|  |
| --- |
| ­Sequence analysis  RNAG: A New Gibbs Sampler for Predicting RNA Secondary Structure for Unaligned Sequences  Donglai Wei1, Charles.E.Lawrence2,\*  1Department of Mathematics, Brown University, Providence, Rhode Island, United States of America  2Division of Applied Mathematics and the Center of Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Motivation:** RNA secondary structures play an important role in the function of many RNAs, and structural features are often key to their interaction with other cellular components. Thus, there has been considerable interest in the prediction of the secondary structures for RNA families. In this paper, we present a new global structural alignment algorithm, RNAG, to predict consensus secondary structures for unaligned sequences. It uses a blocked Gibbs sampling algorithm, which has theoretical advantage in convergence time. This algorithm iteratively samples from the conditional probability distributions P(Structure | Alignment) and P(Alignment | Structure), and in so doing refines the models of both Alignment and Structure. We use a hierarchical clustering method to characterize the shape of the posterior space, γ-centroid estimator to generate a prediction from sampled structures and credibility limits to characterize the uncertainty.

**Results:** An analysis of three publically available datasets shows substantially improved structural prediction based on PPV-SEN curves comparisons. An analysis of 17 RNA families shows that sampled structures are generally compact around their ensemble centroids, at least eleven families with well separated clusters. In general, the distances between the references structures and the predicted structures were large compared to the variation among structures within the ensemble.

**Availability:** [The python implementation of the RNAG algorithm and the repeatable results in Section 3.1, 3.2 in this paper are available at http://ccmbweb.ccv.brown.edu/rnag.html](file:///C:\Users\chip\AppData\Local\Temp\The%20python%20implementation%20of%20the%20RNAG%20algorithm%20and%20the%20repeatable%20results%20in%20Section%203.1,%203.2%20in%20this%20paper%20are%20available%20at%20http:\ccmbweb.ccv.brown.edu\rnag.html)

**Contact:** [Charles\_Lawrence@brown.edu](mailto:Charles_Lawrence@brown.edu)

# introduction

RNA secondary structure plays a key role in the functions of many RNAs including structural RNAs, non-coding RNAs (ncRNA) and RNA regulatory motifs in mRNAs. Accordingly these structural features are often characterized by evolutionarily conserved secondary structures that are critical to their functions. Furthermore, there are often multiple occurrences of these structural elements within one species. Given the recent recognition of many important additional roles that RNAs play in cellular functions predicting the common structural features of a set of RNA sequence is more important than ever.

## 1.1 Structure prediction for single sequence

Three main classes of probabilistic models of P(S|Q) for the prediction of the secondary structure (S) for a single sequence (Q), are currently available: the most popular is a thermodynamic model that supposes that RNA structures may be described by Boltzmann statistics, like Mfold (Zuker *et al.*, 1981). The second model includes phylogenetic information into folding, like PETfold (Seemann *et al.*, 2008). The third method abandons the bio-physical model in favor of machine learning algorithms to empirically infer structure based on probabilistic graphical models like CONTRAfold (Do *et al.*, 2006) or nonparametric methods, like KNETfold (Bindewald *et al.*, 2006).

Early algorithms, Mfold (Zuker *et al.*, 1981) and RNAfold (Hofacker *et al.*, 1994), use dynamic programming to find the most probable structure (MPS), the “minimum free energy structure” (MFE). However, MPS is often not representative of the Boltzmann weighted ensemble of structures. This ensemble can be represented as a large set of binary matrices, which define a high dimensional discrete space. In this high-D space even the most probable structure is likely to have low probability. Furthermore, there is not fundamental principled reason for MPS to be included in the high weight region of the Boltzmann space (Carvalho *et al.*, 2008). Thus, alternative estimators which gain information from the full ensemble of structures have emerged, including Centroid estimators (Ding *et al.*, 2005; Carvalho *et al.*, 2008) and the related maximum expected accuracy (MEA) estimator (Do *et al.*, 2006). A generalization of centroid estimator, γ-centroid (Hamada *et al.*, 2009; 2010), permits the balancing of false positive and false negative errors based on the tunable parameter γ. Moreover, the focus on finding MPS without uncertainty analysis implicitly assumes that RNA exists only in one single stable state, which is not the case for many RNAs, and almost certainly not the case of mRNAs. To address this issue, sampling algorithms, like Sfold (Ding *et al.*, 2005), provide a method to characterize the full ensemble of structures (Mathews 2006) and Bayesian confidence limits, a.k.a. credibility limits, provide a method to delineate the uncertainty of an estimate (Newberg *et al.*, 2009; Webb *et al.*, 2008).

## 1.2 Structure prediction for multiple unaligned sequences

With multiple sequences, the problem becomes harder since the extra unknown alignment (A) of the sequences enters and the model changes to P(S,A|Q). Assuming the alignment of sequences is given the model degenerates into P(S|A,Q), which is similar to the single sequence case. The goal is to identify structural features common to these sequences based on shared statistical patterns. Mutual information (Gutell *et al.*, 1992) and stochastic context-free grammars (SCFG) (Sakakibara *et al.*, 1994; Knudsen *et al.*, 1999) have been effectively used to detect and model complementary covariations that are indicative of conserved base pairing interactions. Maximum weighted matching (MWM), a graph-theoretical approach, were introduced to predict common secondary structures allowing pseudoknots (Cary *et al.*, 1995; Tabaska *et al.*, 1998). RNAalifold (Hofacker *et al.*, 2002; Bernhart *et al.*, 2008) incorporates both thermodynamic parameters and sequence covariation and permits sampling of consensus structures from its probabilistic model.

To solve the alignment problem, there are generic alignment algorithms like ClustalW2 (Ramu *et al.*, 2005) and PROBCONS (Do *et al.*, 2005), which don’t incorporation of structural information and thus model P(A|Q). Algorithms under the model P(A|S,Q), which gain information from the consensus structure, have been shown to improve the alignment of RNA sequences. In one approach, structures of individual sequences are predicted separately and abstractions of these structures are aligned (Giegerich *et al.*, 2004; Steffen *et al.*, 2006; Siebert *et al.*, 2005). Another approach (Ji *et al.*, 2004) applies graph-theory to compare and find stems conserved across multiple sequences first, and then assembles conserved stem blocks to form consensus structures in which pseudoknots are permitted. The probabilistic covariance model (Eddy and Durbin 1994) employs the stochastic context free grammar (SCFG) model to multiply aligned sequences using a given consensus structure. This algorithm iterates between parameters estimation and alignment prediction using Expectation Maximization (EM) algorithm. After convergence it permits sampling of alignments. Eddy and Durbin (1994) also presented and iterative optimization procedure that, like RNAG, iterates between alignment and structure, but instead takes and optimization approach instead of the sampling approach we employ. More recently Yao, *et al.* (2006) described an extension of this Eddy and Durbin algorithm approach and used it to find regulatory motifs.

There is a “chicken and egg” problem for these two classes of algorithms: good RNA sequence alignment (A) depends on a specified consensus structure (S) and good consensus structure (S) prediction dependents of good alignments (A). One approach to solve the dilemma is to simultaneously align and fold a pair of RNA sequences with dynamic programming algorithm (Sankoff, 1985). However its computational complexity is O(n6), too high to be of practical value. Heuristics based on simplifications and additional restrictions of Sankoff algorithm for multiple sequences (more than two) have been developed, such as FoldalignM (Torarinsson *et al.*, 2007), mLocARNA (Will *et al*., 2007), Murlet (Kiryu *et al.*, 2007) and RNA Alignment and Folding (RAF) (Do *et al,* 2008).

Another approach is to iteratively predict structure and alignment conditional on each other. Earlier works focused on finding the optimal solution with EM algorithm (Eddy *et al.*, 1994; Yao et.al. 2006), or simulated (Lindgreen *et al.,* 2007). Recently, approaches that draw samples from probabilistic models using Markov chain Monte Carlo (MCMC) procedures have been described. Meyer *et al*. (2007) employs a Metropolis-Hasting algorithm **that** makes proposals for local alignment and structures changes and accepts them probabilistically. However, the convergence of these local-move algorithms tends to require a large number of sampling steps. Another variation is RNAsampler (Xing *et al.* 2007), which heuristically iterates between the alignment and pieces of possible stems of the multiple sequences.

Gibbs sampling introduced by Geman and Geman (1984), is another popular MCMC procedure. Inspired by a theorem of Liu (1994) concerning accelerated convergence of various Gibbs samplers, here we propose a blocked sampling algorithm that iterates between alignment (A) and structures (S). In Liu's theorem one, three alternative Gibbs sampling approaches are considered: 1) the standard Gibbs samples in which each of the random variables (RV) are sampled individually; 2) the grouped Gibbs sampler in which two or more of the RV are sampled jointly in blocks; and 3) the collapsed Gibbs sampler in which at least one of the RVs is removed from the problem via integrations. He compares their convergence speed rates on their forward operators, Fs, Fg, Fc, respectively. The theorem shows that norms of these operators are ordered as follows ||Fc|| ≤ ||Fg|| ≤ ||Fs||. Thus the expected number of iterations until convergence follows the reverse order. However, as he points out, if the computation required for each iteration to sample blocks or to remove random variables via integration is too large then any improvements in convergence rate may not be worth of the added computational expense. Thus the key is to find efficient procedures for blocking or integrating.

# methods

## 2.1 Sampling Algorithm: Composite probabilistic model

Consider the probabilistic model P(A,S|ΛA,ΛS,Q) for multiple sequences Q, where hidden variables are: A the alignment, S the consensus structure and ΛA,ΛS the corresponding parameters of A,S prediction steps. The goal is to find samples from the joint distribution P(A,S|ΛA,ΛS,Q). The blocked Gibbs sampler, RNAG, described here achieves this by iteratively sampling from the conditional probability P(S(t)|A(t-1), ΛS,Q) and P(A(t)|S(t-1), ΛA ,Q) for the t-th iteration. Notice that our algorithm provides a generic framework, where several of current algorithms can fit into each of these two sampling steps. Specifically, RNAG proceeds as follows:

*2.1.1 Alignment Initialization* In theory, it does not matter if the algorithm starts from an initial alignment or an initial consensus structure. Here we begin with an initial alignment A(0) produced by PROBCONS (Do *et al.*, 2005) under the model P(A|Q).

*2.1.2 Iteration Steps*

1. Sample consensus structure (S(t)) given alignment (A(t-1))  To sample from P(S(t)|A(t-1), ΛS,Q), we employ RNAalifold (Bernhart *et al*., 2008), which combines thermodynamic parameters and empirical parameters estimated from the aligned sequences using a default covariation weight ΛS.
2. Sample alignment (A(t)) given consensus structure (S(t-1))  To sample from P(A(t)|S(t-1), ΛA,Q), we employ the Infernal package (Nawrocki *et al.*,2009). ΛA is a set of empirical parameter estimates (parameters for SCFG model) obtained from P(ΛA |S(t-1),A(t-1),Q) using Expectation Maximization (EM) algorithm. Given ΛA, a multiple alignment is sampled from P(A(t)|ΛA , S(t-1) ,Q) using the SCFG model.

Figure S1 show a diagram of these steps.

## 2.2 Sample analysis: Characterize the posterior space

As described by Mathews (2006), sampling from the Boltzmann weighted ensemble of secondary structures can provide a full characterization of this space. Here the RNAG sampler draws samples from this very high dimensional space of structures and alignments. In this analysis our attention is focused on the sampled structures, though the multiple alignments also evolve during the sampling. We employ clustering analysis to characterize the overall shape of the posteriors space of structures and credibility limits to delineate uncertainty in predicted structures.

2.2.2  *Clustering analysis* Boltzmann weighted ensembles of RNA secondary structure can exhibit complex shapes, which often include multiple modes (Ding *et al.*, 2006). Here we examine the shape of the probabilistically weighted posterior space using a hierarchical clustering procedure like that employed by Ding *et al.* (2006) for a single sequence.

Direct comparison of the sampled consensus structures is impractical because of the dependence of the indices of the bases of sampled structures on the alignment. Thus, we follow the second evaluation procedure in Hamada *et al.*(2010), projecting the consensus structure back onto each sequence and use hierarchical clustering method on the projected structures.

2.2.2  *Centroid Estimator* In this analysis we employ γ-centroid estimators for structure prediction and for the comparison of alternative predictive methods. The γ-centroid, as a generalization of the centroid estimator, provides a means to balance sensitivity and positive predictive value (PPV) and accordingly can be used to compare procedures over the range of this tradeoff. We employ the γ-centroid estimator for such comparisons and the original centroid estimator in calculation of bias and variance.

## Evaluation Metric

*2.3.1 Prediction Accuracy* We compare the predicted structure for each sequence with its reference structure and calculate sensitivity (SEN) and positive predictive value (PPV). SEN is the fraction of known pairs correctly predicted and PPV is the fraction of predicted pairs in the known structure (Mathews, 2004). Using γ-centroid estimation, we can interpolate a curve on PPV-SEN plane based on different γ value (Hamada *et al.*, 2010). Following the president of Do et al. (2008), we report the average of (PPV, SEN) calculated for each test case weighting each sequence equally. For the comparison of the relative performance of the algorithm across RNA families, we use the area under the curve, acquired with linear interpolation, as a qualitative measure.

*2.3.2 Uncertainty Analysis*

1. Credibility Limit: Any prediction of structure provides only a point estimate of secondary structure, but gives no information about the uncertainty of that estimate. We employ Bayesian confidence limits, a.k.a. credibility limits to characterize this uncertainly (Newberg *et al.*, 2009; Webb *et al.*, 2008). These limits compute the radius of the smallest hyper-sphere centered at the estimate containing 95% of the posterior weighted space.
2. Bias-Variance Analysis: In any prediction based on finite data involving comparison with a reference, deviations from the reference involves two components, bias and variance, where the bias measures the square error distance between the mean and the reference, while the variance gives the variation around the mean. In this discrete setting where the secondary structure is treated as a binary matrix with random elements, the mean is almost certainly not a feasible RNA secondary structure because it will almost certainly not be integer valued. Accordingly, here we measure bias as the distance between the structure in the ensemble that is nearest to the mean in the least squares sense, the centroid, and the reference structure (Carvalho and Lawrence, 2008), and the variance around the centroid of the ensemble. As Carvalho and Lawrence show, for binary variables square error distances, pth power error differences, and Hamming distances are equal. We also obtain variances around cluster centroids.
3. Separation index: To assess how well separated the clusters are relative to the variation within clusters we use the following separation index

 (1)

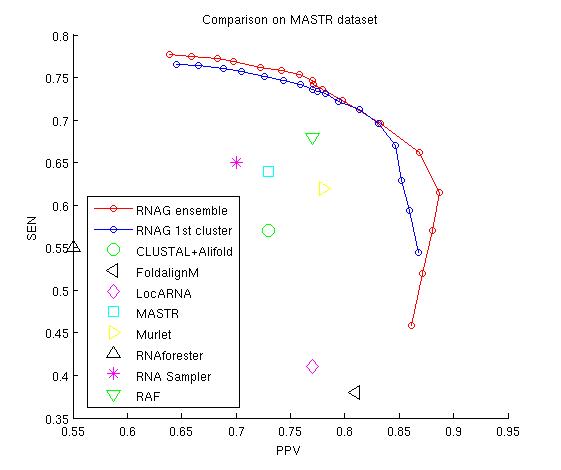
,where *D* is the Hamming distance between the two centroids of the two largest clusters, i.e. the total number of paired bases contained in one of the two centroid structures but not in the other, and  are the 95% credibility limits around the two largest cluster centroids. When the index is at least 1 no more than 5% of the structures from either cluster are within the 95% credibility limit of the other cluster, and thus we say the two largest clusters are well separated.

# results

For all the experiments below, we employed RNAG for a burning period of 1,000 sampling iterations and the next 1,000 sampled structures are used in our analyses. We followed Hamada *et al.* (2010) and picked 17 γ-centroid estimators where γ∈{2k: −5 ≤ k ≤ 10,k∈Z}∪{6} to interpolate the curve on the PPV-SEN plane.

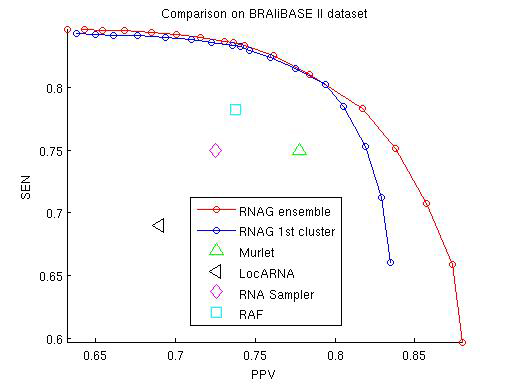
## Comparison of accuracy

In our first accuracy assessment, we evaluated RNAG using the benchmarking dataset of MASTR RNA secondary structure prediction program. Results of current algorithms for the MASTR dataset were given in Do *et al.* (2008) and the sensitivity-PPV frontier form RNAG was plotted on PPV-SEN plane in Figure 1.



**Fig 1**. Average performance of different secondary structure prediction methods in PPV-SEN plane for MASTR dataset. PPV=TP/P= TP/(TP+FP), SEN=TP/T=TP/(TP+FN).

We also tested and compared different align-fold algorithms on the first dataset from BRAliBASE II, which contains collections of around 100 five-sequence sub-alignments sampled from five specific Rfam families (5S rRNA, gourp II intron, SRP, tRNA and U5). We did not use SRP for quality issues described on their website. For comparison, the results in Do *et al.* (2008) were averaged over the four RNA families and are plotted on PPV-SEN plane along with the RNAG frontier in Figure 2.



**Fig 2**. Average performance of different secondary structure prediction methods in PPV-SEN plane for 4 RNA families (5SRNA, group II intron, tRNA and U5) from BRAliBASE II dataset.

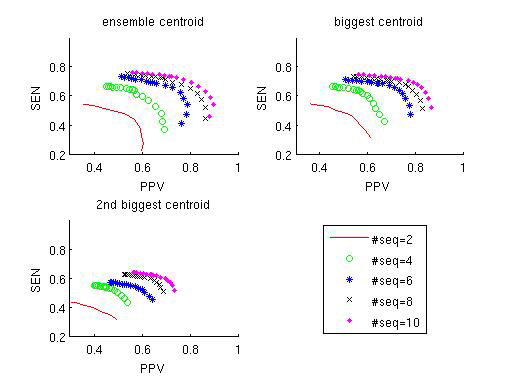
The fact that in both of these comparisons extant procedures lie below the RNAG frontier indicates that on average RNAG provides a better tradeoff for PPV and sensitivity. Not surprising, this isn’t always the case. For the BRAliBASE II dataset, Do *et al.* (2008) gives results for four methods in each of these four RNA families. As shown in figure S2, 14 of these 16 predictions below the RNAG frontier and two are somewhat above this frontier.

## A sample survey of Rfam

In order to have a larger sample of Rfam, to explore RNAG’s properties, we used the benchmarking dataset of Murlet (Kiryu *et al.* 2007), which contains 85 reference alignments of 10 sequences taken from 17 RNA families in the Rfam database (Griffiths-Jones *et al.*, 2005). This dataset spans a range of sequence lengths from 51 to 291 bases, and a range of sequence identity from 40% to 94%, including nine families with identities under 60%. Details of the dataset are presented by Kiryu *et al.* (2007). Kiryu *et al.*(2007) used this data set for comparison of algorithms that predict a consensus structure for an aligned set of sequences. Perhaps not surprisingly, as shown in figure S3 RNAG also performs better these procedures including CentroidAlifold (Hamada *et al.*, 2010), a state-of-the-art algorithm. However, our focus here is on the variation of performance with number of sequences in the alignment and over various families of RNAs.

* + 1. *Investigation on the number of unaligned sequences*

To assess the contribution from multiple sequences, we took *N* (2≤*N*≤10) random sequences from each of the 85 sub-alignments in Kiryu *et al*. (2007), ran RNAG on these subsets of sequences and averaged over 10 independent runs except for N=10. As Figure 3 shows, additional sequence improves prediction of the reference structure but with decreasing increment as indicated by the small improvement from 8 to 10 sequences. But as shown in figure S4 and table S1 this finding differs when between sequence sets under and over 60% identity with larger gains and an indication for perhaps further room for improvement with more sequences for those with sequence identity under 60%. Notice in Table 1 that the bias decreases with the number of sequences in the alignment but with decreasing gains, which is in agreement with improvements in the area under the PPV-SEN curves.



**Fig 3.** Improvement of PPV-SEN curves with number of sequences involved in the prediction

*3.2.2 A Detailed Look into each Family*

The above results describe the overall performance of RNAG for this dataset, but don't reveal differences across the families. In Table 2, we list the bias-variance statistics, area under PPV-SEN curve and cluster statistics for each family. As this table indicates, there is considerable variability in the biases and areas under the PPV-SEN curves between the families, which reflects the fact that the ability to predict the references structure varies widely between families. Figure 4 highlights this variability and shows that there is strong correlation between bias and the area under the PPV-SEN curve. Notice that the normalized 95% ensemble credibility limits are under 10% for eleven of the families, which indicates that in the majority of the families the probabilistically weighted ensembles are quit tightly compact around the centroid of the full ensemble. Normalization was obtained by dividing Hamming distances by the lengths of the sequences. In spite of this eleven families have a separation score of at least 1 as indicated by the last column in Table 2. This indicates that the cluster centroids are well separated for these eleven families since the distanced between the centroids of the two clusters is at least as large as the sums of the 95% credibility limits of two clusters. Finally notice that the biases, which give the distances between the predicted structures and the references structure, are more than twice as large as the standard deviations of the distances of ensemble members around the predicted structures.

**Fig 4.** 2D plot of the 17 RNA families grouped by their types on the plane of Bias per base pair (bp) and Area under the PPV-SEN of Ensemble Centroid

* 1. **Detecting the structure of a Riboswitch:**

A riboswitch is part of an [mRNA](http://en.wikipedia.org/wiki/Messenger_RNA) molecule that binds a [small target molecule](http://en.wikipedia.org/wiki/Small_molecule), and changes its conformation. It has been reported that the existence of bimodality in structure space is the indication of a riboswitch at least for the SAM family of riboswitches (Giegerich *et al.*,1999; Freyhult *et al.*, 2007).

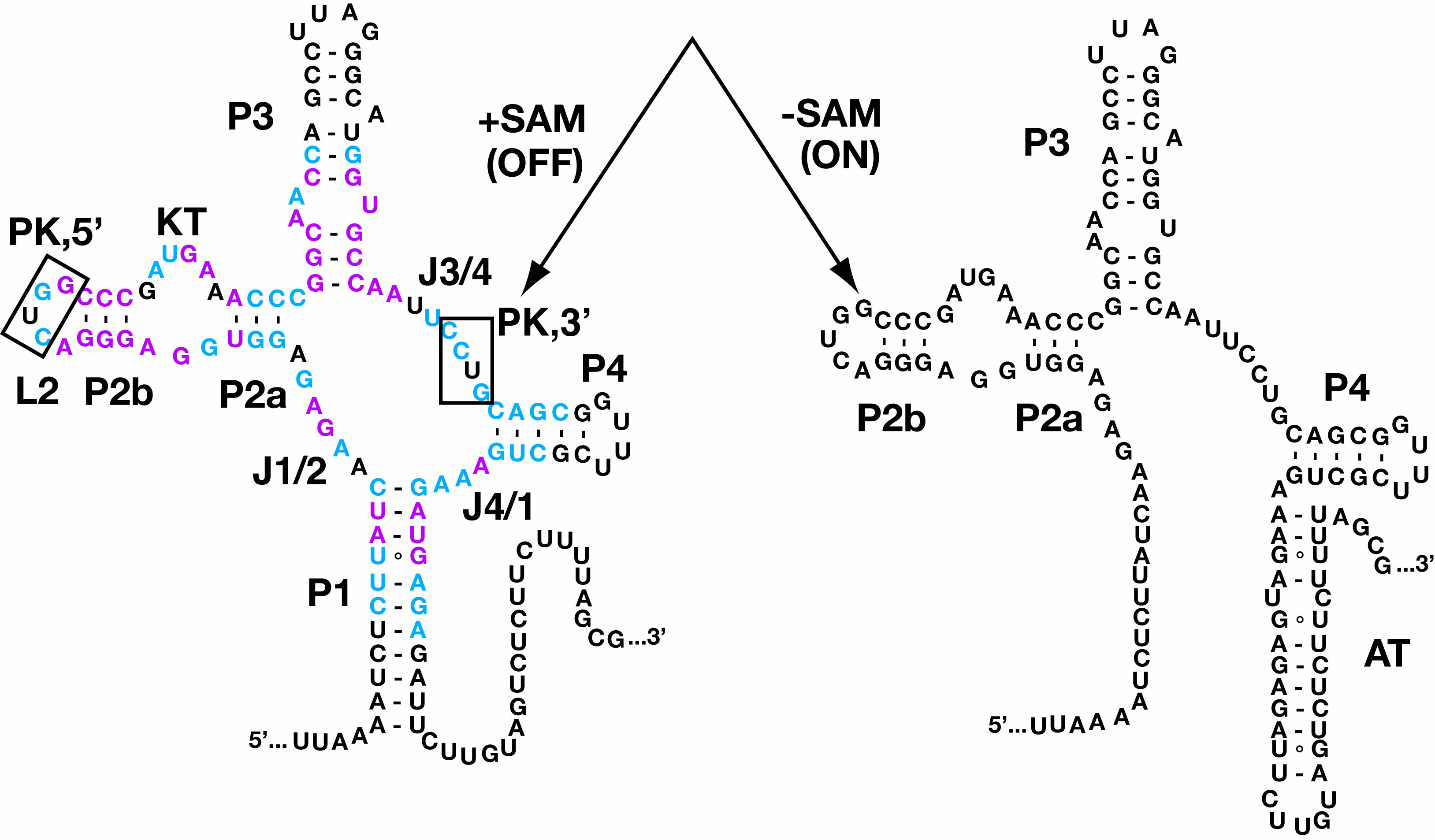
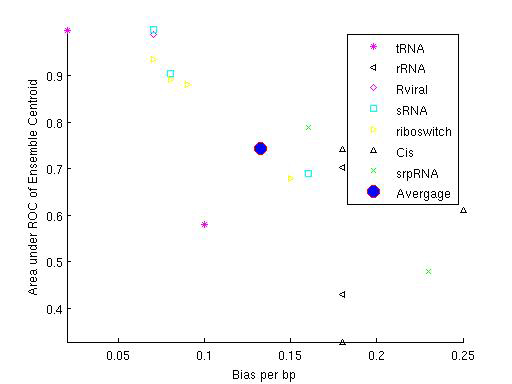
Here, we examine the 105nt SAM riboswitch with EMBL accession number AE016750.1/132874-132778, used in Freyhult *et al.*, (2007). The test dataset consists of 5 sub-alignments of 10 sequences from SAM family. In each sub-alignment, the first sequence is the target RNA mentioned above and the rest are manually chosen uniformly from the phylogenetic tree of SAM family. As shown in Table 3, the bias of the larger cluster centroid estimators from RNAG decreases with the increase of the number of sequences, and agrees with the X-ray structure with SAM-off to within a normalized Hamming distances of less that 7% in all five sets.

**Table 3.** Bias: Hamming distance from the centroids of the two largest clusters to the reference structure in Rfam. The value for the 2nd cluster is in the bracket.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test#/#seqs | 2 | 5 | 7 | 10 |
| 1 | 0.18(0.11) | 0.04(0.06) | 0.06(0.08) | 0.05(0.14) |
| 2 | 0.17(0.24) | 0.04(0.07) | 0.06(0.08) | 0.07(0.18) |
| 3 | 0.08(0.27) | 0.05(0.03) | 0.04(0.01) | 0.01(0.03) |
| 4 | 0.11(0.08) | 0.06(0.16) | 0.05(0.02) | 0.07(0.05) |
| 5 | 0.15(0.06) | 0.03(0.14) | 0.04(0.13) | 0.06(0.04) |

The centroid of the larger of the two clusters is almost always closer to the reference than that from the smaller. As the table indicates, with 10 sequences the centroid of the larger cluster predicts the references structure in Rfam quite well, while the centroid of the second structure doesn’t.

None of the predictions of from previous studies of these molecule nor those of RNAG predict the SAM-on Xray structure well. Apparently the conformation of the SMA is altered too much by the binding of SAM to be captured with models that consider the energy of the RNA alone. In addition the suggestion of previous works that the presents of a bimodal posterior spaces is indicative of a riboswitch is problematic as our finding of well separated clusters for eleven of the 17 families and two of the four riboswitch families in the Kiryu *et al.*(2007) dataset indicates that the existence of distinct clusters is quite common in RNA families.



**Fig 5.** Experiment result of two distinct secondary structures for SAM RNA family 2D plot to cluster different RNA families. The left structure is close to the reference structure in Rfam and right one cannot be predicted by current models which do not capture the interaction with the extra molecule.

# Discussion:

From Figure 1 and 2, we can see that extant procedures yield a combination of sensitivity and PPV that are considerably below the RNAG frontier in this plane. Some features of RNAG suggest an explanation for this behavior. RNAG not only inherits the advantage of the sampling method but also enjoys theoretically convergence advantage over the Metropolis-Hasting algorithm employing local moves. Since it samples full valid secondary structure, RNAG enjoys advantage over iteration algorithms that perform this step heuristically. Also, since recursions samples directly the full spaces of alignments and structure in each of its two steps it avoids the need to use a reduced model that is common to several extant procedures. However, since it is an MCMC procedure there is no means to assure that it has converged to its target distribution.

**4.1 Limitations of comparison datasets and training**

We specifically selected three published datasets and compared RNAG’s performance to the published performance of other methods in order to avoid self serving selection biases and biases that can arise with less than ideal application of extant methods. Of these three the most extensive dataset is that of Kiryu *et al*. (2007), which includes 17 Rfam families. We accept that in our field it is almost always difficult or impossible to obtain a truly representative dataset. Nevertheless it is important to recognize that available datasets have limitations. Specifically, generalizations from these seventeen plus the datasets from the other two comparison groups to the population of RNAs should be drawn with some caution as the combined sample size is not large and these sets are not random samples. We did very little to train RNAG in this implementation. First, the auxiliary packages that RNAG uses (PROBCONS, RNAalifold, Infernal) are used with their default options and default parameters. Furthermore, we used the Kiryu *et al.* (2007) data set exclusively to select among the small number of available algorithms for each of the three components of RNAG in this implementation. Thus, Kiryu data set can be seen as a training data set, and the MASTR and BRAliBASE II data sets are test data sets.

**4.2 Potential Improvement of RNAG**

There are several potential means to improve RNAG. Since we have done no training to select options or parameters for the algorithms components in this implementation, the performance of RNAG can potentially be improved by exploiting the full strength of these packages and by choosing parameters via cross-validation. Moreover, RNAG is only a framework for computation and the auxiliary packages above can be replaced by any other algorithms that are designed for P(A|Q), P(S|A,Q) and P(A|S,Q). Furthermore, RNAG now takes the theoretical advantage of blocked Gibbs Sampler by grouping parameters to sample into S and A. A further increase in the convergence rate may be available by integrating out A from the model to take the advantage of collapsed Gibbs Sampler. There are several other options for improving the algorithms speed including the use of better stopping rules, parallel implementation, and the use of more advances sampling methods such as parallel tempering.

## 4.3 An alternative goal of these algorithms

Our findings of substantial biases, suggest that current alignment and structural models are deficient, that we haven’t sampled long enough to achieve convergence, or that several of the references structures in 17 Rfam families are not reflective of the structural and sequence features common to RNA families. As shown in table S2, only two of the reference structures in the Kiryu *et al.* (2007) dataset were obtained by covariation analysis and thirteen by X-ray, NMR. Thus, nearly 76% of the reference structures in this data set were determined by in vitro methods. Structures from in vitro experiments may not reflect structure features common among family members as key interacting factors are not present in these experiments. This suggests an alternative goal for align-fold algorithms aimed at RNA family identification: correct classification of sequences to families, similar to that reported by Webb *et al*. (2002) for protein sequences. As the database of Rfam families have been obtained based on alignments to specify “reference structure”, it will be a particularly difficult challenge to demonstrate that there is an alternative structure, which is superior in the identification of family members. Thus comparisons of performances in family membership may require the use of reference sets obtained through independent experiments, such as experiments using immunoprecipitation (IP) methods. Finally, the existence of small variances indicates that an alternative estimator that trades of variance to reduce bias may yield lower overall deviations.

## 4.4 Confusion of maximum expected accuracy (MEA)

In recent publications (Do *et al.*, 2006; Kiryu *et al.*, 2007), maximum expected accuracy (MEA) estimators are widely used as a better representative than the previous MFE estimator. However, we find the name of MEA misleading. If the MEA is calculated on the basis of base pairs instead of individual bases then this estimator corresponds to the centroid or γ-centroid. But our findings of large biases of these estimators indicate that expected "accuracy" is misleading in that there is no assurance that these estimators are close to an outside reference structure. However, these estimators do return estimates that have minimum variance, and thus in the least squared sense they are the most reproducible of all estimators in the posterior weighted space. Accordingly, they would be better described as Maximum expected precision (MEP) estimators, or perhaps preferably by the non buoyant name that defines them as centroid or γ-centroid estimates.

# Conclusion

In this study, we introduce a blocked Gibbs Sampler (RNAG) to predict secondary structure for unaligned RNA sequences. RNAG confronts the high time complexity of the align-fold problem by capitalizing on Liu’s findings on blocked Gibbs sampling. As Figure 1 and 2 show that the new algorithm delivers substantial improvement during PPV-SEN performance. However, as with any MCMC procedure, evidence of convergence of the burn in can’t be guaranteed. Also, in the current implementation of this algorithm little has been done to obtain fast code or an efficient stopping rule. So improvements in implementation speed, need to be considered. While the results with the two available datasets and another shown in Figure S3 are encouraging, these do not assure that this procedure will perform this well for all RNA sequence sets. Furthermore, this procedure and others like it may not be ideal for structure prediction since if it works perfectly, it will only capture structural and sequential feature common to a set of input sequences, much as motif finding algorithms capture sequence characteristics common to transcription factor binding sites in multiple sequences. Nevertheless, here we show that RNAG does a better job at predicting reference structures than extant procedures while providing a fuller characterization of the shape of the posterior space including characterization of multimodal features and ascertainment of uncertainty in structural predictions. Even if RNAG does continue to perform well at this task, several more steps will be necessary to develop a fully Bayesian RNA motif finder.

acknowledgementS

We thank Lauren Alpert, Sean Eddy, Dave Mathews, William Thompson, and the referees for their helpful suggestions. This work was partially supported by DOE grant DOE: DE-FG02-04ER63942

References

Bafna V,et al Consensus folding of unaligned RNA sequences revisited. J. Comput. Biol. 200613:283-295.

Bindewald, E and Shapiro B.A (2006): RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA. 12(3): 342-352 (2006).

Bernhart,S.H. et al*.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474–484.

Carvalho L, Lawrence C. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. Proc. Natl Acad. Sci. USA. 2008105:3209-3214.

Cary RB, Stormo GD.(1995) Graph-theoretic approach to RNA modeling using comparative data. Proc. Int. Conf. Intell. Syst. Mol. Biol. 19953:75-80

Ding Y, Chan CY, and Lawrence CE. (2005) RNA Secondary Structure Prediction by Centroids in a Boltzmann Weighted Ensemble. RNA, 11 (8):1157-1166.

Ding Y, Chan CY, and Lawrence CE (2006) Clustering of RNA secondary structures with application to messenger RNAs, Journal of Molecular Biology, 359: 554–571

Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. (2005) PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment. Genome Research 15: 330-340.

Do, C.B., Woods,D.A., and Batzoglou, S. (2006) CONTRAfold: RNA Secondary Structure Prediction without Energy-Based Models. *Bioinformatics*, 22(14): e90-e98.

Do, C.B., Foo, C.-S., and Batzoglou, S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics 24: i68-i76

Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res.22:2079-2088.

Freyhult .E, Moulton.V, Clote P. (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection. , Bioinformatics. 2007 Aug 1523(16):2054-62. Epub Jun 14.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Analysis and Machine Intelligence 6: 721-741

Giegerich R., Haase D. and Rehmsmeier M. (1999) Prediction and visualization of structural switches in RNA, Pac. Symp. Biocomput., Pages: 126-137,

Giegerich R, et al (2004) Abstract shapes of RNA. Nucleic Acids Res.32:4843-4851.

Gorodkin J, et al (2001) Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res.29:2135-2144.

Griffiths-Jones S, et al (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res.33:D121-D124.

Gutell R.R.,Power, A., Hertz, G.Z., Putz, E.J. and Stormo, G.D.(1992) Nucleic Acids Res. 20: 5785-5795

Hamada M, Sato K, Asai K (2010) Improving the accuracy of predicting secondary structure for aligned RNA sequences. Nucleic Acids Res. Sep 15. doi: 10.1093/nar/gkq792

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., & Asai, K. (2009),“Prediction of RNA Secondary Structure Using Generalized Centroid Estimators,” Bioinformatics,25(4), 465–473. doi:10.1093/bioinformatics/btn601.

Hofacker IL, Fekete M, Stadler PF. (2002) Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 2002319:1059-1066.

Hofacker,I.L. et al. (2004) Alignment of RNA base pairing probability matrices.Bioinformatics, 20, 2222–2227.

Hofacker I.L. et.al (1994) Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 125: 167-188

Holmes,I. and Rubin,G.M. (2002) Pairwise RNA structure comparison with stochastic context-free grammars. Pac. Symp. Biocomput., 163–174.

Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. B*MC Bioinformatics*, **6**, 73–77.

Ji Y, Xu X. and Stormo, G.D.(2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics 200420:1591-1602

Kiryu,H. et al.(2007) Murlet: a practical multiple alignment tool for structural RNAsequences. *Bioinformatics*, **23**, 1588–1598.

Kiryu,H. Tabei,Y.,Kub,T. and Asai,K. (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics 200723:434-441.

Knudsen B, Hein J (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 199915:446-454.

Lindgreen S., Gardner P.P. and Krogh A. (2007): MASTR: Multiple alignment and structure prediction of non-coding RNAs using simulated annealing, Bioinformatics, 23(24):3304-11

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association, 89: 958-966.

Mathews D.H. (2006) Revolutions in RNA Secondary Structure Prediction J. Mol. Biol. 359, 526–532

Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. 317:191-203.

Meyer I. M. and Miklos I., SimulFold: Simultaneously Inferring RNA Structures Including Pseudoknots, Alignments, and Tress Using a Bayesian MCMC Framework, PLoS Computational Biology (2007) 3(8): e149

MEME (2011), <http://meme.sdsc.edu/meme4_6_0/intro.html>

Montange, R.K, Batey R.T. (2006) Structure of the S-adenosylmethionine riboswitch regulatory mRNA element Nature 441, 1172-1175

Nawrocki E. P., Kolbe D. L., and Eddy S. R.,(2009) Infernal 1.0: Inference of RNA alignments , Bioinformatics 25:1335-1337

Newberg L. and Lawrence C. (2009) Exact calculation of distributions on integers, with application to sequence alignment. Journal of Computational Biology, 16(1), 1-18

Ramu C.et al., (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31 (13):3497-5000

Sakakibara Y, et al (1994) Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res. 22:5112-5120.

Sankoff D Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. 198545:810-825.

Seemann SE, Gorodkin J, Backofen, R.( 2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Research, 36(20):6355-6362,

Siebert S, Backofen R MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics 200521:3352-3359.

Steffen P, et alRNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 200622:500-503.

Tabaska JE, et al An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 199814:691-699.

Thompson WA.Newberg, LA. Conlan SP.McCue, LA. Lawrence, CE. (2007) The Gibbs Centroid Sampler, *Nucleic Acids Res,* 35(Web Server issue): W232-W237.

Torarinsson,E. *et al*. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932

Webb BJ, Liu JS, Lawrence CE (2002), BALSA: Bayesian Algorithm for Local Sequence Alignment. Nucleic Acids Res, 30(5):1268-1277.rescu,A.

Webb-Robertson, B.-J.M., McCue, L.A., and Lawrence,C.E. (2008) Measuring global credibility with application to local sequence alignment. PLoS Comput. Biol. 4, e1000077.

Will,S. *et al.* (2007) Inferring noncoding RNAfamilies and classes by means of genomescale structure-based clustering,. *PLoS Comput. Biol.*, **3**, e65–e65.

Xing Xu, Yongmei Ji, and Gary D. Stormo (2007) RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment Bioinformatics, August 23: 1883 - 1891.

Yao Z,et al*. Cmfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 2006;22:445-452.*

Zuker M, Stiegler P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. Jan 109(1):133–148

**Table 1.**Effects of number of sequences in the alignment on the structure prediction results. For each row, we not only calculate the average area under PPV-SEN curve for accuracy comparison, but also summarize the bias-variance statistics and the size of the two biggest clusters to visualize the clustering results. In order to normalize bias, standard deviation(std) and credibility limit with respect to the sequence length, we divide them by the averaged sequence length for the family.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #seqs | Area under PPV-SEN curve | | | Bias | Std | #samples | | | 95% Credibility Limit | | |
| ensemble | 1st cluster | 2nd cluster | 1st cluster | 2nd cluster | 1st +2nd cluster | ensemble | 1st cluster | 2nd cluster |
| 2 | 0.44 | 0.46 | 0.37 | 0.27 | 0.04 | 728.13 | 150.76 | 878.89 | 0.21 | 0.14 | 0.11 |
| 3 | 0.58 | 0.59 | 0.49 | 0.20 | 0.03 | 793.15 | 124.94 | 918.09 | 0.14 | 0.10 | 0.07 |
| 4 | 0.58 | 0.58 | 0.48 | 0.20 | 0.03 | 791.66 | 115.00 | 906.66 | 0.14 | 0.09 | 0.06 |
| 5 | 0.62 | 0.63 | 0.51 | 0.17 | 0.03 | 802.20 | 113.24 | 915.44 | 0.12 | 0.08 | 0.05 |
| 6 | 0.67 | 0.67 | 0.54 | 0.16 | 0.03 | 800.50 | 111.66 | 912.16 | 0.11 | 0.07 | 0.05 |
| 7 | 0.70 | 0.69 | 0.57 | 0.15 | 0.03 | 795.52 | 111.92 | 907.44 | 0.10 | 0.07 | 0.05 |
| 8 | 0.73 | 0.71 | 0.60 | 0.15 | 0.03 | 797.56 | 116.19 | 913.75 | 0.10 | 0.07 | 0.04 |
| 9 | 0.73 | 0.73 | 0.60 | 0.14 | 0.02 | 790.59 | 122.38 | 912.97 | 0.09 | 0.06 | 0.04 |
| 10 | 0.75 | 0.74 | 0.63 | 0.13 | 0.02 | 792.85 | 125.11 | 917.96 | 0.09 | 0.06 | 0.04 |

**Table 2.** A detailed look into RNAG results on 17 RNA families. We group the selected families by their type and calculate the average PPV-SEN area for accuracy comparison and statistics like bias, standard deviation (std), credibility limit and separation index from cluster analysis to better understand the posterior space of the secondary structure space.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RNA Family | RNA type | mean length  (percent identity) | Bias | Std | Credibility Limit | | | PPV-SEN Area | | | # Samples | | | Separation  Index |
| ensemble | 1st cluster | 2nd cluster | ensemble | 1st cluster | 2nd cluster | 1st+2nd | 1st cluster | 2nd cluster |
| T-box | tRNA | 244(45) | 0.10 | 0.01 | 0.06 | 0.04 | 0.02 | 0.58 | 0.55 | 0.47 | 926 | 826 | 100 | 1.00 |
| t-RNA | tRNA | 73(45) | 0.02 | 0.01 | 0.03 | 0.01 | 0.01 | 1.00 | 0.99 | 0.91 | 949 | 888 | 61 | 2.50 |
| 5S-rRNA | rRNA | 116(57) | 0.17 | 0.02 | 0.07 | 0.05 | 0.03 | 0.70 | 0.70 | 0.67 | 922 | 751 | 171 | 0.88 |
| 5-8S-rRNA | rRNA | 154(61) | 0.18 | 0.03 | 0.14 | 0.10 | 0.08 | 0.43 | 0.42 | 0.26 | 907 | 744 | 163 | 0.56 |
| Retroviral-psi | Rviral | 117(92) | 0.07 | 0.05 | 0.15 | 0.11 | 0.05 | 0.99 | 0.99 | 0.47 | 981 | 952 | 29 | 1.25 |
| U1 | sRNA | 157(59) | 0.16 | 0.02 | 0.06 | 0.06 | 0.02 | 0.69 | 0.69 | 0.63 | 988 | 928 | 60 | 1.13 |
| U2 | sRNA | 182(62) | 0.08 | 0.02 | 0.05 | 0.05 | 0.02 | 0.90 | 0.90 | 0.71 | 981 | 941 | 40 | 1.14 |
| Sno-14q-I-II | sRNA | 75(64) | 0.07 | 0.03 | 0.12 | 0.08 | 0.07 | 1.00 | 0.92 | 0.86 | 838 | 636 | 202 | 0.47 |
| Lysine | riboswitch | 181(49) | 0.07 | 0.02 | 0.06 | 0.05 | 0.03 | 0.94 | 0.93 | 0.84 | 983 | 923 | 60 | 0.88 |
| RFN | riboswitch | 140(66) | 0.15 | 0.03 | 0.11 | 0.06 | 0.06 | 0.68 | 0.64 | 0.60 | 820 | 574 | 246 | 0.58 |
| THI | riboswitch | 105(55) | 0.08 | 0.02 | 0.07 | 0.06 | 0.02 | 0.89 | 0.88 | 0.75 | 968 | 869 | 99 | 1.13 |
| S-box | riboswitch | 107(66) | 0.09 | 0.02 | 0.07 | 0.03 | 0.03 | 0.88 | 0.87 | 0.74 | 945 | 806 | 139 | 1.17 |
| IRES-HCV | Cis | 261(94) | 0.25 | 0.05 | 0.21 | 0.16 | 0.08 | 0.61 | 0.58 | 0.44 | 936 | 877 | 59 | 1.00 |
| SECIS | Cis | 64(41) | 0.17 | 0.02 | 0.08 | 0.02 | 0.02 | 0.74 | 0.71 | 0.72 | 840 | 679 | 161 | 1.50 |
| UnaL2 | Cis | 54(73) | 0.18 | 0.03 | 0.06 | 0.02 | 0.02 | 0.33 | 0.62 | 0.61 | 867 | 752 | 115 | 1.00 |
| SRP-bact | srpRNA | 93(47) | 0.16 | 0.03 | 0.12 | 0.04 | 0.04 | 0.79 | 0.78 | 0.70 | 834 | 646 | 188 | 2.75 |
| SRP-euk-arch | srpRNA | 291(40) | 0.23 | 0.01 | 0.04 | 0.03 | 0.02 | 0.49 | 0.48 | 0.47 | 921 | 837 | 84 | 0.80 |
| Avg |  | 142 | 0.13 | 0.02 | 0.09 | 0.06 | 0.04 | 0.76 | 0.74 | 0.63 | 926 | 826 | 100 | 0.90 |

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)