A conditional random fields method for RNA sequence–structure relationship modeling and conformation sampling

Zhiyong Wang* and Jinbo Xu*

Toyota Technological Institute at Chicago, IL, USA

ABSTRACT

Accurate tertiary structures are very important for the functional study of non-coding RNA molecules. However, predicting RNA tertiary structures is extremely challenging, because of a large conformation gap to be explained. Thus, an accurate score function differentiating the native structure from decoys. The fragment-based conformation sampling method (e.g. FARNA) bears shortcomings that the limited size of a fragment library makes it infeasible to represent all possible conformations well. A recent dynamic Bayesian network method, BARNACLE, overcomes the issue of fragment assembly. In addition, neither of these methods makes use of sequence information in sampling conformations. Here, we present a new probabilistic graphical model, conditional random fields (CRFs), to model RNA sequence–structure relationship, which enables us to accurately estimate the probability of an RNA conformation from sequence. Coupled with a novel tree-guided sampling scheme, our CRF model is then applied to RNA conformation sampling. Experimental results show that our CRF method can model RNA sequence–structure relationship well and sequence information is important for conformation sampling. Our method, named as TreeFolder, generates a much higher percentage of native-like decoys than FARNA and BARNACLE, although we use the same simple energy function as BARNACLE.

Contact: zywang@ttic.edu; j3xu@ttic.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

RNA has become an important research subject in recent years, and there is an increasing study of non-coding RNA in biology and health. Its growing important role appears in various life domains, including regulating gene expression (Backofen et al., 2007; Hiller et al., 2007; Ray et al., 2009; Souchek, 1985), interaction with other ligands (Badorrek et al., 2006; Bićek et al., 2005) and stabilizing itself (Reymond et al., 2009). To elucidate the function of an RNA molecule, it is essential to determine its 3D structure. However, there are a great number of RNA sequences without solved structures. Experimental methods for RNA 3D structure determination are time-consuming, expensive and sometimes technically challenging. By far, there are ~29 million RNA molecules with (predicted) secondary structure in the RFam database (Gardner et al., 2009), but only 4816 of them have tertiary structures in the nucleotide database (Berman et al., 1992). Therefore, we have to fill this large gap by predicting the 3D structure of an RNA using computational methods.

*To whom correspondence should be addressed.

RNA tertiary structure prediction does not gain as much attention as secondary structure prediction (Akutsu, 2000; Alkan et al., 2006; Backofen et al., 2009; Bindaewald and Shapiro, 2006; Eddy and Durbin, 1994; Ferretti and Sankoff, 1989; Gardner and Giegerich, 2004; Hamada et al., 2009; Hovgaard et al., 2007; Hofacker, 2003; Knudsen and Hein, 2003; Mathews, 2006; Mathews and Turner, 2002; Mathews and Turner, 2006; Poolsap et al., 2009; Will et al., 2007; Zhang et al., 2008; Zuker, 2003; Zuker and Sankoff, 1984). Both molecular dynamic methods (Bindewald and Shapiro, 2006; Hajdin et al., 2010; Sharma et al., 2008) and knowledge-based statistical methods (Das and Baker, 2007; Das et al., 2010; Frellesen et al., 2009) have been proposed to fold RNA molecules. The knowledge-based statistical methods for RNA tertiary structure prediction consist of two major components: an algorithm for conformation sampling and an energy function for differentiating the native structure from decoys. Fragment assembly, a knowledge-based method widely used for protein structure prediction (Hasel et al., 2003; Lee et al., 2004; Simons et al., 1997), has been implemented in FARNA (Das and Baker, 2007) for RNA 3D structure prediction. However, this method has a couple of limitations: (i) there is no guarantee that any region of an RNA structure can be accurately covered by structure fragments in the RNA solved structure database, which currently contains only a limited number of non-redundant solved RNA structures; and (ii) sequence information is not employed in FARNA for conformation sampling. MC-Sym (Parisien and Major, 2008) is a motif assembly method for RNA 3D structure prediction, which uses a library of nucleotides cyclic motifs (NCM) to construct an RNA structure. MC-Sym has a time complexity exponential with respect to RNA length (i.e. the number of nucleotides), so MC-Sym may not be used to predict the tertiary structure for a very large RNA. As reported in Laing and Schlick (2010), MC-Sym also fails in the case when the secondary structure of RNA lacks cyclic motifs. Recently, Frellesen et al. (2009) have proposed a probabilistic model (BARNACLE) of RNA conformation space. BARNACLE uses a dynamic Bayesian network (DBN) to model RNA structures, but this DBN method does not take into consideration any sequence information. In addition, BARNACLE models the interdependency between the local conformations of only two adjacent nucleotides, but not of more nucleotides. Other RNA three dimensional structure prediction methods can be found in Abraham et al. (2008); Das and Baker (2007); Das et al. (2010); Ding et al. (2008); Flores et al. (2010); Frellesen et al. (2009); Gillespie et al. (2009); Hajdin et al. (2010); Jonakova et al. (2009); Laing and Schlick (2010); Parisien and Major (2008); Sharma et al. (2008); Tang et al. (2005) and Wexler et al. (2006). This article presents a novel probabilistic method conditional random fields (CRFs) (LaFerty et al., 2001) to model RNA sequence–structure relationship. Different from BARNACLE...
Conditional random fields for RNA conformation sampling

modeling only RNA structures, our CRF method models the sophisticated relationship among primary sequence, secondary structure and 3D structure, which enables us to more accurately estimate the probability of RNA conformations from its primary sequence and thus sample RNA conformations more efficiently.

We have already successfully applied CRF to model protein sequence–structure relationship and conformation sampling (Zhao et al., 2008, 2009, 2010). However, our CRF method for proteins cannot be directly applied to RNA. In order to apply CRF to RNA modeling, we have to employ a different method to represent an RNA 3D structure and model RNA bond torsion angles. We also have to face the challenge that there are a lot fewer solved RNA structures than the solved protein structures for CRF model training. By exploiting the secondary structure information of an RNA molecule, we have also developed a novel tree-based sampling scheme that can simultaneously sample conformations for two segments far away from each other along the RNA sequence. In contrast, our protein conformation sampling method can sample conformations for only one short segment at a given time. Finally, we also have to employ a totally different energy function for RNA folding. To the best of our knowledge, CRF has also been applied to RNA secondary structure prediction (Do et al., 2006) and alignment (Sato and Sakakibara, 2005), but not modeling the relationship between RNA sequence and 3D structure.

Our method TreeFolder is more effective in sampling native-like decoys than FARNA and BARNACLE, although we use the same simple energy function as BARNACLE, which contains only base-pairing information. Tested on 11 RNA molecules, TreeFolder obtains much better decoys for most of them. Our results imply that TreeFolder models RNA sequence–structure relationship well, which it is feasible to sample RNA conformations without using fragments and that sequence information is important for RNA conformation sampling. Experiments also show that TreeFolder works well with predicted secondary structures generated by tools such as CONTRAfold (Do et al., 2006).

2 METHODS

2.1 Representation of an RNA structure and conformation state

We can represent an RNA 3D structure using a sequence of torsion angles, as shown in Figure 1. Every nucleotide has in total seven bonds that rotate freely. Six of them lie on the backbone: \( \text{P–O}_5^\prime \), \( \text{O}_5^\prime –\text{C}_5^\prime \), \( \text{C}_5^\prime –\text{C}_4^\prime \), \( \text{C}_4^\prime –\text{C}_3^\prime \), \( \text{C}_3^\prime –\text{O}_3^\prime \) and \( \text{O}_3^\prime –\text{P} \). The seventh bond connects a base to atom \( \text{C}_{1^\prime} \). As shown in Figure 2 torsion angle \( \chi \) around the seventh bond has a small variance, so we assume that it is independent of the other angles and has a normal distribution. The planar angles between two adjacent bonds on the backbone are almost constants, so are the lengths of the bonds.

We use a simplified representation so that we can reduce the number of torsion angles needed for the local conformation of a nucleotide (Cao and Chen, 2005; Duarte and Pyle, 1998; Hershkovitz et al., 2006; Zhang et al., 2008). In particular, we use the torsions \( \tau_1 \) and \( \tau_2 \) on pseudo-bonds \( \text{P–C}_4^\prime \) and \( \text{C}_4^\prime –\text{P} \) (see pink lines in Figure 1). However, to determine coordinates of the six backbone atoms of a nucleotide, we also need two planar angles \( \theta, \phi \) and another torsion \( \mu \) on bond \( \text{P–O}_5^\prime \). Overall, we use a five-tuple \( (\tau_1, \tau_2, \theta, \phi, \mu) \) to represent the local conformation of a nucleotide. The torsion angles are separated in several groups in the whole angle space, as shown in Figure 3. Although there are many different methods to represent an RNA conformation, this simplified representation enables us to rapidly rebuild backbone atoms from angles. Similar representations have also been...
Table 1. Accuracy of the structures rebuilt from the native torsion angles, assuming the bond lengths are constants.

| PDB ID | RMSD | PDB ID | RMSD |
|-------|------|-------|------|
| 1esy  | 0.77 | 1xpr  | 0.55 |
| 1lk4  | 0.35 | 1zhb  | 0.22 |
| 1t2a  | 0.41 | 28sp  | 1.00 |
| 1lgl  | 0.28 | 2x3d  | 0.34 |

Then, we calculate the mean and variance in each group and model the conditional probability of an RNA structure, represented as a conformation generative probability of an RNA structure, our CRF model estimates the cluster which the angles at this nucleotide belongs to. It is also the index of determining RNA tertiary structures. The output label at each nucleotide is a conformation state (also called label in CRF). It is also the index of the torsion angles are independent of one another in Gaussian distribution. Later we will show how to make angle sampling easy and fast, we assume the torsion angles as a conformation state. Given the conformation state of a nucleotide, we as a linear function of \( \sum_{k=1}^{N} \psi(y_{i},x_{k}) = \sum_{k=1}^{N} L_{k} \Phi_{i} \). \( \Phi_{i} \) is the weight for transition from state \( i \) to \( j \).

In contrast to BARNACLE (Frellsen et al., 2009) estimating the generative probability of an RNA structure, our CRF model estimates the conditional probability of an RNA structure, represented as a conformation state vector \( y \), from the input feature vector \( x \) as follows:

\[
P(Y=y|X=x) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^{t} \sum_{k=1}^{L} \psi(y_{i},x_{k}) + \sum_{i=1}^{L-1} \Phi_{i} \right)
\]

\[
Z(x) = \sum_{y} \exp \left( \sum_{i=1}^{t} \sum_{k=1}^{L} \psi(y_{i},x_{k}) + \sum_{i=1}^{L-1} \Phi_{i} \right)
\]

Meanwhile, \( Z(x) \) is the partition function, \( y \) is the feature vector at position \( i \), \( \Phi_{i} \) is the weight for transition from state \( i \) to \( j \), \( \psi \) is the weight factor for predicting state \( i \) from an input feature \( x \), \( L \) is the length of RNA, i.e. the number of nucleotides. The function \( \psi \) describes dependency between a conformation state and the input features and thus, called a label feature function. The function \( \Phi \) describes dependency between two adjacent states and thus called an edge feature function.

Figure 4 shows a linear-chain CRF model for the sequence–structure relationship of an artificial RNA with five nucleotides. We also extend \( y \) to a linear combination of features of the adjacent nucleotides in a sliding window. That is, \( \bar{y} \) is a linear function of \( \bar{y} = [x_{t-1}, x_{t}, x_{t+1}] \). \( W_{i} \) is the window size to be determined later.

Once the CRF model is trained, we can calculate the (marginal) probability of a conformation state at a given position, using the forward-backward algorithm as follows.

\[
P(Y=y|X=x) = \frac{1}{Z(x)} \sum_{u} F(L,u,x)
\]

\[
F(t,y,x) = \begin{cases} 
    \sum_{u} F(t-1,u,x)e^{\psi(y_{t},x_{u})+\Phi_{t-1}} & t > 1 \\
    e^{\psi(y_{t},x_{t})} & t = 1
\end{cases}
\]

\[
B(t,y,x) = \begin{cases} 
    \sum_{u} F(t+1,u,x)e^{\psi(y_{t+1},x_{u})+\Phi_{t}} & t < L \\
    e^{\psi(y_{t},x_{t})} & t = L
\end{cases}
\]

\[
Z(x) = \sum_{y} \prod_{t} F(L,u,x)
\]

We train our CRF model by maximizing the occurring probability of a set of training RNAs with solved structures. In order to avoid overfitting, we also enforce regularization on the model parameters. As such, we train the model parameters by maximizing the following regularized log-likelihood:

\[
\log \left( \prod_{t} P(Y=y|X=x) + \lambda \|W\|_2 + \mu \|\Phi\|_2 \right)
\]

Meanwhile, \( y \) and \( x \) are the conformation state vector and input feature vector of the \( t \)-th training RNA. \( W \) and \( \Phi \) are model parameters defined in Equation (1) and \( \lambda \) and \( \mu \) are the regularization factors. This maximization problem can be solved to optimal using the L-BFGS algorithm (Liu and Nocedal, 1989).

We also extend the first-order CRF model to the second-order model so that we can capture dependency among three adjacent nucleotides. As in Figure 5, two adjacent positions are combined to a single super node. All the algorithms for the first-order CRF model can be easily extended to the second-order model.

---

**Z.Wang and J.Xu**
2.3 A tree-guided conformation sampling algorithm

Once the CRF model is trained, we can use it to sample conformations for a segment in an RNA molecule. By combining this segment conformation sampling algorithm with a tree representation of the RNA base pairing information, we can have a tree-guided conformation sampling scheme, which enables us to sample conformations for two segments far away from each other along the sequence.

**Building a guide tree for conformation sampling:** the guide tree represents the base pairing information in an RNA, which can be predicted using a secondary structure prediction method or determined by experimental methods. In the case of pseudo-knots, we remove the minimal number of base pairings so that a tree can be built. Since the pseudo-knots do not occur frequently, removal of a small number of base pairings does not impact our method. Note that all the base pairings are taken into consideration in calculating the energy of a sampled conformation. Therefore, removal of some base pairs in tree construction will not impact the formation of pseudo-knots, since we also use energy function to guide the folding simulation. Given the base pairing information, we build a guide tree as follows. The root node in the tree corresponds to the whole RNA. Given a base pair \((i, j)\), we have one node in the tree corresponding to the segment between \(i\) and \(j\). One node \(A\) is the child of the other node \(B\) if and only if the segment corresponding to \(B\) is the minimal segment containing the segment corresponding to \(A\). In case that one node has more than two child nodes, we can always add some intermediate nodes so that any node has at most two child nodes. For example, suppose node \(A\) corresponding to segment \((i, j)\), has three child nodes \(A_1, A_2, A_3\), where \(A_3\) corresponds to segment \((i, j)\) and \(i < j < i < j < j < j < j\). We can add an intermediate node \(C\) for segment \((i, j)\) so that \(C\) becomes the parent node of \(A_1, A_2, A_3\) and \(B\).

**Segment conformation sampling algorithm:** This sampling algorithm consists of two steps: sampling a label for each nucleotide, in the segment, by the probability calculated from the CRF model and sampling real-valued angles from Gaussian distribution corresponding to a label. We use a forward-backward algorithm to sample the label sequence of a segment from position \(i\) to \(j\). The algorithm iteratively draws a conformation label of the last position from the conditional probability as follows.

\[
P(y_{i:j}|x_{i:j}) = \frac{\sum_{y_i} F(t, y_{i:j}, x_{i:j}) \exp(\phi(y_{i:j})/\beta)}{\sum_{y_i} F(t, y_{i:j}, x_{i:j})}.
\]

\[
F(t, y_{i:j}, x_{i:j}) = \begin{cases} 
\sum_{y_{i-1:j}} F(t-1, y_{i-1:j}, x_{i-1:j}) \exp(\phi(y_{i:j})/\beta) & t > i \\
\sum_{y_{i-1:j}} F(t-1, y_{i-1:j}, x_{i-1:j}) & t = i, i > 1 \\
\phi(y_{i:j})/\beta & t = i, i = 1
\end{cases}
\]

Meanwhile, \(Z(t)\) is the partition function and can be calculated using the forward-backward algorithm. After the conformation state at position \(j\) is sampled, the algorithm replaces \(y_j\) by \(j-1\) and repeats the sampling process until position \(i\) is sampled. Once the labels of the segment are sampled, we can sample the real-valued angles from the Gaussian distribution associated with a label.

**Folding simulation:** the folding simulation begins with a heating up process, in which we repeatedly sample conformations for the whole RNA using the above-mentioned segment conformation sampling algorithm. This heating up procedure terminates if one conformation without steric clashes is generated. In our experiments, we usually can obtain a conformation without clashes very quickly, which is used as the initial conformation of the simulated annealing optimization (Andree et al., 2003; Zhao et al., 2010).

To resample conformations of an RNA, we build a conformation sampling guide tree based upon the base pairing information in the RNA and all the nodes in the tree are marked as ‘ undone’. The torsion angles of the RNA are resampled using a bottom-up method along the tree as follows. We randomly pick up an ‘ undone’ node \(A\) in the tree, which is either a leaf node or a node with all the child nodes being marked as ‘ done’.

(i) If \(A\) is a leaf node, we resample the angles for the segment corresponding to \(A\) using the segment conformation sampling algorithm.

(ii) If \(A\) has one or two child nodes, by cutting out the segments corresponding to the child nodes, we have at most three separate segments \(A_1, A_2, A_3\) for which we use the segment conformation sampling algorithm to generate angles separately.

The new conformation is accepted if its energy is lower. Otherwise it is accepted by a probability \(\exp(\Delta E/T)\), where \(\Delta E\) is the energy difference between current and the new conformations and \(T\) is the annealing temperature. This sampling procedure is repeated 3000 times and then node \(A\) is marked as ‘ done’. The folding simulation process ends when the root node is marked as ‘ done’.

**Energy function:** different from the complex energy function in FARNA, we adopt a simple energy function used by BARNACLE (Frøllsen et al., 2009) as follows.

\[
E = \frac{1}{N} \sum_{i \neq j} \left( d_{ij} - \delta_{ij} \right)^2
\]

where \(N\) is the number of hydrogen bonds formed in the secondary structure (every A–U and G–U pair contributes two distances, and every C–G pair contributes three distances), \(d_{ij}\) is the distance between the donor and the acceptor of the \(i\)-th hydrogen bond and \(\delta_{ij}\) is the average length of hydrogen bonds of the same type. The smaller this value is, the more the decoy is consistent with its secondary structure. The energy is measured in Å, and the ideal base pair energy of 0 Å is only obtained for conformations with perfect base pairing.

We employ such a simple energy function (without any tuned parameters) so that we can carefully examine the performance of our sampling algorithm and perform a well-controlled comparison with other sampling methods such as BARNACLE. More sophisticated energy items, such as Mg\(^{2+}\) ion interaction and stacking effect of base pairs, can be taken into account in future study.

2.4 CRF model training

**Training data:** we build our training dataset from the RNA structure classification database DARTS (Abraham et al., 2008), which collects 244 structures representing 1333 solved RNA structures and groups them into 94 clusters. Our training set comes from the 94 cluster representative structures, which have \(\sim 6000\) nucleotides in total. We use all 94 cluster representative structures to build empirical distributions of bond torsion angles. To make sure our training dataset does not overlap with the 11 benchmark RNA molecules, we exclude the representative structures in the same cluster as the 11 benchmark RNA. With the remaining 83 structures, we use 3-fold cross-validation to determine the CRF model regularization factors \(\lambda\) and \(\mu\) and the proper window size. In each fold validation, two thirds of structures are used for training and the remaining for test.
Finally, we decide to use the normal distribution for directly from the training data without using any mathematical modeling. Model 2 samples \( \chi \) from models with different distributions of torsion \( \chi \).

Fig. 7. The 5% quantiles of the RMSD distributions for decoys sampled from models with different distributions of torsion \( \chi \). Model 1 uses a fixed value of \( \chi \). Model 2 uses log normal distribution. Model 3 is the normal distribution. Model 4 is the empirical distribution of all values in training data. Models 2 and 3 are fit from all training data.

We also investigate the sampling performance of the CRF model with 50 states. Fixing the number of conformation states in a CRF model is defined as the number of correctly predicted states divided by the total number of positions. Fig. 7 shows the 5% quantiles of RMSD values for 300 decoys sampled using four different CRF models. As shown in Figure 6, the model with 50 states generates better decoys than others.

3 RESULTS

We use 11 RNAs tested by both BARNACLE and FARNA to benchmark our method TreeFolder. These RNAs contain 12–46 nucleotides and are not homologous to any structures in our training dataset. In case an RNA has multiple NMR structures, we use the first structure in the PDB file as its native structure.

Using different methods to model the distribution of torsion, \( \chi \), makes a slight difference on the quality of sampled decoys. Figure 7 shows the 5% quantiles of RMSD values for 300 decoys sampled using four different \( \chi \) distributions with a well-trained CRF model. In Model 1, we fix \( \chi \) as the mean of the training data. Model 2 samples \( \chi \) from a log normal distribution. Model 3 samples \( \chi \) from a normal distribution. Model 4 uses sample \( \chi \) directly from the training data without using any mathematical modeling. Finally, we decide to use the normal distribution for \( \chi \), to yield a bit of variance.

Table 2. Comparison between FARNA and our method TreeFolder

| ID       | Len | Best cluster centroid RMSD | Lowest cluster centroid RMSD | No. of decoys | Best cluster centroid RMSD | Lowest cluster centroid RMSD | #Decoys |
|----------|-----|-----------------------------|-----------------------------|--------------|-----------------------------|-----------------------------|---------|
| 1a4d     | 41  | 3.43                        | 3.65                        | 28            | 2.69                        | 7168                        |         |
| 1esy     | 19  | 1.44                        | 2.00                        | 69            | 1.52                        | 22529                       |         |
| 1kka     | 17  | 2.08                        | 3.71                        | 81            | 2.44                        | 24934                       |         |
| 1zii     | 27  | 3.11                        | 4.71                        | 47            | 3.97                        | 15380                       |         |
| 1qwa     | 27  | 2.65                        | 4.76                        | 68            | 3.54                        | 15415                       |         |
| 1jqr     | 11  | 2.11                        | 6.25                        | 24            | 3.53                        | 18838                       |         |
| 1zih     | 12  | 1.03                        | 7.05                        | 117           | 1.99                        | 7168                        |         |
| 28sp     | 32  | 2.31                        | 2.96                        | 46            | 1.91                        | 17177                       |         |
| 2a43     | 26  | 2.79                        | 4.52                        | 49            | 3.47                        | 18432                       |         |
| 2888     | 34  | 2.41                        | 3.33                        | 36            | 2.77                        | 12230                       |         |

The results of FARNA are taken from Table 1 in Diao and Baker (2007). Column ‘Best cluster centroid’ lists the RMSD of the best cluster centroid of the top 1% decoys with the lowest energy. Column ‘No. of decoys’ is the number of decoys generated by the methods. Bold fonts indicate better results.

Our TreeFolder generates better decoys than FARNA: we compare FARNA and TreeFolder in terms of the quality of the decoy clustering centroids. Similar to FARNA clustering only on the top 1% decoys with the lowest energy, we run MaxCluster to cluster the top 1% of our decoys with the lowest energy into five clusters. As shown in Table 2, TreeFolder can generate decoys with better cluster centroids for nine RNAs: 1a4d, 1esy, 1kka, 1qwa, 1jqr, 1zih, 28sp, 2a43 and 2888. By the way, even if a significantly smaller number of decoys is generated by us, the lowest RMSD decoys by our TreeFolder for 1a4d, 1zih and 28sp still have smaller RMSD than those by FARNA.

Our TreeFolder generates better decoys than BARNACLE: Table 3 displays the 5% and 25% quantiles of the RMSD distributions for decoys generated by BARNACLE and TreeFolder. The quantiles by BARNACLE are taken from Supplementary Table S4 in Frellsen et al. (2009). BARNACLE considers only decoys with energy < 1, since this kind of decoys are likely to have more correct base pairings. We use exactly the same energy function as BARNACLE, so we also consider only decoys with energy < 1 to ensure a fair comparison. We did not generate as many decoys as BARNACLE, so we also consider only decoys with energy < 1. In this case, we use decoys with energy < 2. On the 10 RNAs shown in Table 3, TreeFolder yields better RMSD.
To calculate the median RMSD, for each RNA we generate 300 decoys using the two CRF models. For the CRF model without sequence features, the regularization factor is set to 5 (i.e. $\lambda = 5$). While for the CRF model utilizing sequence information, the regularization factor are set to 5 and 10 (i.e. $\lambda = 5$, $\mu = 10$). To calculate the median RMSD, for each RNA we generate 300 decoys using the two CRF models.

Sampling real-valued angles generates better decoys; in order to show the detailed difference between our TreeFolder and FARNA, we look into the decoys of 1esy. We choose it because that FARNA and TreeFolder yield the largest difference on this RNA among all the 11 tested RNA molecules. As shown in Figure 8, TreeFolder can generate a much larger percentage of decoys with RMSD $< 4 \text{Å}$ than FARNA. We also compute local RMSD of each position in the decoys, which is defined as the RMSD of the segment of four consecutive nucleotides starting with this position, as compared to the native structure. We calculate the correlation between the local RMSD of each position with the global RMSD, as shown in Figure 9. Among the decoys generated by both FARNA and TreeFolder, the local RMSD at position 13 has the highest correlation with the global RMSD. We also calculate the angle error at each position by $\text{Error} = \frac{1}{4} \sum_{i=1}^{4} \sqrt{(\theta_{i} - \theta_{i}^{\text{native}})^{2}}$, where $\theta_{i}$ is the angle vector of a decoy at one position and $\theta_{i}^{\text{native}}$ is the native angle vector at the same position.

Figure 10 shows the angle error histograms in three positions 13, 14 and 15. The angles at these three positions determine the conformation of the segment starting at position 13. At positions 13 and 15, the angle errors by our method TreeFolder are significantly smaller than those by FARNA. As Figure 10 shows, the angle errors by FARNA are distributed around several separated peaks, which may be caused by the limited number of fragments used in FARNA. In contrast, the angle errors by TreeFolder are distributed more smoothly, possibly because we can sample real-valued angles.

Folding RNA using predicted secondary structures: we use the secondary structures predicted by CONTRAfold (Do et al., 2006) and sample 1000 decoys for each RNA. The quantiles of their RMSD values are shown in Table 5. On 6 of the 10 tested RNA, decoys generated from native secondary structures are better than those from predicted secondary structures. On the other four RNAs, the difference between the two types of decoys is small, because of...

### Table 3. The 5 and 25% quantiles of the RMSD distributions for decoys generated by our method TreeFolder and BARNACLE

| PDB ID | Len | Bps | 5% | 25% | 5% | 25% | # Energy <1 | # Energy <2 |
|--------|-----|-----|----|-----|----|-----|----------|----------|
| 1esy   | 19  | 6   | 2.99 | 3.28| 2.19 | 2.60 | 577   | 2.25 | 2.78 | 1102 |
| 1lkka  | 17  | 6   | 4.40 | 5.02| 3.75 | 4.30 | 349   | 3.8  | 4.39 | 776  |
| 1lxz   | 27  | 8   | 5.43 | 6.88| 4.80 | 5.42 | 486   | 4.61 | 5.07 | 1025 |
| 1q9a   | 27  | 6   | 4.80 | 5.42| 4.55 | 5.05 | 466   | 3.9  | 4.51 | 884  |
| 1qwa   | 21  | 4   | 4.06 | 4.64| 3.65 | 4.26 | 407   | 3.9  | 4.51 | 884  |
| 1xjr   | 46  | 15  | 4.72 | 6.08| -    | -    | 215   | 4.64 | 5.48 | 26   |
| 2a43   | 26  | 7   | 3.23 | 3.76| 2.88 | 3.43 | 152   | 2.93 | 3.58 | 563  |

Bold numbers indicate better distributions. Columns ‘Energy <1’ and ‘Energy <2’ list the number of decoys with energy <1 and <2, respectively. ‘Bps’ is the number of base pairings.

### Table 4. Comparison between the CRF models using or without using sequence information

| PDB ID | Median RMSD value | Median RMSD value |
|--------|-------------------|-------------------|
|        | With seq. feature | Without seq. feature | With seq. feature | Without seq. feature |
| 1zhh   | 2.68              | 4.56              | 2.83              | 4.17              |
| 1esy   | 3.73              | 6.17              | 3.74              | 6.16              |
| 1lkka  | 5.49              | 6.67              | 2a43              | 10.62             |
| 1qwa   | 5.58              | 5.99              | 1lxz              | 11.01             |
| 2a43   | 6.56              | 9.55              | 2f88              | 13.02             |

For 10 of the 11 tested RNAs, the model using sequence information yields decoys with much smaller median RMSD. Bold numbers indicate smaller RMSD values.
Fig. 10. The angle error histograms at positions 13, 14 and 15. At positions 13 and 15, the decoys by our TreeFolder have much smaller angle errors than those by FARNA.

Table 5. Comparison between folding with native and predicted secondary structure

| PDB ID | Native SS | Predicted SS |
|--------|-----------|--------------|
|        | 5%        | 25%          | 5%        | 25%          |
| 1esy   | 2.25      | 2.78         | 3.90      | 4.35         |
| 1kka   | 3.80      | 4.39         | 4.57      | 5.46         |
| 1l2x   | 5.44      | 8.08         | 15.23 (3.53) | 17.32 (3.88) |
| 1q9a   | 4.61      | 5.07         | 4.65      | 5.01         |
| 1q9a   | 3.90      | 4.51         | 3.45      | 4.31         |
| 1ixj   | 8.84      | 9.79         | 9.17      | 9.79         |
| 1izh   | 1.36      | 1.88         | 3.56      | 4.02         |
| 28sp   | 2.93      | 3.58         | 2.71      | 3.63         |
| 2a43   | 4.64      | 5.48         | 21.22 (3.89) | 21.99 (4.35) |
| 28sp   | 3.85      | 4.57         | 3.58      | 4.21         |

The four numerical columns list the RMSD values of the 5 and 25% quantiles of the decoys with energy values <2. Bold numbers indicate better results.

accurate secondary structure prediction. The results for 1l2x and 2a43 from predicted secondary structures are quite bad, since all of their base pairs are contained in a H-type pseudoknot and only half of their base pairs are recovered by CONTRAfold. However, our TreeFolder generates decent conformations for half of the pseudoknot with predicted base pairs, as shown in brackets. In particular, TreeFolder generates decent structures for 2a43 from nucleotides 1 to 14 and for 1l2x from nucleotides 1 to 18, respectively. In order to improve sampling performance on the whole structures of 1l2x and 2a43, we need an energy function like what is used in FARNA to guide the folding simulation.

Comparison with MC-Sym on the large RNA molecules: our TreeFolder is much faster than the MC-Fold and MC-Sym pipeline (Parisien and Major, 2008) for folding large RNA molecules as shown in Table 6. The running times in this table were obtained on a workstation with 96 GB RAM and 24 computing cores [2.67 GHz Intel(R) Xeon(R)].

Overlay examples: Figure 11 shows three overlay examples of 1q9a, 2a43 and 1xjr (from left to right) with their native structures (blue). These three RNA molecules have lengths of 27 nt, 26 nt and 49 nt, respectively. Pictures in blue display native, while in red the best centroids produced by our algorithm. As shown in this figure, our algorithm recovered a pseudoknot for 2a43.

4 CONCLUSIONS

We have presented a new method TreeFolder for modeling RNA sequence-structure relationship and conformation sampling using CRFs and a tree-guided sampling scheme. Our CRF method not only captures the relationship between sequence and angles, but also models the interdependency among the angles of three adjacent nucleotides. Our conformation sampling method distinguishes from FARNA in that we do not use fragments to build RNA conformations, so that we do not need to worry about if there are a sufficient number of structure fragments to cover all the possible local conformations. Our TreeFolder also differs from both FARNA and BARNACLE, in that we use primary sequence to estimate the probability of backbone angles, while the latter two do not. In addition, we also use a tree, built from (predicted) secondary structure, to guide conformation sampling so that at one moment we can simultaneously sample conformations for two segments far away from each other along the RNA sequence. In contrast, both FARNA and BARNACLE can only sample conformations for a single short segment at any time. The results indicate that our
TreeFolder indeed models sequence–structure relationship well and
compares favorably to both FARNA and BARNACLE, even if we
use only the same simple energy function as BARNACLE.

We will extend our TreeFolder further. For example, we can
incorporate information in sequence homologs into our CRF
model so that we can estimate the conformation probability more accurately
and thus improve the sampling accuracy. Information in homologs
has been successfully used in RNA secondary structure and should
be useful for 3D structure prediction. Information in homologs
has also been used for protein conformation sampling (Zhao et al.,
2010). Currently TreeFolder works well when the native base
pairing information is used to calculate the energy function (same as
BARNACLE) and to build the sampling guide tree. Not all the RNAs
without 3D structures have the native base pairing information. Our
next step is to further improve TreeFolder with the predicted base
pairings. In particular, we need to design an energy function similar
to what is used in FARNA to guide the folding simulation so that
TreeFolder works well even if the predicted secondary structure is
not very accurate. To tolerate errors in the predicted base pairing
information, we will use the predicted confidence as the weight of
each item in the energy function and only use those base pairings
with high confidence to build the conformation sampling guide tree.
We can also take another strategy to circumvent possible impact of errors
in the predicted base pairings. In particular, we will extend our
CRF method so that we can simultaneously sample base pairings
and 3D conformations so that errors in the predicted base pairings will
be corrected in the folding simulation process.

Currently, we use a very simple energy function to guide the folding
simulation. We will develop a more sophisticated energy
function to guide the formation of hydrogen bonds in a better way,
just like what FARNA does. Thus, we cannot only generate decoys with
better RMSD, but also with better hydrogen bonds.

ACKNOWLEDGEMENTS

The authors are grateful to the open science grid and to TeraGrid for
the computational resources.

Funding: National Institutes of Health (grant R01GM087553, to J.X.); National Science Foundation (grant DBI-0960390, to J.X.); Open Science Grid and TeraGrid (grants TG-MCB100062 and TGGCR100005, to J.X.).

Conflict of Interest: none declared.

REFERENCES

Abraham, M. et al. (2007) Analysis and classification of RNA tertiary structures. RNA, 14, 2274–2289.
Akitaya, T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Appl. Math., 104, 45–62.
Akutsu, T. et al. (2006) RNA secondary structure prediction via energy density minimization. Res. Comput. Mol. Biol., 2009, 130–142.
Andreatta, C. et al. (2005) An introduction to MCMC for machine learning. Mach. Learn., 50, 5–43.
Backofen, R. et al. (2007) RNAs-everywhere: genome-wide annotation of structured RNAs. J. Exp. Zool., 308B, 1–25.
Backofen, R. et al. (2009) Sparse RNA folding: time and space efficient algorithms. Comput. Pattern Matching, 8797, 249–262.
Bakudov,C.S. et al. (2006) Structure of an RNA switch that enforces stringent retroviral protein minus-strand template fidelity. Proc. Natl Acad. Sci., 103, 751–759.
Bender, A.H. et al. (2005) Structural perspective on the activation of RNase P RNA by protein. Nat. Struct. Mol. Biol., 12, 958–964.
Cao, S. and Chen, S. (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA, 11, 1884–1897.
Das, B. and Raker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc. Natl Acad. Sci., 104, 14666–14669.
Das, R. et al. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat. Methods, 7, 201–204.
Ding, F. et al. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA, 14, 1164–1173.
Do, C.B. et al. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 22, 490–498.
Duarte, C.M. and Pyle, A.M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. J. Mol. Biol., 284, 1465–1478.
Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. Nucleic Acids Res., 22, 2079–2088.
Ferretti, V. and Sankoff, D. (1989) A continuous analog for RNA folding. B. Math. Biol., 51, 167–171.
Flores, S.C. et al. (2010) Predicting RNA structure by multiple template homology modeling. In Pacific Symposium on Biocomputing. World Scientific Publishing, Co., Hawaii, pp. 216–227.
Frellsen, J. et al. (2009) A probabilistic model of RNA conformational space. PLoS Comput. Biol., 5, 1000486.
Gauthier, P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics, 5, 140.
Guthrie, P. et al. (2005) Rifam updates to the RNA families database. Nucleic Acids Res., 37, 136.
Gwirath, D.T. et al. (1987) Secondary structure of 5S RNA: NMR experiments on RNA molecules partially labeled with nitrogen-15. Biochemistry, 26, 5213–5220.
Gillespie, J. (2009) RNA folding on the 3D triangular lattice. BMC Bioinformatics, 10, 369.
Hahn, C.E. et al. (2010) On the significance of an RNA tertiary structure prediction. RNA, 16, 1340–1349.
Hamadi, M. et al. (2009) Predictions of RNA secondary structure by combining homologous sequence information. Bioinformatics, 25, 330.
Hase, T. et al. (2003) Reducing the computational complexity of protein folding via fragment folding and assembly. Protein Sci., 12, 1177–1187.
Havgaard, J.H. et al. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput. Biol., 3, e193.
Hershkovitz, E. et al. (2006) Statistical analysis of RNA backbone, IEEE/ACM T. Comput. Biol. Bioinformatics, 3, 33.
Hinrichs, M. et al. (2007) PriRNA secondary structures influence exon recognition. PLoS Genet., 3, e204.
Hofacker, I.L. (2003) Vienna RNA secondary structure server. Nucleic Acids Res., 31, 3420–3431.
Hominick, M. et al. (2000) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA, 15, 189–199.
Koehn, D. and Heniz, J. (2003) PriFOLD: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res., 31, 3420–3428.
Lafrance, M.J. et al. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In ICML 2001: Proc. Eighteenth Intl Conf. Mach. Learn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.
Laing, C. and Schlick,T. (2010) Computational approaches to 3D modeling of RNA. J. Phys. Condens. Matter, 22, 283101.
Lee, J. et al. (2004) Prediction of protein tertiary structure using PROFESYS, a novel method based on fragment assembly and conformational space annealing. Proteins: Struct. Func., Bioinformatics, 56, 704–714.
Lee, B.C. and Noell, J. (1989) On the limited memory BFGS method for large scale optimization. Math. Program., 45, 503–528.
Mathews, D.H. (2006) Revolutions in RNA secondary structure prediction. J. Mol. Biol., 359, 526–532.
Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol., 327, 191–203.
Mathews, D.H. and Turner, D.H. (2006) Prediction of RNA secondary structure by free energy minimization. Curr. Opin. Struct. Biol., 16, 270–278.
Parisme, M. and Mayo,E. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature, 452, 51–55.
Z. Wang and J. Xu

Vuotog, U. et al. (2009) Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC Bioinformatics*, 10, S36.

Ray, J. S. et al. (2009) A stress-responsive RNA switch regulates VEGFA expression. *Nature*, 457, 915–919.

Reymond, C. et al. (2009) Modulating RNA structure and catalysis: lessons from small cleaving ribozymes. *Cell. Mol. Life Sci.*, 66, 3937–3950.

Sato, K. and Sakakibara, Y. (2005) RNA secondary structural alignment with conditional random fields. *BMC Bioinformatics*, 21, i237–i242.

Sharma, S. et al. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24, 1951–1952.

Simons, K. T. et al. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268, 209–225.

Solnick, D. (1985) Alternative splicing caused by RNA secondary structure. *Cell*, 43, 667–676.

Tang, X. et al. (2005) Using motion planning to study RNA folding kinetics. *J. Comput. Biol.*, 12, 862–881.

Wexler, Y. et al. (2006) A study of accessible motifs and RNA-folding complexity. *Res. Comput. Mol. Biol.*, 3009, 473–487.

Wills, S. et al. (2007) Inferring noncoding RNA families and clusters by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3, e65.

Wohrner, J. et al. (1999) Direct identification of NH...N hydrogen bonds in non-canonical base pairs of RNA by NMR spectroscopy. *Nucleic Acids Res.*, 27, 3104–3110.

Zhang, J. et al. (2008) Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *J. Chem. Phys.*, 128, 125107.

Zhao, F. et al. (2008) Discriminative learning for protein conformation sampling. *Proteins: Struct., Funct., Bioinformatics*, 73, 228–240.

Zhao, F. et al. (2009) A probabilistic graphical model for ab initio folding. *Res. Comput. Mol. Biol.*, 5041, 59–73.

Zhao, F. et al. (2010) Fragment-free approach to protein folding using conditional neural fields. *Bioinformatics*, 26, i110–i117.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31, 3406–3415.

Zuker, M. and Sankoff, D. (1984) RNA secondary structures and their prediction. *B. Math. Biol.*, 46, 591–621.

Zwahlen, C. et al. (1997) Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: an application to a bacteriophage λ N-peptide/boxB RNA complex. *J. Am. Chem. Soc.*, 119, 6711–6721.