Sequence analysis

RNAG: A New Gibbs Sampler for Predicting RNA Secondary Structure for Unaligned Sequences

Donglai Wei[1], Charles.E.Lawrence[2,*]

[1]Department of Mathematics, Brown University, Providence, Rhode Island, United States of America

[2]Division 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

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 algorithm, RNAG, to predict consensus secondary structures for unaligned sequences using the blocked Gibbs sampler, 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 17 RNA families shows substantially improved structural prediction based on PPV-SEN curves comparisons, the compactness of sampled structures around their ensemble centroids for all but two families, 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 3.1.1 in this paper are available at http://ccmbweb.ccv.brown.edu/rnag.html

Contact: Charles_Lawrence@brown.edu

1 INTRODUCTION

Non-coding RNAs (ncRNA) and RNA regulatory motifs in mRNAs play important roles in gene regulation and other cellular functions. They are often characterized by evolutionarily conserved secondary structures that are critical to their functions. The identification of conserved RNA secondary structures in related sequences provides a promising avenue for the characterization of novel ncRNA genes and RNA regulatory motifs.

Three main classes of probabilistic models of P(S|R) for the prediction of the secondary structure(S) for a single sequence (R), 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.*, 2008) 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 in the high dimensional discrete spaces, where even the most probable structure has incredibly low probability, and 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.*, 2008). Also, 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. Most current algorithms focus almost exclusively on generating a single best prediction without uncertainty analysis and implicitly assume that RNA exists only in one single stable state, which is often 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).

Several procedures are available for RNA secondary structure prediction based on multiple aligned sequences. The goal of these procedures 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 effec-

*To whom correspondence should be addressed.

tively used to detect and model complementary covariations that are indicative of conserved base pairing interactions. Maximum weighted matching (MWM), a graph-theoretical approach, was introduced to predict common secondary structures allowing pseudoknots (Cary *et.al.*, 1995; Tabaska *et.al.*, 1998). RNAalifold (Hofacker *et.al.*, 2002) incorporates both thermodynamic parameters and sequence covariation for prediction and permits sampling of consensus structures from its probabilistic model.

Numerous algorithms are available to align multiple sequences, but most of them aren't specific to RNA sequences and don't incorporate structural information. In one approach involving RNA secondary structure from multiple sequences, 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 *et.al.*, 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, and after convergence permits sampling of structures.

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). Several approaches have been taken to address this problem. Sankoff (1985), described a dynamic programming algorithms that simultaneously aligns and folds RNA sequences. However its computational complexity is $O(n^6)$, too high to be of practical value. Heuristics based on simplifications and additional restrictions of Sankoff algorithm have been developed, like Foldalign (Gorodkin *et.al.*, 2001) and Dynalign (Mathews *et.al.*, 2002). Stemloc (Holmes *et.al.*, 2002), PMcomp (Hofacker,et al., 2004), RNAscf (Bafna *et.al.*, 2006) and CARNAC (Touzet *et.al.*, 2004) improve the prediction performance by simplifying Sankoff algorithm in different ways, including SCFG, RNA base pairing matrix alignment and stem comparison to simultaneously achieve consensus structures and structural alignments.

More recently, approaches that draw samples from probabilistic models using Markov chain Monte Carlo (MCMC) procedures that address the alignment and the structure simultaneously have been described. They employ a Metropolis-Hasting algorithm (Lindgreen *et.al.* 2007) 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 very large number of sampling steps in the high dimensional space of secondary structures. 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 *et.al.* (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 1, 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, F_s, F_g, F_c, respectively. The theorem shows that norms of these operators are ordered as follows $||F_c|| \leq ||F_g|| \leq ||F_s||$. Thus the expected number of iterations until convergence follows the reverse order. However, as he points out, if the computation required at each iteration to either 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.

2 METHODS

2.1 The Sampling Algorithm: the composite probabilistic model

Consider the probabilistic model $P(A,S|\Lambda_A,\Lambda_S,Q)$ for multiple sequences Q, where hidden variables are: A the alignment, S the consensus structure and Λ_A,Λ_S the corresponding parameters of A,S prediction steps. The goal is to find samples from the joint distribution $P(A,S|\Lambda_A,\Lambda_S,Q)$. The blocked Gibbs sampler, RNAG, described here achieves this by iteratively sampling from the conditional probability $P(S^{(t)}|A^{(t-1)}, \Lambda_S,Q)$ and $P(A^{(t)}|S^{(t-1)}, \Lambda_A ,Q)$. 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).

2.1.2 Iteration Step:

(1) Sample consensus structure($S^{(t)}$) given alignment($A^{(t-1)}$):

To sample from $P(S^{(t)}|A^{(t-1)}, \Lambda_S,Q)$, we employ RNAalifold (Hofacker *et.al.*, 2002), which combines thermodynamic parameters and empirical parameters estimated from the aligned sequences using a default covariation weight Λ_S.

(2) Sample alignment($A^{(t)}$) given consensus structure($S^{(t-1)}$):

To sample from $P(A^{(t)}|S^{(t-1)}, \Lambda_A,Q)$, we employ the Inferno package (Nawrocki *et.al.*,2009). Λ_A is the empirical parameter estimates (parameters for SCFG model) obtained from $P(\Lambda_A |S^{(t-1)},A^{(t-1)},Q)$ using Expectation Maximization (EM) algorithm. Given Λ_A, a multiple alignment is sampled from $P(A^{(t)}|\Lambda_A , S^{(t-1)} ,Q)$ using the SCFG model.

2.2 Sample analysis: Characterize the posterior space

As described by Mathews (2006), sampling from the Boltzmann weighted space 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 (Newberg *et.al.*, 2009; Webb *et.al.*, 2008) to delineate uncertainty in predicted structures.

2.2.1 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 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 Estimation: Centroid Estimator

In this analysis we employ γ-centroid estimators for structure prediction and for the caparison of alternative predictive methods. Ding et.al (2005) first proposed the use of the centroid estimator for the prediction of RNA secondary structure. Carvalho *et.al.* (2008) employed statistical decision theory to show the advantages of centroids estimators over traditional highest scoring estimators such as the minimum free energy structure. Hamada *et.al.* (2009) introduced the γ-centroid as a generalization of the centroid estimator that 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.

2.2.3 Evaluation:

(1) Area under PPV-SEN curve:

Using the γ-centroid estimator above, we plot PPV (positive prediction value) v.s. SEN (sensitivity) and use these curves for comparison among various methods (Hamada *et.al.*, 2010). We use the area under the curve, acquired with linear interpolation, as a qualitative measure for the comparison the relative performance of the algorithm across RNA families.

(2) 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 sphere centered at the estimate containing 95% of the posterior weighted space.

(3) 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 distance between the mean and the reference, while the variance gives the variation around the mean. In this discrete setting, the mean is almost certainly not a feasible RNA secondary structure because it is very unlikely to 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 *et.al.*, 2008), and the variance around the centroid of the ensemble. We also obtain variances around cluster centroids.

(4) Separation index

To assess how well separated the clusters are relative to the variation within clusters we use the following separation index

$$S=\frac{D}{C_1+C_2} \tag{1}$$

,where D is the Hamming distance between the two centroids of the two largest clusters, and C_1, C_2 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.

3 RESULTS

3.1 Predicting RNA secondary structures for families of sequences

To permit comparison with extant procedures we employed the dataset from 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 choice permits comparison with the performance of the following algorithms (CentroidAlifold-Alipfold-McCaskill model(al-mc), CentroidAlifold-Alipfold-CONTRAfold model(al-ct), Centroid-Alifold-McCaskill model(mc), CentroidAlifold-CONTRAfold model(ct), RNAalipfold-Centroid, PETfold, RNAalifold) as described in Hamada et.al (2010). We employed RNAG for a burning period of 1000 sampling iterations and the next 1000 sampled structures were used for analyses.

3.1.1 Average performance on secondary structure prediction In order to evaluate the centroid estimators described in Section 2, we counted the total number of True Positive (TP), False Positive (FP), True Negatives (TN) and False Negative (FN) for the 17 γ-centroid estimators where $\gamma \in \{2^k : -5 \le k \le 10, k \in Z\} \cup \{6\}$ from the 85 sets of sequence to calculate the points on the PPV-SEN plane. The interpolated curves for RNAG and other popular methods are shown in Figure 1, which indicates the RNAG enjoys significant advantage over existing methods.

3.1.2 Investigation on the number of unaligned sequences To assess the contribution from multiple sequences, we took N ($2 \le N \le 10$) 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 2 shows, additional sequence improves prediction of the reference structure but with decreasing increment as indicated by the small improvement from 8 to 10 sequences. This effect is also reflected in the area under the PPV-SEN curve as shown in Table 1. 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. Also,

notice that the standard deviations, $\sqrt{\text{variance}}$, around the centroid estimators also decrease with the number of sequences in the alignment and are consistently small compared to the biases. In addition, notice that the 95% credibility limits are 4 to 5 times greater than the standard deviations, indicating that the distances from the centroids are far from normally distributed, and that the area under the PPV-SEN curve for the ensemble centroid is consistently higher than those for either of the cluster centroid.

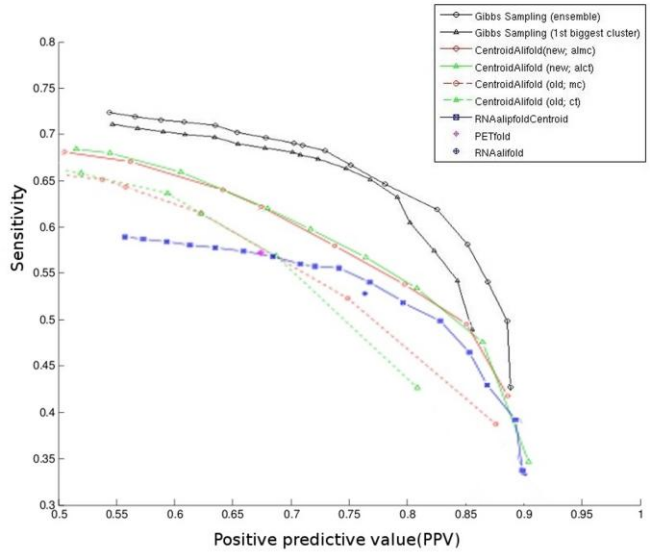

Fig 1. PPV-SEN curve performance for different secondary structure prediction methods (average over 85 RNA families). PPV=TP/P= TP/(TP+FP), SEN=TP/T=TP/(TP+TN)

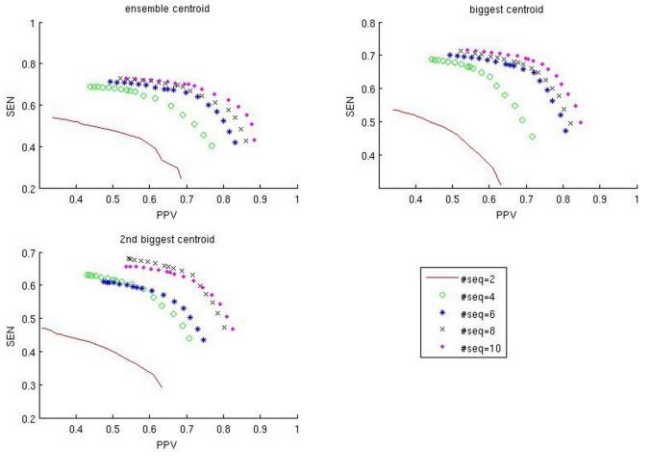

Fig 2. Improvement of PPV-SEN curves with number of sequences involved in the prediction

3.1.3 A Detailed Look into each Family

The above results describe the overall performance of RNAG for this data set, 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 3 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 all but four of the families, which indicates that in most 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 nine 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.

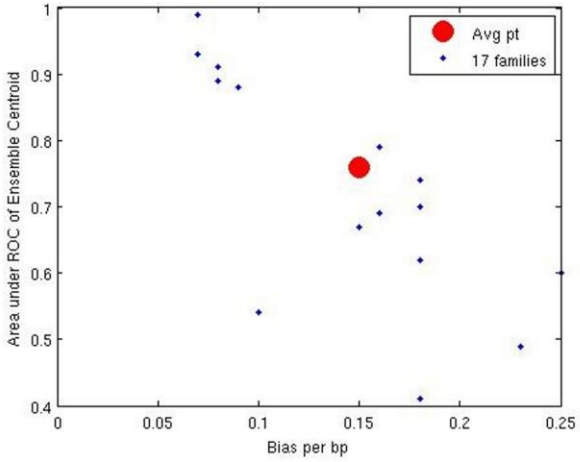

Fig 3. 2D plot to cluster different RNA families by Bias per base pair (bp) and Area under the PPV-SEN of Ensemble Centroid

3.2 Detecting Riboswitch:

Riboswitch is a part of an mRNA molecule that binds a small target 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 (Gieger-rich *et.al.*,1999; Freyhult *et.al.*, 2007).

Here, we examine the 105nt SAM riboswitch with EMBL accession number AE016750.1/132874-132778. 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 cluster centroid estimators from RNAG decreases with the increase of the number of sequences. The centroid of the larger of the two clusters is almost always closer to the reference that that from the smaller. As the table indicates with 10 sequences the centroid of the larger cluster predicts the references structure quite well, while the centroid of the second structure doesn't. However, neither centroid corresponds to the structure predicted for this ri

Table 3. 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)

boswitch when SAM is bound (Montange *et.al.*, 2006). Apparently, binding of SAM alters the structural ensemble substantially. Our finding of well separated clusters for eleven families and two of the four riboswitch families indicates that the existence of distinct clusters is not sufficient to indentify riboswitches.

4 DISCUSSIONS

RNAG not only inherits the advantage of the sampling method but also enjoys theoretically convergence advantage over the Metropolis-Hasting algorithm employing local moves, which may easily get stuck in a local energy well. Furthermore, since it samples full valid secondary structure, RNAG enjoys advantage over iteration algorithms that perform this step heuristically.

4.1 Validity of the reference structure

Our findings of substantial biases, suggest that either 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 Rfam are not reflective of the structural and sequence features common to RNA families. Specifically, many of the reference structures in Rfam are obtained from crystal structures or from in vitro experiments and 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. Of course arguments based on deviations from in-vitro experiments do not hold for reference structures obtained by covariation analysis. 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. The existence of small variances indicates that an alternative estimator that trades of variance to reduce bias may yield lower overall deviations.

4.2 Confusion of maximum expected accuracy(MEA)

In recent publications, 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 about these estimators having minimum departure from 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.

5 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 shows 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 there is substantial room for improvement in implementation speed. 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.

ACKNOWLEDGEMENTS

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).

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

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.

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.

Ji Y, et al (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) 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.

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

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.

Touzet H,Perriquet O (2004) CARNAC: folding families of related RNAs. Nucleic Acids Res. 32:W142-W145.

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.

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.

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. Eeffects of the number of sequences. 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.

	Area under PPV-SEN curve					#samples			95% Credibility Limit		
#seqs	ensemble	1st cluster	2nd cluster	Bias	Std	1^{st} cluster	2^{nd} cluster	$1^{st}+2^{nd}$ cluster	ensemble	1^{st} cluster	2^{nd} cluster
2	0.44	0.46	0.41	0.27	0.04	728.13	150.76	878.89	0.21	0.14	0.11
3	0.61	0.60	0.55	0.20	0.03	793.15	124.94	918.09	0.14	0.10	0.07
4	0.63	0.61	0.56	0.20	0.03	791.66	115.00	906.66	0.14	0.09	0.06
5	0.65	0.64	0.59	0.17	0.03	802.20	113.24	915.44	0.12	0.08	0.05
6	0.67	0.67	0.59	0.16	0.03	800.50	111.66	912.16	0.11	0.07	0.05
7	0.70	0.68	0.63	0.15	0.03	795.52	111.92	907.44	0.10	0.07	0.05
8	0.70	0.68	0.68	0.15	0.03	797.56	116.19	913.75	0.10	0.07	0.04
9	0.71	0.69	0.66	0.14	0.02	790.59	122.38	912.97	0.09	0.06	0.04
10	0.72	0.71	0.66	0.13	0.02	792.85	125.11	917.96	0.09	0.06	0.04

Table 2. Statistics for each of 17 RNA families. Bias, standard deviation(std) and credibility limit are normalized with respect to the sequence length as they are in Table 1.

					Credibility Limit			PPV-SEN Area			# Samples			Sepa-ration
RNA Family	RNA type	mean length	Bias	Std	ensemble	1st cluster	2nd cluster	ensemble	1st cluster	2nd cluster	1st+2nd	1st cluster	2nd cluster	Index
T-box	tRNA	244	0.10	0.01	0.06	0.04	0.02	0.54	0.52	0.47	926	826	100	1.00
t-RNA	tRNA	73	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	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	0.18	0.03	0.14	0.10	0.08	0.41	0.40	0.26	907	744	163	0.56
Retroviral-psi	Rviral	117	0.07	0.05	0.15	0.11	0.05	0.99	0.99	0.45	981	952	29	1.25
U1	sRNA	157	0.16	0.02	0.06	0.06	0.02	0.69	0.69	0.63	988	928	60	1.13
U2	sRNA	182	0.08	0.02	0.05	0.05	0.02	0.91	0.90	0.70	981	941	40	1.14
Sno-14q-I-II	sRNA	75	0.07	0.03	0.12	0.08	0.07	1.00	0.91	0.84	838	636	202	0.47
Lysine	riboswitch	181	0.07	0.02	0.06	0.05	0.03	0.93	0.93	0.83	983	923	60	0.88
RFN	riboswitch	140	0.15	0.03	0.11	0.06	0.06	0.67	0.64	0.59	820	574	246	0.58
THI	riboswitch	105	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	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	0.25	0.05	0.21	0.16	0.08	0.60	0.57	0.44	936	877	59	1.00
SECIS	Cis	64	0.17	0.02	0.08	0.02	0.02	0.74	0.71	0.72	840	679	161	1.50
UnaL2	Cis	54	0.18	0.03	0.06	0.02	0.02	0.62	0.62	0.61	867	752	115	1.00
SRP-bact	srpRNA	93	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	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