a sequence's window is less than the centroid, support from the sequence is considered weak.

A cluster is validated with the following steps:

1. Set the total score T to the number of sequences in the initial cluster before refinement
2. For each missing sequence from the initial cluster, run MFE folding on the window defined in Section 3.4.2
3. Count the number base pairs g in the MFE substructure
4. Score the sequence by comparing g with the number of base pairs in the cluster centroid
5. Add each sequence's score to T
6. If total score T is above zero, keep the cluster; else, delete the cluster

The base pairings in a window is assesed by running minimum free energy (MFE) calculations on the window of interest, with constraints that the 5' and 3' ends cannot base pair with themselves. The sequence is assigned a score of 1 if the resulting minimum free energy structure has at least the number of base pairs as the centroid; a score of 0 if it has less base pairs than the centroid but at least half; a score of -1 if it contains less than half the number of base pairs as the centroid; and a score of -2 if the MFE structure has zero energetically favorable base pairs. Sequence scores are summed to assess overall support for the cluster, with any cluster with a zero or negative score eliminated as unfavorable.

3.6 Resampling

To further locate missing FN features, the sequences are resampled with constraints based on the clusters found thus far. Clusters with broad sequence support define an initial set of features from every sequence that can tentatively be considered TP. The features from each sequence that are not part of a cluster, then, can be considered FP: predicted by Boltzmann sampling, but not backed up by enough structural support across all sequences to be significant. Since these FP features preclude other, potentially native helices from forming, the sequence is resampled again while prohibiting the FP features. The basic steps in resampling are:

1. Form constraints file forbidding all features not contained in cluster by end of cluster validation
2. If the first iteration, resample each sequence using the constraints file
3. If not the first iteration, check whether any clusters are not present in previous iteration; if so, resample with constraints file
4. If no new clusters are formed between the present iteration and the last, terminate the program and output the final list of clusters

However, prohibiting the FP features could result in a sample significantly less energetically favorable than the original, to the point of being implausible. Hence, if the new sample's MFE structure is more than one standard deviation below from the median free energy of the old, then the new sample is considered implausible and not used, with the original sample being employed instead.

The entire algorithm is run again after Boltzmann sampling with constraints: profiling, clustering, refinement, and validation. If, after the second iteration, any new clusters are found, then the resampling occurs again, incorporating the data from the new clusters. If no new clusters are found, then the process terminates, with the last set of clusters presented as the final output. Termination is guaranteed, because all previously found TP features are retained; either the shrinking set of 'new' features or the increasing free energy suboptimality will limit the number of new clusters to be discovered.

At termination, a list of clusters in outputted with its set of sequence/feature pairs, and its representative centroid stem (i_c, j_c, k_c, l_c). In the tRNA example, `ConsensusStems` terminates with four clusters having 30, 27, 29, and 28 sequence/feature pairs, with respective centroid stems: $(1, 71.5, 7, 7)$, $(10, 25, 4, 4)$, $(22, 48, 9, 10)$, and $(47.5, 65, 6, 6.5)$. These centroids are very close to the Rfam consensus stems of $(1, 70, 7, 7)$, $(10, 24, 4, 4)$, $(26, 42, 5, 5)$, and $(48, 64, 5, 5)$.

Indeed, we shall see that the high accuracy of `ConsensusStems` in predicting the native stems is reflected across all tested Rfam families.

4 Results

Given a set of homologous RNA sequences as input, `ConsensusStems` outputs a list of clusters denoted by their centroids. These $[i_c, j_c, k_c, l_c]$ quadruples are the consensus stem predictions. The sequence/feature pairs associated with each cluster indicate the support for this prediction across the family. Hence, we evaluate consensus prediction accuracy at three levels of structural granularity: base pair, stem, and centroid. Additionally, the dependence on number of sequences in the test family is analyzed.

Accuracy is measured by precision, recall, and Mathews Correlation Coefficient (MCC) [28]. These depend on the number of TP, FP, and FN structural elements, denoted tp, fp, and fn. Precision is calculated as P $= \frac{tp}{tp+fp}$ and recall as R $= \frac{tp}{tp+fn}$ while MCC is approximated [1, 29] as their geometric mean: MCC $\approx \sqrt{PR}$. Hence, to evaulated accuracy, we must define, and then count, TP, FP, and FN at each level of granularity. To illustrate, we return to our initial tRNA *T.brucei* example.

4.1 At the base pair level

Although `ConsensusStems` harnesses the power of structural abstraction [22], base pair accuracies are measured for two reasons. First, this is the usual standard [30], so confirms that this new approach does no worse than other methods. Second, and more importantly, the contrast in accuraccies validates the choice of structural granularity. Base pairing interactions which are perceived as contradictory become a unified pattern when consolidated into a single stem, i.e. extended helix. Thus, the *exact same data* at higher abstraction/lower granularity is a much stronger and more accurate structural signal.

A predicted pairing is a TP if it appears in the Rfam alignment for that sequence. However [1], it is only classified as a FP if it actively contradicts the Rfam structure (although still counted in the prediction size). A canonical, noncrossing native base pair is a FP if it is not predicted; Boltzmann sampling does not predict noncanonical and/or pseudoknotted pairings so these are not counted.

To generate predicted base pairs for a given sequence, each feature (i, j, k) associated with a cluster is decomposed as $(i, j), \ldots, (i + k - 1, j - k + 1)$. *T.brucei* has two features, $(46, 66, 7)$ and $(50, 60, 4)$, associated with the same cluster output by `ConsensusStems`. The first contains the native helix $(48, 64, 5)$ from the Rfam alignment for *T.brucei*. Hence, for this cluster and this sequence, $tp = 5$, $fp = 2{+}4$, and $fn = 0$.

The base pair accuracy for the 11 Rfam test families is listed in Table 2. It was calculated by summing over each cluster and each sequence. Hence, the precision denominator is the total number of predicted base pairs, with multiplicity, across the entire family. These accuracies are comparable to other state-of-the-art consensus prediction methods [11, 15, 25].

Family	Precision	Recall	MCC
tRNA+	0.52	0.76	0.63
THF	0.54	0.91	0.70
TPP*	0.44	0.74	0.57
5S*+	0.50	0.76	0.62
FMN	0.28	0.78	0.47
U1	0.31	0.58	0.42
ykoK+	0.52	0.80	0.65
glmS+	0.41	0.84	0.58
IRES cripavirus	0.35	0.65	0.48
IRES HCV*	0.25	0.46	0.34
metazoa SRP	0.58	0.73	0.65
Avg	0.43	0.73	0.56
Stdev	0.11	0.13	0.11

Table 2. Base pair accuracy as described in Section 4.1. Values are comparable to other consensus methods. Note the low average precision relative to recall.

Family	Precision	Recall	MCC
tRNA+	0.97	0.91	0.94
THF	0.96	0.96	0.96
TPP*	0.98	0.84	0.91
5S*+	0.99	0.98	0.98
FMN	0.98	0.88	0.93
U1	0.95	0.90	0.93
ykoK+	1.00	0.97	0.98
glmS+	0.80	0.85	0.82
IRES cripavirus	0.83	0.85	0.84
IRES HCV*	0.99	0.81	0.90
metazoa SRP	0.91	0.71	0.80
Avg	0.94	0.88	0.91
Stdev	0.07	0.08	0.06

Table 3. Stem accuracy according to Section 4.2. Predictions, especially precision, have improved measurably with the reduction in granularity.

4.2 At the stem level

The *T.brucei* helices $(46, 66, 7)$ and $(50, 60, 4)$ are in conflict at the base pair level, since their coordinates overlap. However, they reinforce a clear, common structural signal at a higher level of abstraction — that the 5' and 3' regions of the stem $(46, 66, 8, 10)$ interact.

To consolidate this information, all features for a given sequence in a particular cluster are grouped under a single stem with coordinates (i, j, k, l). This indicates that all cluster pairings (i', j') for this sequence have endpoints with $i \leq i' \leq i + k - 1$ and $j - l + 1 \leq j' \leq j$ for the shortest possible segments. Hence, the stem $(46, 66, 8, 10)$ communicates that regions $[46, \ldots, 53]$ and $[57, \ldots, 66]$ of the *T.brucei* sequence interact, although the specific base pairings may belong to either feature $(46, 66, 7)$ or $(50, 60, 4)$, or even some other helix.

This abstraction also addresses the situation when one helix extends another in close enough succession to be clustered together.

A predicted stem is a TP if it intersects at least 50% of both 5' and 3' regions of a native stem [12, 15]. Otherwise, it is a FP. As with base pairs, predicted stems that are not in the native but do not contradict it are excluded from being counted as a FP; at least 50% of the stem must not be contradictory for this rule to apply. A native stem which is not so intersected by a predicted one is a FN. According to the Rfam alignment, native stems for *T.brucei* include $(48, 64, 5, 5)$. Since this is a subset of $(46, 66, 8, 10)$, the prediction is a TP.

The stem accuracy of each family, listed in Table 3, was calculated like base pairs, by summing over each cluster and each sequence. At this level of granularity, the wisdom of not trying to resolve competing base pairings is clear. Instead, those signals have been consolidated into a single, coherent regional interaction, and the resulting increase in accuracy, especially in precision as illustrated in Figure 9, is substantial.

4.3 At the centroid level

The previous section evaluated prediction accuracy for a family according to the pooled sequences' stem accuracies. We now consider the consensus stem predictions; recall that the cluster centroid (i_c, j_c, k_c, l_c) is the median of its constituent sequences' (i, j, k, l) stem coordinates.

The native consensus stems were determined from the Rfam consensus structure for each family. A quadruple (i, j, k, l) is the maximal set such that all base pairs (i', j') are located within the stem ($i \leq i' \leq i + k - 1$ and $j - l + 1 \leq j' \leq j$), are nested within each other (if $i_A < i_B$, then $j_B < j_A$), and with no non-nested base pairs (i_n, j_n) occurring such that

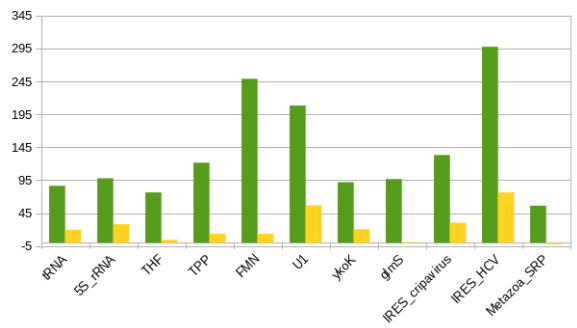


Fig. 9. Percentage accuracy increases with increasing abstraction from base pairs to stems for precision (green) and recall (yellow). While detailed prediction remains difficult, a clear structural signal emerges at the higher level of structural abstraction.

Family	Precision	Recall	MCC
tRNA+	1.00	1.00	1.00
THF	1.00	1.00	1.00
TPP*	1.00	1.00	1.00
5S*+	1.00	1.00	1.00
FMN	1.00	0.80	0.89
U1	1.00	1.00	1.00
ykoK+	1.00	1.00	1.00
glmS+	0.75	0.75	0.75
IRES cripavirus	0.80	1.00	0.89
IRES HCV*	1.00	1.00	1.00
metazoa SRP	0.86	0.86	0.86
Avg	0.95	0.95	0.94
Stdev	0.08	0.10	0.08

Table 4. Cluster centroid accuracy as in Section 4.3. At this scale, the method has perfect precision and recall for 64% of the test families.

either i_n or j_n is located in the regions $[i, i+k-1]$ or $[j-l+1, j]$. Visually, this is represented in the nested parentheses found in final line of the Rfam consensus alignment as an (i, j, k, l) region containing base pairs of the same symbol: '{}', '[]', '<>', or '()'.

Rfam's tRNA consensus structure includes the stem $(47, 63, 5, 5)$ while `ConsensusStems` outputs the centroid $(47.5, 65, 6, 6.5)$ for the cluster which contains the *T.brucei* stem $(46, 66, 8, 10)$. The cluster centroid is a TP since it overlaps with more than 50% of a native consensus stem. If no centroid so overlapped, the native would be a FN. If the centroid did not overlap with any native (or overlapped a native by less than 50%), it would be a FP.

The results in Table 4 show that 7 of 11 families tested have perfect consensus stem prediction accuracy. Average recall increased slightly over the summed sequence values in Table 3, while average precision remained essentially constant. The increase in recall is consistent with the value of consensus structure prediction; homologous sequences compensate for FN predictions.

The fact that accuracy values can drop as well as increase reflects the very different numbers of predictions when moving from the sequence/stem pairs to cluster centroids. For example, the glmS family has 77 TP, 15 FP, and 11 FN stem predictions but only 3 TP, 1 FP, and 1 FN centroids. Hence, the fraction of errors is higher for centroids since the set sizes are an order of magnitude smaller.

4.4 Tests with reduced sequence sets

The accuracy of `ConsensusStems` does depend on a large enough pool of Boltzman samples. This effect was assessed by running six additional trials per family: three sets of size 15 and then three of size 8 randomly chosen sequences from the original pool (summarized in Table 1).

Dropping the set size to 15 decreases the average centroid precision, recall and MCC by -2.6%, 10.8%, and 5.1%. Reducing the number further to 8 results in average decreases of 0.9%, 21.3%, and 13.8% respectively.

The corresponding changes at the base pair level were 3.5%, 38.4%, and 23.1% and then 3.8%, 67.1%, and 44.0%. Unsurprisingly, recall is affected much more severely than precision by decreases in the initial amount of information. This is especially apparent at high granularity.

5 Discussion

Results demonstrate that Boltzmann sampling of homologous sequences filtered by RNA profiling and noise-sensitive clustering resolves the consensus structure problem at the stem level of granularity. Namely, `ConsensusStems` output clusters recover the native regions of interaction with a high degree of accuracy on average, and low standard deviations, for a comprehensive test set of 11 Rfam families.

The cluster centroids, which are the consensus stem predictions for the entire family, have an average (approximate) Mathews Correlation Coefficient (MCC) of 95%, with an 8% standard deviation. MCC is useful as a summary statistic since it incorporate both precision (how many predicted elements are native) and recall (how may native elements are predicted) into a single value. The associated individual sequence stems, which are the 'alignment' produced by this method, similarly have an average MCC of 91% with a 6% standard deviation. In comparison, the base pair level MCC average is just 56% with std $= 11\%$.

Improvements in RNA prediction accuracy achieved by structural elements at lower granularity than base pairs is a known phenomenon [22]. Here, profiling's methodology has been extended from maximal helices to stems, a higher level of structural abstraction. Stems consolidate competing helices often predicted by thermodynamic optimization [30] into a single coherent structural signal — two regions interacting with high probability. By turning 'competitors' into 'allies', `ConsensusStems` achieves high accuracy, correctly predicting the forest by not trying to resolve each tree.

The power of abstraction is such that resonable consensus predictions can be achieved on relatively small sets of sequences. That recall suffers disproportionately is not surprising; the method extracts the signal present with high precision, but cannot consolidate what is not there. Hence, sufficiently many homologous sequences (26 on average for the 11 Rfam families tested) are needed for full recall.

In comparison to other stem-based consensus prediction approaches, `ConsensusStems` has been more comprehesively tested [12] and more rigorously evaluated [15]. The `comRNA` program [12] identifies conserved helices in unaligned sequences using a graph-theoretic approach. Three familes were tested, with average precision of 86.7% but no reported recall. The longest sequences were $\sim$200 nucleotides (nt), and most number of native stems were 5. Here, 11 families were tested with sequence lengths up to 300 nt, and up to 10 native consensus stems. The `RNAscf` [15] program achieved average precision and recall of 88.4% and 92.6% respectively over 12 test families with sequence lengths up to 200 nt and up to 5 stems. However, these values counted any overlap between predicted and native stems as a true positive, not the 50% minimum required here.

Moreover, while others [12, 15] perform exhaustive searches to find all possible helices, `ConsensusStems` harnesses the power of Boltzmann sampling to generate only the most probable ones. Their analysis is then very fast; RNA profiling extracts features from a set of structures in time linear in their size [18], and the noise-sensitive clustering algorithm DBSCAN has an average complexity of $O(n \log n)$ [23]. While the cubic runtime of Boltzmann sampling [17] is the bottleneck here, it is orders of magnitude lower than many consensus prediction methods [6, 8].

Additionally, if speed-ups are desired, it would be straight-forward to parallelize the sampling component of this approach since the individual input sequences have no data dependencies. Thus, `ConsensusStems` embodies the best of both worlds, achieving high accuracy in an efficient manner.

6 Conclusion

Predicting a common structure for a family of RNA sequences is is an old and important open problem in computational biology. It is challenging in no small part because RNA structures are much more strongly conserved than sequence identity. Our new `ConsensusStems` method addresses this challenge by (1) finding the right balance between precision and recall, and (2) selecting on an appropriate level of structural granularity.

First, Boltzmann sampling of sufficiently many homologous sequences eliminates nearly all false negative predictions, while noise-sensitive clustering of RNA profiling features filters almost all false positive ones. Second, focusing on the 'forest' of interacting sequence segments, rather than the 'trees' of specific base pairs, yields clear and accurate predictions, both for the cluster centroids as consensus stems as well as the supporting sequence/feature pairs across the family.

Thus, our method succeeds in predicting the consensus stems with a high, often perfect degree of accuracy, as tested on a diverse group of families whose lengths span the range where the thermodynamic model is the most accurate. Even if a finer grained consensus prediction is desired, `ConsensusStems` should be used to make the initial lower granularity prediction. The predicted consensus stems can then be fine tuned, either by applying sequence alignment tools [31] to the features of the cluster, or by using the predicted stems as the known structural input to an alignment method [32].

Hence, both on its own and as a initial step toward accurate base pair level prediction, `ConsensusStems` represents a significant advance to the state-of-the-art of consensus structure prediction.

Funding

Funding for this paper was provided in part by the Burroughs Wellcome Fund [CASI #1005094 to C.E.H.], and by the NIH [R01 6M126554].

References

[1]P P Gardner and R Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinfor*, 5(1):140, 2004.

[2]P P Gardner, A Wilm, and S Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–2439, 2005.

[3]A Wilm, I Mainz, and G Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algor for Mol Biol*, 1(1):19, 2006.

[4]J H Havgaard and J Gorodkin. RNA structural alignments, part I: Sankoff-based approaches for structural alignments. *Meth Mol Biol*, 1097:275–290, 2014.

[5]K Asai and M Hamada. RNA structural alignments, part II: non-Sankoff approaches for structural alignments. *Meth Mol Biol*, 1097:291–301, 2014.

[6]S Lalwani, R Kumar, and N Gupta. Sequence-structure alignment techniques for RNA: A comprehensive survey. *Advances in Life Sciences*, 4(1):21–35, 2014.

[7]M Hochsmann, T Toller, R Giegerich, and S Kurtz. Local similarity in RNA secondary structures. In *Bioinfor Conf, 2003. CSB 2003. Proc of the 2003 IEEE*, pages 159–168. IEEE, 2003.

[8]J Liu, J TL Wang, J Hu, and B Tian. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC bioinfor*, 6(1):89, 2005.

[9]I L Hofacker, S HF Bernhart, and P F Stadler. Alignment of RNA base pairing probability matrices. *Bioinfor*, 20(14):2222–2227, 2004.

[10]D A Sorescu, M Möhl, M Mann, R Backofen, and S Will. CARNA – alignment of RNA structure ensembles. *Nucleic Acids Res*, 40(W1):W49–W53, 2012.

[11]X Xu, Y Ji, and G D Stormo. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinfor*, 23(15):1883–1891, 2007.

[12]Y Ji, X Xu, and G D Stormo. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinfor*, 20(10):1591–1602, 2004.

[13]J H Havgaard, R B Lyngsø, G D Stormo, and J Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinfor*, 21(9):1815–1824, 2005.

[14]Y Tabei, K Tsuda, T Kin, and K Asai. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinfor*, 22(14):1723–1729, 2006.

[15]V Bafna, H Tang, and S Zhang. Consensus folding of unaligned RNA sequences revisited. *Jour of Comp Biol*, 13(2):283–295, 2006.

[16]M Hamada, K Tsuda, T Kudo, T Kin, and K Asai. Mining frequent stem patterns from unaligned RNA sequences. *Bioinfor*, 22(20):2480–2487, 2006.

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

[18]E Rogers and C E Heitsch. Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble. *Nucleic Acids Res*, 42(22):e171, 2014.

[19]D H Mathews. Revolutions in RNA secondary structure prediction. *J Mol Biol*, 359(3):526–532, 2006.

[20]B A Shapiro, Y G Yingling, W Kasprzak, and E Bindewald. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol*, 17(2):157–165, 2007.

[21]E Rogers, D Murrugarra, and C Heitsch. Conditioning and robustness of RNA Boltzmann sampling under thermodynamic parameter perturbations. *Biophysl Jourl*, 113(2):321–329, 2017.

[22]E Rogers and C Heitsch. New insights from cluster analysis methods for RNA secondary structure prediction. *Wiley Inter Reviews: RNA*, 7(3):278–294, 2016.

[23]M Ester, H Kriegel, J Sander, X Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[24]S Griffiths-Jones, A Bateman, M Marshall, A Khanna, and S R Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–441, 2003.

[25]S Lindgreen, P P Gardner, and A Krogh. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinfor*, 23(24):3304–3311, 2007.

[26]L Kaufman and P J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[27]A Joshi and R Kaur. A review: Comparative study of various clustering techniques in data mining. *Intl Jour of Advd Res in Comp Sci and Software Eng*, 3(3), 2013.

[28]P Baldi, S Brunak, Y Chauvin, C AF Andersen, and H Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinfor*, 16(5):412–424, 2000.

[29]J Gorodkin, S L Stricklin, and G D Stormo. Discovering common stem–loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, 29(10):2135–2144, 2001.

[30]K J Doshi, J J Cannone, C W Cobaugh, and R R Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinfor*, 5(1):105, 2004.

[31]R C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[32]H-P Lenhof, K Reinert, and M Vingron. A polyhedral approach to RNA sequence structure alignment. *J of Comp Bio*, 5(3):517–530, 1998.