PROBABILISTIC STRUCTURAL ALIGNMENT OF RNA SEQUENCES

A. Ozgun Harmanci[1], *Gaurav Sharma*[1,2]

[1]Dept. of Electrical and Computer Engineering,
University of Rochester,
Hopeman 204, RC Box 270126,
Rochester, NY 14627, USA
{arharman, gsharma}@ece.rochester.edu

David H. Mathews[2,3]

[2] Dept. of Biostat. and Comput. Biology
[3] Dept. of Biochemistry and Biophysics,
University of Rochester Medical Center
Rochester, NY 14642, USA
David_Mathews@urmc.rochester.edu

ABSTRACT

We propose an algorithm for estimating the common secondary structure, alignment, and posterior base pairing probabilities for two RNA sequences. A definition of structural alignment is presented based on a novel concept of *matched helical regions* that generalizes the common secondary structure and alignment constraints used in prior work. A probabilistic framework for scoring structural alignments is developed based on a pseudo free energy model. Utilizing the model, maximum *a posteriori* probability estimates of secondary structure and alignment, and *a posteriori* probabilities for base pairing are computed using an efficient dynamic programming algorithm. Experimental results demonstrate that the proposed method offers significant improvements in structure and alignment prediction accuracy in comparison with single sequence thermodynamic methods for secondary structure prediction and purely sequence based alignment.

Index Terms**—** structural alignment, RNA secondary structure, posterior base pairing probability

1. INTRODUCTION

Computational methods for the estimation of RNA secondary structure predict the pairing of complementary bases in the RNA molecular chain, which is mediated by the formation of hydrogen bonds and stacking of neighboring pairs, since the base pairing causes the linear chain to fold back on itself, the estimation process is commonly referred to as RNA *folding*. The input for RNA folding methods is the primary linear structure of the RNA molecule, specified as a sequence of nucleotides A, U, G, and C in the $5'$ to $3'$ direction. RNA folding methods are of significant research interest due to the increasing awareness of noncoding roles of RNA in cellular processes. The methods can be classified as techniques that operate on a single sequence [1, 2] and techniques that operate on multiple homologous sequences [3, 4, 5]. The comparative analysis between sequences implicit in multi-sequence methods provides valuable information, mimicking the biologist's approach to this problem. Among currently available algorithms, multi-sequence methods are among the most promising and are more accurate than single sequence methods [6].

An algorithm for joint prediction of secondary structure over multiple homologous sequence was first proposed by Sankoff [7] who formulated a dynamic programming solution under a "pseudoknot free" constraint on the structure. Versions of the Sankoff algorithm have been developed under a thermodynamic free energy minimization framework [6, 8] and under a probabilistic modeling framework utilizing stochastic context free grammars [1, 9]. These methods allow computation of the common secondary structures that maximize the posterior probability of structural alignment given the sequence data (or equivalently minimize the free energy). In addition, several of the algorithms provide estimates of possible suboptimal structures with the K largest *a posteriori* probability values, for some choice of K.

A limitation of these approaches is that they provide no estimates of confidence in the predicted base pairs. For methods utilizing single sequence free energy based folding, this has been addressed through computation of the partition function [10, 11] but for multiple sequences, the problem received only limited attention [12].

In this paper we present an algorithm for joint prediction of secondary structure and alignment of two RNA sequences [12]. The algorithm represents an extension of Sankoff's work [7] in that it incorporates a more sophisticated scoring mechanism for alignment and allows computation of *a posteriori* base pairing probabilities.

Results demonstrate that the method provides a significant improvement over single sequence free energy minimization. Base pairs predicted with high confidence for the proposed method exhibit greater sensitivity compared to single sequence partition function method while maintaining high positive predictive value. Base pairs that are predicted with high confidence are useful to experimentalists because they can use these predictions to constrain the folding space of the corresponding RNA sequences when determining their secondary structure by alternate experimental methods.

2. RNA STRUCTURAL ALIGNMENT

To formulate the problem of simultaneous alignment and folding of two RNA sequences, we introduce the concept of a structural alignment of two sequences. For this purpose, we begin with definitions of sequence alignment and secondary structure which we then combine using a new concept of "matched helical regions" to define a structural alignment.

Denote the two RNA sequences by $\mathbf{x}_1$ and $\mathbf{x}_2$ and then lengths as N_1 and N_2, respectively. A *sequence alignment* is defined as sequence of 3-tuples:

$$A = [(i_0, k_0, m_0), (i_1, k_1, m_1), (i_2, k_2, m_2), \ldots, (i_L, k_L, m_L)] \tag{1}$$

where $0 = i_0 \leq i_1 \leq i_2 \leq \ldots \leq i_L = N_1$ and $0 = k_0 \leq k_1 \leq k_2 \leq \ldots \leq k_L = N_2$ and $m \in \{\text{ALN}, \text{INS1}, \text{INS2}\}$ and 3-tuples satisfy following conditions:

- if $(i_n, k_n, ALN) \in \mathbf{A}$ then $i_{n-1} = i_n - 1, k_{n-1} = k_n - 1$

1-4244-1484-9/08/$25.00 ©2008 IEEE

- if $(i_n, k_n, INS1) \in \mathbf{A}$ then $i_{n-1} = i_n - 1, k_{n-1} = k_n$
- if $(i_n, k_n, INS2) \in \mathbf{A}$ then $i_{n-1} = i_n, k_{n-1} = k_n - 1$

Note that this definition of alignment is consistent with a hidden Markov model of the alignment process [1, 3] and that the sequence indices $(i_1, k_1), (i_2, k_2), \ldots, (i_L, k_L)$ define a co-incidence path for an alignment between $\mathbf{x}_1$ and $\mathbf{x}_2$ [3] and L denoted the length of sequence alignment.

A secondary structure on a sequence $\mathbf{x}$ of length N is defined as a set $\mathbf{S}$ of base pairs (i, j), $1 \leq i < j \leq N$ satisfying the (pseudo knot free) condition that there exist no two pairs (i, j), (i', j') in $\mathbf{S}$ for which $i \leq i' \leq j \leq j'$ [7].

Homologous sequences share common secondary structure. This "commonality" of the structure, however, does not imply that the patterns of base pairing are identical, rather the topology of the induced shapes is matched [13]. Thus when exploring the set of possible common secondary structures, we need to adopt a definition that agrees with the above notion of commonality. When considering sequence alignment and common secondary structure together, it can be readily seen that these two elements are not independent. Fixing the sequence alignment between the two sequences restricts the set of common secondary structures allowable and vice versa; given a common secondary structure the alignments between the sequences are restricted. To jointly and consistently handle sequence alignment and common secondary structures, it is useful to introduce the concept of a *structural alignment* between two sequences. For this purpose, a structural motif called a *matched helical region* is introduced next.

Given two sequences $\mathbf{x}_1$and $\mathbf{x}_2$, corresponding secondary structures $\mathbf{S}_1$and $\mathbf{S}_2$, and an inter sequence alignment $\mathbf{A}$ between sequences, for $\tau < \frac{j-i}{2}$ and $\mu < \frac{l-k}{2}$ we say the segments $([i, i+\tau], [j, j-\tau])$ and $([k, k+\mu], [l-\mu, l])$ constitute a *matched helical region* if

- $\exists (i', j') \in \mathbf{S}_1 \ni i \leq i' \leq i + \tau \ , \ j - \tau \leq j' \leq j$
- $\exists (k', l') \in \mathbf{S}_2 \ni k \leq k' \leq k + \mu \ , \ l - \mu \leq l' \leq l$
- $\forall (i', j') \in \mathbf{S}_1 \ni i \leq i' \leq i + \tau \ , \ j - \tau \leq j' \leq j$
    1. Aligned Base Pairs: $\exists (k', l'), k \leq k' \leq k + \mu \ , \ l - \mu \leq l' \leq l \ni (i', k', \text{ALN}) \in \mathbf{A}$ and $(j', k', \text{ALN}) \in \mathbf{A}$
    2. Base pair aligned to two unpaired bases: $\exists (k', l') \notin \mathbf{S}_2, k \leq k' \leq k+\mu \ , \ l-\mu \leq l' \leq l \ni (i', k', \text{ALN}) \in \mathbf{A}$ and $(j', k', \text{ALN}) \in \mathbf{A}$
    3. Base pair Insertion in $\mathbf{x}_1$: $\exists (k-1) \leq k' \leq k + \mu, (l - \mu - 1) \leq l' \leq l \ni (i', k', INS1) \in \mathbf{A}$ and $(j', l', INS1) \in \mathbf{A}$
- $\forall (k', l') \in \mathbf{S}_2 \ni k \leq k' \leq k + \mu \ , \ l - \mu \leq l' \leq l$
    1. Base pair aligned to two unpaired bases: $\exists (i', j') \notin \mathbf{S}_1, i \leq i' \leq i+\tau \ , \ j-\tau \leq j' \leq j \ni (i', k', \text{ALN}) \in \mathbf{A}$ and $(j', k', \text{ALN}) \in \mathbf{A}$
    2. Base pair insertion in $\mathbf{x}_2$: $\exists (i-1) \leq i' \leq i + \tau, (j - \tau - 1) \leq j' \leq j \ni (i', k', INS2) \in \mathbf{A}$ and $(j', l', INS2) \in \mathbf{A}$
- Only unpaired bases aligned with base pairs are allowed in $\mathbf{x}_1$: $\forall i', i \leq i' \leq i + \mu \ni \mathbf{S}_1$ has no base pair including i' $\exists j', j' \leq j' \leq j$ and $(k', l') \in \mathbf{S}_2, k \leq k' \leq k + \mu, l - \mu \leq l' \leq l$ such that $(i', k', \text{ALN}) \in \mathbf{A}, (j', l', \text{ALN}) \in \mathbf{A}$
- Only unpaired bases aligned with base pairs are allowed in $\mathbf{x}_2$: $\forall k', k \leq k' \leq k + \tau \ni \mathbf{S}_2$ has no base pair including k' $\exists l', l' \leq l' \leq l$ and $(i', j') \in \mathbf{S}_1, i \leq i' \leq i + \tau, j - \tau \leq j' \leq j$ such that $(i', k', \text{ALN}) \in \mathbf{A}, (j', l', \text{ALN}) \in \mathbf{A}$

A structural alignment between two RNA sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ is defined by the 4-tuple $(\mathbf{A}, \mathbf{S}_1, \mathbf{S}_2, \mathbf{H})$ where $\mathbf{A}$ is an inter sequence alignment, $\mathbf{S}_1$ and $\mathbf{S}_2$ are secondary structures on $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively and $\mathbf{H}$ is a collection of matched helical regions (with respect to $\mathbf{A}$, $\mathbf{S}_1$ and $\mathbf{S}_2$) that includes all the base pairs in $\mathbf{S}_1$ and $\mathbf{S}_2$. Each structural alignment represents a unique combination of a sequence alignment and a conformal common secondary structure for the two sequences and vice versa each sequence alignment with a conforming secondary structure defines a unique structural alignment.

3. PROBABILISTIC MODEL FOR STRUCTURAL ALIGNMENT BASED ON PSEUDO FREE ENERGY

Given two sequences $\mathbf{x}_1$ and $\mathbf{x}_2$, we would like to determine a structural alignment that maximizes the *a posteriori* probability (of the structural alignment given the sequence data) and *a posteriori* probabilities of base pairing (for the individual sequences). Empirical benchmarking of the different methods indicates that thermodynamic approaches based on free energy minimization tend to offer the best accuracy in predicting base pairs [3]. These methods, however, do not directly incorporate alignment in a rigorous fashion [12]. We therefore propose a new probabilistic model for scoring structural alignments that combines precomputed posterior probabilities of base pairing and alignment through the definition of a pseudo free energy for each structural alignment. The pseudo free energy of a structural alignment $\mathcal{S} = (\mathbf{A}, \mathbf{S}_1, \mathbf{S}_2, \mathbf{H})$ is defined as:

$$\begin{aligned}\Delta G(\mathcal{S}) = &- \sum_{(i,j) \in \mathbf{S}_1} \log(\pi_{p_1}(i,j)) \\ &- \sum_{(k,l) \in \mathbf{S}_2} \log(\pi_{p_2}(k,l)) - \sum_{i \in \Upsilon_1} \log(\pi_{u_1}(i)) \\ &- \sum_{k \in \Upsilon_2} \log(\pi_{u_2}(k)) - \sum_{(i,k,m) \in \mathbf{A}} \log(\pi_a(i,k,m)) \quad (2)\end{aligned}$$

where Υ_1 and Υ_2 correspond to the sets of unpaired bases in structures of $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively, $\pi_{p_q}(r, s)$is the precomputed base pairing probability of nucleotides at indices r and s in $\mathbf{x}_q$, and $\pi_{u_q}(r)$ is the precomputed unpairing probability of nucleotide at index r in $\mathbf{x}_q$ and unpairing probability of nucleotide at index k in $\mathbf{x}_2$ respectively. $\pi_a(i, k, m)$ is the precomputed probability of alignment state m at alignment position (i, k). Using the pseudo free energy in a manner analogous to the thermodynamic free energy, we can obtain the probability of a structural alignment $\mathcal{S}$ as

$$p(\mathcal{S}) = \frac{1}{Z} e^{-\Delta G(\mathcal{S})} \quad (3)$$

where $Z = \sum_{\mathcal{S}} e^{-\Delta G(\mathcal{S})}$ denotes the (pseudo) Boltzmann partition function.

It follows that under this probabilistic model, the maximum *a posteriori* probability (MAP) estimate of the structural alignment for the two sequences corresponds to the structural alignment with the lowest pseudo free energy. Furthermore the *a posteriori* probability that nucleotide positions i and j in the first sequence are paired (given two sequences and the model) is given by;

$$p_a^1(i,j) = p(i \square j | \mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathcal{S}: \{(i,j) \in \mathbf{S}_1\}} p(\mathcal{S}) \quad (4)$$

where $i \Box j$ denotes the event of pairing of nucleotides at indices i and j, $\mathbf{S}_1$ denotes the secondary structure corresponding to the first sequence in the structural alignment $\mathcal{S}$. *A posteriori* probabilities of base pairing for the second sequence, $p_a^2(i,j)$, are similarly determined.

One limitation of our proposed scoring model is that it implicitly assumes that the precomputed posterior probabilities of base pairing and alignment correspond to independent events. This does not hold in practice. We, however, adopt this approximation in order to simplify computations.

4. EFFICIENT STRUCTURAL ALIGNMENT

The number of possible structural alignments is exponential in the length of the shorter sequence [13]. Thus a brute force evaluation of either the MAP structural alignment or the partition function is infeasible for typical lengths of interest.

Fortunately, the problem of enumerating all structural alignments exhibits the *overlapping subproblems* property [14]; the enumeration process can be broken down into structural alignment enumeration of subsequences and solutions of these subproblems can be reused to enumerate structural alignments of bigger subsequences. As a consequence, the computation of the partition function Z and determination of the MAP structural alignment can be efficiently accomplished by dynamic programming.

Denoting by ${}_i^j\mathbf{x}$ the subsequence of $\mathbf{x}$ from indices i through j (in $5'$ to $3'$ order). The MAP structural alignment is then obtained by determining the minimum (pseudo) free energy structural alignment for subsequences ${}_i^j\mathbf{x}_1$ and ${}_k^l\mathbf{x}_2$, starting with all possible single nucleotide sequences ($i = j = 1, 2, \ldots, N_1$ and $k = l = 1, 2, \ldots, N_2$) and growing the length of the sequence by one in each dynamic programming step till the subsequences incorporate the full sequence. In implementation, memory savings can be accomplished by performing this in a two step process: A forward process that tracks minimum free energy for the "best" structural alignment of the subsequence and a traceback step which recovers the minimum free energy structural alignment $\mathcal{S}$. MAP estimates of the (common) secondary structures $\mathbf{S}_1$ and $\mathbf{S}_2$ and the alignment $\mathbf{A}$ can then be obtained from $\mathcal{S}$.

Computation of the *a posteriori* probabilities in (4) is more involved and requires computation of the partition function Z and the probability sums over all structural alignment in which i pairs with j. The latter objective is accomplished (through dynamic programming) by performing two sets of calculations corresponding to *internal* and *external* fragments of the sequence [1]. Denote by ${}_i^j\tilde{\mathbf{x}}$, the fragments of the sequence $\mathbf{x}$ excluding the nucleotide indices between $(i+1)$ and $(j-1)$, i.e. ${}_i^j\tilde{\mathbf{x}} = [{}_1^i\mathbf{x}, {}_j^N\mathbf{x}]$ where N is the length of $\mathbf{x}$ then

$$p_a^1(i,j) = \frac{1}{Z} \sum_{\substack{1 \leq k \leq N_2 \\ k < l \leq N2}} \alpha(i,j,k,l)\ \beta(i-1, j+1, k-1, l+1) \quad (5)$$

where

$$\alpha(i,j,k,l) \quad = \sum_{\mathcal{S}({}_i^j\mathbf{x}_1, {}_k^l\mathbf{x}_2) : \{(i,j) \in \mathbf{S}_1\}} e^{-\Delta G(\mathcal{S}({}_i^j\mathbf{x}_1, {}_k^l\mathbf{x}_2))} \quad (6)$$

$$\beta(i,j,k,l) \quad = \sum_{\mathcal{S}({}_i^j\tilde{\mathbf{x}}_1, {}_k^l\tilde{\mathbf{x}}_2)} e^{-\Delta G(\mathcal{S}({}_i^j\tilde{\mathbf{x}}_1, {}_k^l\tilde{\mathbf{x}}_2))} \quad (7)$$

The terms $\alpha(i,j,k,l)$ and $\beta(i,j,k,l)$ can be recursively computed through dynamic programming recursions analogous to the MAP case. By thresholding the base pairing probabilities at a suitably high threshold one can obtain base pairs predicted with high confidence. This is particularly useful for experimentalists looking to have predictions that contain pairs of high confidence.

Even though dynamic programming makes the structural alignment problem significantly simpler than brute force optimization, the computational complexity of the resulting algorithm is still rather high, $O(N^6)$, where N is the length of the smaller sequence. Thus, in practice, heuristic pruning of the search space is necessary to realize implementations that run in reasonable time on current hardware. We use a strategy that prunes the allowable space of base pairs and alignments by excluding very low probability alignment and base pairing positions [3], where each of these are based on the precomputed probabilities that form the input to the algorithm. A description of the recursions along with additional implementation details and results can be found in a companion paper[17].

5. RESULTS

We evaluate the proposed methods and compare their performance against free energy based single sequence structure prediction methods and [18, 11] and pure sequence alignment [3]. All methods are run over a test set of 2000 randomly chosen tRNA [19] pairs and 2000 randomly chosen 5S RNA [20] pairs. For each pair of (homologous) sequences, posterior probabilities for base pairing were computed for each of the sequences in the pair using a (single sequence) thermodynamic free energy model [11] and posterior probabilities of alignment states were computed using a pairwise HMM [3]. These were then utilized in the proposed algorithm in order to obtain: a) MAP estimates of S_1, S_2 the common secondary structures for the two sequences, b) the MAP estimate of the alignment A, and c) *a posteriori* probabilities of base pairing. The predictions were then scored against ground truth data for secondary structure and alignment obtained from the corresponding databases [19, 20].

Table 1 shows the overall accuracy of structure and alignment prediction of these methods. The prediction accuracy is reported in terms of sensitivity and positive predictive value (PPV). Sensitivity for structure prediction (alignment) represents the ratio of number of base pairs (aligned positions) that are predicted correctly to total number of base pairs (aligned positions) in the correct secondary structure (alignment). PPV for structural prediction (alignment) represents the ratio of number of base pairs (aligned positions) that are predicted correctly to total number of base pairs (aligned positions) in predicted structure (alignment). For the estimation of high confidence base pairs in the AP algorithm in Table 1, a threshold value of $\mathrm{P_{threshold}} = 0.9999$ was utilized. For the AP algorithm we do not currently obtain alignment estimates, the corresponding entries are therefore denoted by 'N/A' in Table 1. The entries in the alignment column of Table 1 for the single prediction method represent the sensitivity and PPV for MAP alignment using a pairwise hidden Markov model [3].

We also compare the proposed AP algorithm against single sequence AP [10, 11] by plotting the sensitivity vs PPV for different choices of $\mathrm{P_{threshold}}$. The resulting plot (analogous to an ROC curve) is shown in Fig. 1. From Fig. 1 it can be seen that the proposed (two sequence) AP algorithm offers a significant improvement.

[1] This is analogous to a forward-backward calculation for HMMs [15, 16]

		Structure	Alignment
MAP Algorithm	Sens	0.729	0.895
	PPV	0.775	0.900
Single Prediction	Sens	0.599	0.857
	PPV	0.541	0.860
AP Algorithm	Sens	0.309	**N/A**
($P_{\text{thresh}} = 0.9999$)	PPV	0.895	**N/A**
Single AP	Sens	0.027	**N/A**
($P_{\text{thresh}} = 0.9999$)	PPV	0.937	**N/A**

Table 1. Sensitivity and PPV for structure and alignment predictions. The MAP algorithm and AP algorithm rows correspond to the proposed method.

6. CONCLUSION AND DISCUSSION

The proposed algorithm for joint prediction of common secondary structure and alignment of two RNA sequences generalizes the constraints imposed in prior work on this problem and provides MAP estimates of common secondary structure and alignment, as well as *a posteriori* base pairing probabilities. The predictions obtained with the proposed algorithm offer a significant improvement in accuracy over single sequence secondary structure prediction and over pure sequence alignment.

The estimates of posterior probabilities of base pairing obtained from the proposed algorithm are particularly valuable since these can indicate high confidence base pairs that biologists can investigate further using experimental methods.

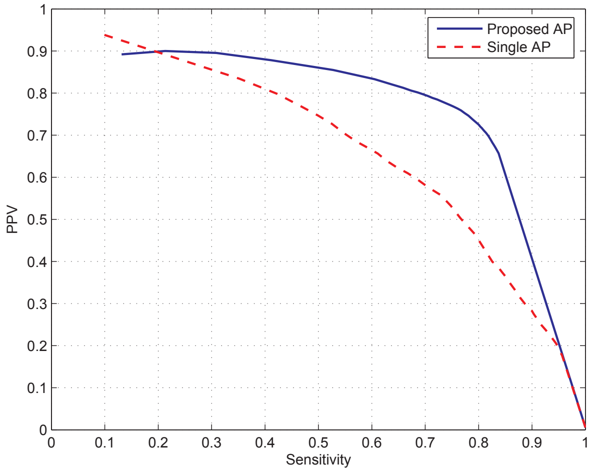

Fig. 1. Structure Prediction Sensitivity versus PPV for the proposed AP algorithm (solid) and single prediction (dashed) obtained by varying $P_{\text{threshold}}$.

7. REFERENCES

[1] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1999.

[2] Byung-Jun Yoon and P. P. Vaidyanathan, "RNA secondary structure prediction using context-sensitive hidden markov models," in *Biomedical Circuits and Systems, 2004 IEEE International Workshop on*, 1-3 Dec. 2004, pp. S2/7/INV–S2/7/1–4.

[3] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign," *BMC Bioinformatics*, vol. 8, no. 130, April 2007.

[4] A. Ozgun Harmanci, Gaurav Sharma, and David H. Mathews, "Toward turbo decoding of RNA secondary structure," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, Apr. 2007, vol. I, pp. 365–368.

[5] I. Holmes, "Accelerated probabilistic inference of RNA structure evolution," *BMC Bioinformatics*, vol. 6, no. 1, pp. 73, March 2005.

[6] D. H. Mathews, "Revolutions in RNA secondary structure prediction," *J Mol Biol*, vol. 359, pp. 526–532, 2006.

[7] D. Sankoff, "Simultaneous solution of RNA folding, alignment and protosequence problems," *SIAM J. App. Math.*, vol. 45, no. 5, pp. 810–825, Oct. 1985.

[8] J. H. Havgaard, E. Torarinsson, and J. Gorodkin, "Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix," *PLoS Computational Biology*, 2007, in press.

[9] R. D. Dowell and S. R. Eddy, "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction," *BMC Bioinformatics*, vol. 5, no. 1, pp. 71, 2004.

[10] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105 – 1119, November 1988.

[11] D. H. Mathews, "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization," *RNA*, vol. 10, pp. 1178–1190, 2004.

[12] Ivo L. Hofacker and Peter F. Stadler, "The partition function variant of Sankoff's algorithm," in *Lecture Notes in Computer Science, Computational Science - ICCS 2004*, pp. 728–735. Cold Spring Harbor Laboratory Press, 2004.

[13] Robert Giegerich, Björn Voß, and Marc Rehmsmeier, "Abstract shapes of RNA," *Nucleic Acids Res*, vol. 32, pp. 4834–4851, 2004.

[14] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms*, The MIT Press, second edition, September 2001.

[15] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Info. Theory*, vol. IT-20, no. 2, pp. 284–287, Feb. 1974.

[16] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[17] Arif O. Harmanci, Gaurav Sharma, and David H. Mathews, "PARTS: Probabilistic alignment for RNA joinTs secondary structure prediction," *Nucleic Acids Res*, 2007, Manuscript Accepted.

[18] D. H. Mathews, M. D. Disney, J. L. Childs, S. J Schroeder, M. Zuker, and D. H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," *Proc Natl Acad Sci USA*, vol. 101, pp. 7287–7292, 2004.

[19] M Sprinzl, C Horn, M Brown, A Ioudovitch, and S Steinberg, "Compilation of tRNA sequences and sequences of tRNA genes," *Nucleic Acids Res*, vol. 26, pp. 148–153, 1998.

[20] M Szymanski, M. Z. Barciszewska, J Barciszewski, and V. A. Erdmann, "5S ribosomal RNA database Y2K," *Nucleic Acids Res*, vol. 28, pp. 166–167, 2000.