TOWARD TURBO DECODING OF RNA SECONDARY STRUCTURE

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 iterative probabilistic algorithm for estimation of RNA secondary structure using sequence data from two homologous sequences. The method is intended to exploit inter-sequence correlations "encoded" in the form of probabilistic models for alignment and for common secondary structure. In analogy with turbo-decoding in digital communications, we formulate a maximum *a posteriori* probability objective function for joint structural prediction and sequence alignment using iterations over individual structural and sequential alignment models with soft-input soft-output estimators. As a preliminary step toward realizing this methodology, we present results obtained from incorporating (hard) constraints based on posterior sequence alignment probabilities in joint secondary structure prediction. Through experimental evaluations over available databases of known secondary structure, we demonstrate that this results in a significant decrease in computation time while simultaneously providing a marginal increase in structural prediction accuracy.

Index Terms— RNA, secondary structure , pairwise alignment, structural alignment, Iterative probabilistic decoding.

1. INTRODUCTION

Since the three-dimensional shape of biological molecules determines their physiological function, the estimation of molecular structure constitutes one of the fundamental problems in biology. Accurate estimates of structure can help in understanding interactions among different biomolecules, which in turn can assist in drug discovery and in development of alternate cures. Computational approaches that estimate the structure from more readily obtainable genome/proteome sequence data are particularly attractive in this respect because of their significantly lower cost in comparison to experimental procedures (e.g. X-ray crystallography). In this paper we address the problem of computationally estimating *secondary structure* for RNA molecules. Once the *primary structure* of an RNA molecule, i.e. the sequence of the bases adenine (A), guanine (G), cytosine (C), and uracil (U) that determines the linear chain forming the "backbone" of the molecule, is obtained from sequencing[1], the next step in the structural estimation progression is the determination of secondary structure, which is defined as the set of base pairings $A-U$, $G-C$, and $G-U$ formed through hydrogen bond interactions between the nucleotides in the linear chain. The sequential estimation process mirrors the hierarchy of RNA structure formation, which is commonly referred to as *folding* [2]. Examples of secondary structure for RNA molecules are shown in Fig 1.

Computational methods for RNA secondary structure prediction can be classified in two major categories based on the information they utilize: a) *Single sequence* prediction methods that attempt to infer the structure of a single RNA strand, or b) *Multi-Sequence* methods that work on multiple RNA sequences to infer their common homologous structure. Multi-sequence methods jointly perform the tasks of aligning and predicting the common secondary structure for the multiple sequences. The inter-sequence *comparative analysis* inherent in this process provides a major benefit and leads to significant improvement in structural prediction accuracy scores over single sequence prediction methods [3]. Computational requirements, on the other hand, are substantially higher for multi-sequence methods and grow with increasing number of sequences. Therefore, a majority of the current methods work with two sequences, though limited effort has also been directed to extending these methods to more than two sequences [4].

Current promising techniques for the prediction of RNA secondary structure are based either on thermodynamic models that predict common secondary structure using free-energy minimization as a predictor of structure likelihood [5] or on statistical learning techniques, primarily, stochastic context free grammars that provide probabilistic estimates by utilizing a model trained on a dataset with known alignment and secondary structure [6, 7]. In both cases, the problem is rendered computationally tractable by the use of dynamic programming, an approach first proposed by Sankoff [8] for the joint problem of alignment and secondary structure predic-

[1]For introductory background on sequencing, see for example [1].

tion for multiple sequences. As compared to the exponential complexity of brute force evaluation, Sankoff's algorithm and its aforementioned variants have a computational complexity of $O(N^6)$ in time and $O(N^4)$ in memory, where N is the length of the smaller of the two sequences. Despite the significant improvement (of polynomial complexity), for typical sequence lengths of interest, further constraints are normally required in order to make the run time reasonable on current hardware. The constraints range from a heuristic restriction of the alignment region to a band based on a guess of the maximum insertion length [8, 9] to restraints on alignment and folding to the union of the sets of K-best sub-optimal alignments and folds [7], for some choice of K.

2. PROBABILISTIC FRAMEWORK FOR JOINT ALIGNMENT AND SECONDARY STRUCTURE PREDICTION

Given two RNA sequences $\mathbf{x}_1$ and $\mathbf{x}_2$, a maximum likelihood (ML) formulation for the problem of joint alignment and secondary structure prediction is obtained as

$$(\mathcal{A}^*, \mathcal{S}^*) = \arg\max_{\mathcal{A},\mathcal{S}} \; p(\mathbf{x}_1\mathbf{x}_2 \mid \mathcal{A}, \mathcal{S}) \tag{1}$$

where $\mathcal{A}$, denotes an *alignment* of the two sequences and $\mathcal{S}$ denotes a *common* secondary structure. Formal definitions of alignment and common (secondary) structure may be found in [8]. We confine ourselves to observing that an alignment between the two RNA sequences arranges the bases in the sequences along a single "time-axis", as shown by an example in Fig. 2, where at each point there is a base from one or both of the sequences and gaps, denoted by $\sqcap$, occupy positions at where there is no nucleotide from a sequence. The term *common* secondary structure, following biological convention, refers to topological equivalence rather than exact match between the structures. Figure 1 illustrates two tRNA molecules with common secondary structure.

Although, dynamic programming allows a polynomial time solution for the ML formulation of joint alignment and secondary structure prediction, the approach suffers from a couple of limitations in practice. Firstly, though fairly sophisticated models exist for prediction of alignment, only relatively simple methods are readily integrated in the joint formulation (e.g. a linear or affine penalty for gaps and possibly mismatch penalties). Similarly, models for secondary structure of individual sequences are often more sophisticated than those used for the joint problem. Secondly, the computational complexity of the ML problem remains too high for practical deployment on typical RNA molecules and therefore practical variants of the Sankoff algorithm described in Section 1 perform only a restricted search. For example, in [7] the search is limited to the set of K-best (for some choice of K) alignments and folds (i.e. secondary structure) where the former are determined purely from sequence alignment models and the latter from single sequence folding methods.

As an alternative to the ML formulation, we propose a base-pair by base-pair maximum *a posteriori* probability (MAP) approach for the problem of joint secondary structure and alignment, working ultimately toward the development of an iterative decoding method that alternates between the alignment and structural models. The *a posteriori* probability of base pairing with respect to *sequence alignment model*, $\mathcal{M}_A$ and *structural model*, $\mathcal{M}_S$, can be written as

$$\mathrm{P}(i \;\square\; k \mid \mathbf{x}_1, \mathbf{x}_2, \mathcal{M}_A, \mathcal{M}_S) \tag{2}$$

where $\square$ denotes pairing of nucleotide positions i and j in sequences $\mathbf{x}_1$. A similar expression can be obtained for the base pairing probabilities in the second sequence $\mathbf{x}_2$. Likewise the *a posteriori* probability for alignment of nucleotide position i in the first sequence with nucleotide position k in the second sequence can be expressed as $\mathrm{P}(i \;\Leftrightarrow\; k \mid \mathbf{x}_1, \mathbf{x}_2, \mathcal{M}_A, \mathcal{M}_S)$, where $\Leftrightarrow$ denotes alignment.

Computation of these probabilities (and the MAP solution) is computationally demanding. We therefore propose a heuristic simplification by making analogy with iterative probabilistic (turbo) decoding techniques in digital communications: We treat the problems separately for the alignment and the secondary structure models and iterate over these through the exchange of soft information in order to obtain an approximate solution to the joint MAP problems. At each iteration, the resulting subproblems require calculation of the posterior probabilities $\mathrm{P}(i \Leftrightarrow k \mid \mathbf{x}_1, \mathbf{x}_2, \mathcal{M}_A)$ and $\mathrm{P}(i \Leftrightarrow k \mid \mathbf{x}_1, \mathbf{x}_2, \mathcal{M}_S)$ while incorporating soft information from each other in the form of "pseudo-priors". Furthermore, in yet another heuristic modification, we can incorporate more sophisticated models for the alignment and secondary structure than is feasible in the joint model. Note that we express these problems in terms of alignment posterior probabilities since the alignment forms the primary source of inter-sequence information, whereas the secondary structure is composed of intra-sequence base pairs. The posterior probability of base pairing may be evaluated once iterations are completed in order to obtain an estimated structure.

In order to accomplish iterative probabilistic decoding of secondary structure, we need soft-input soft-output estimators for both models. Hidden Markov models are a natural choice for the alignment model $\mathcal{M}_A$. In the next section, we present a brief outline of the computation of posterior alignment probabilities under the alignment model. The computation of posterior alignment and pairing probabilities under the structural model can be performed through a computation of the Boltzmann partition function that we intend to undertake in future work. As a preliminary result, we also apply the posterior probability estimates to obtain improved (though hard) constraints for current joint structure and alignment methods.

2.1. Posterior Pairwise Alignment Probabilities

Alignment between two sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ each representing bases along an RNA molecular chain can be effectively

modeled by a Hidden Markov Model [10, 3]. The Model uses 3 hidden states $M = \{\text{ALN, INS1, INS2}\}$, where each state outputs an ordered pair of symbols from the alphabet $\{A, C, G, U, \sqcap\}$. The first four symbols in output alphabet correspond to nucleotides and last symbol corresponds to a *gap*. The ordered outputs of each state form two sequences. Individual output sequences are denoted by lowercase boldface letters namely $\mathbf{x}_1$ and $\mathbf{x}_2$ for the case of two sequences. Specific nucleotides or subsequences selected from a sequence are indicated by prescripts: ${}_{n_1}\mathbf{x}$ denotes the $n_1{}^{\text{th}}$ nucleotide of the 1st sequence and ${}_{n_1}^{n_2}\mathbf{x}$ denotes the subsequence of nucleotides from index n_1 to n_2 in sequence $\mathbf{x}$.

Posterior probabilities, $\mathrm{P}(n_1 \leftrightarrow n_2, m \mid \mathbf{x}_1, \mathbf{x}_2)$, corresponding to co-incidence [11] of nucleotides n_1 and n_2 in state m, are efficiently computed in terms of recursions involving a *forward-variable* and a *backward-variable*. Denote by $S_m(n_1, n_2)$ the event that the state is m at the point when n_1 nucleotides corresponding to the first sequence and n_2 nucleotides corresponding to the second sequence have been emitted. The forward-variable is then defined as the joint probability

$$\alpha_m(n_1, n_2) = \mathrm{P}(S_m(n_1, n_2),\ {}_1^{n_1}\mathbf{x}_1,\ {}_1^{n_2}\mathbf{x}_2), \qquad (3)$$

i.e., the probability that the subsequence ${}_1^{n_1}\mathbf{x}_1$ of n_1 nucleotides is emitted in the first sequence, the subsequence ${}_1^{n_2}\mathbf{x}_2$ of n_2 nucleotides is emitted in the second sequence, and the state (of the alignment Markov process) is m. The backward variable is defined as the conditional probability

$$\beta_m(n_1, n_2) = \mathrm{P}({}_{n_1+1}^{N_1}\mathbf{x}_1,\ {}_{n_2+1}^{N_2}\mathbf{x}_2 \mid S_m(n_1, n_2)), \qquad (4)$$

i.e., the probability that subsequences ${}_{n_1+1}^{N_1}\mathbf{x}_1$ and ${}_{n_2+1}^{N_2}\mathbf{x}_2$ are observed given that state is m when n_1 and n_2 nucleotides have been emitted in the first and second sequence, respectively. Here N_1 and N_2 represent the lengths of sequences $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The forward variable can be computed recursively as:

$$\alpha_{m'}(n_1, n_2) = \sum_{m \in M} \tau(m, m') \cdot \gamma_{m'}(u_1(n_1, m'), u_2(n_2, m')) \cdot \pi_{m'}(n_1, n_2) \cdot \alpha_m(n_1'(n_1, m), n_2'(n_2, m)) \qquad (5)$$

where $n_1'(n_1, m)$ and $u_1(n_1, m)$ are:

$$(n_1', u_1)(n_1, m) = \begin{cases} (n_1, \sqcap) & \text{if } m = \text{INS2} \\ (n_1 - 1, {}_{n_1}\mathbf{x}_1) & \text{otherwise} \end{cases} \qquad (6)$$

and $n_2'(n_2, m)$ and $u_2(n_2, m)$ are similarly defined. The term $\tau(m, m')$ corresponds to probability of transition from state m to m', $\gamma_{m'}(\cdot)$ corresponds to emission probability of the ordered symbol pair $(u_1(n_1, m'), u_2(n_2, m'))$ by state m' and $\pi_{m'}(n_1, n_2)$ corresponds to *prior* probability of state at indices (n_1, n_2) estimated from structural alignment.

Recursions for the backward variable are similarly established.

	PPV	Sensitivity
Dynalign (Previous)	0.796	0.862
Single Prediction	0.609	0.687
Dynalign (Proposed)	0.803	0.865

Table 1. Average structural prediction accuracy statistics for the three methods over 2000 random tRNA and 2000 random 5S RNA alignments.

3. PROBABILISTIC CONSTRAINTS IN SECONDARY STRUCTURE PREDICTION

As a first step in iterative probabilistic estimation of secondary structure, we consider the incorporation of probabilistically derived hard constraints from sequence alignment model into secondary structure prediction. The pairwise hidden Markov model and constraint calculation is implemented [11]. For joint structure prediction we use the *Dynalign* [12] program.

Constraints are incorporated in Dynalign as a set of nucleotide position pairs from the two sequences that maybe co-incident [11]. Accordingly, constraints are defined by thresholding posterior co-incidence probabilities:

$$\mathbf{C} = \{(n_1, n_2) \mid \mathrm{P}(n_1 \leftrightarrow n_2 \mid \mathbf{x}_1, \mathbf{x}_2) > P_{\text{thresh}}\} \qquad (7)$$

where $\leftrightarrow$ denote co-incidence event and co-incidence probability is defined as the sum:

$$\mathrm{P}(n_1 \leftrightarrow n_2 \mid \mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{m \in M} \alpha_m(n_1, n_2)\beta_m(n_1, n_2)}{\sum_{m \in M} \alpha_m(N_1, N_2)} \qquad (8)$$

where N_1 and N_2 correspond to length of 1st and 2nd sequence respectively. Denominator in (8) corresponds to $P(\mathbf{x}_1, \mathbf{x}_2)$ The threshold probability P_{thresh} mediates a trade-off between the possibility of missing alignment positions and the computational complexity [11].

4. RESULTS

Probabilistic alignment constraints are tested on 2000 randomly chosen 5S RNAs and tRNA pairs for structural prediction accuracy. The performance is compared across three methods: a) Single sequence structure prediction [13] b) Dynalign with previous banded alignment constraints and c) Dynalign with the proposed probabilistic alignment constraints. Table 1 summarizes the average structural prediction accuracy results on these two sets.

Results are tested in terms of sensitivity ($\approx$ probability of detections) and positive predictive value (PPV) ($\approx 1 -$ false detection probability). It can be seen that there is a marginal increase in *both* sensitivity and positive predictive value with the proposed method. Table 2 summarizes timing requirements of probabilistic alignment constraints versus banded alignment constraints.

It can be seen that the proposed method offers a significant speedup. There is around $1.25 - 3$ timing saving on average

	tRNA	5S RNA
Dynalign (Previous)	38.53	236.03
Dynalign (Proposed)	30.77	84.95

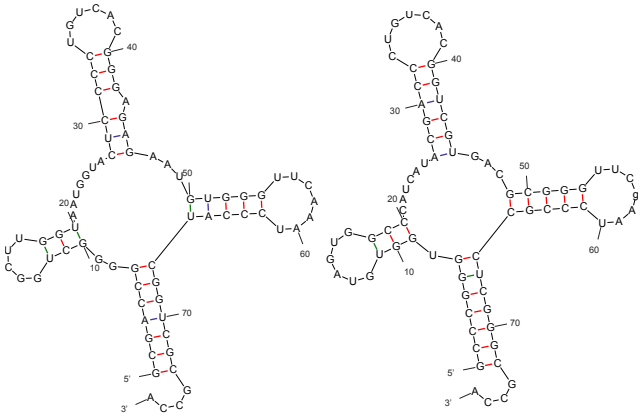
Table 2. Average CPU times (in seconds) for Dynalign with M (= 7) constraint and Dynalign with probabilistic alignment constraints over 100 randomly chosen 5S RNA and tRNA alignments each from [14] and [15]. A 3.0 GHz Intel Pentium 4 system with 1 GBytes of main memory running Linux Fedora Core 4 was utilized for the timing experiments.

using probabilistic alignment constraints with Dynalign. It should be noted that the timing gain for 5S RNAs is higher compared to that of tRNAs where 5S RNAs are about 1.5 times length of tRNAs on the average.

5. CONCLUSIONS

In this paper, we proposed an iterative approach for solving the problem of joint structure prediction and alignment for RNA sequences. The method is motivated by *soft-input soft-output* probabilistic iterations which are used in turbo decoding algorithms.

As a first step in this direction, we presented a method for determining constraints for joint prediction of secondary structure and alignment based on estimates of *a posteriori* coincidence probabilities estimated using a Hidden Markov Model. When integrated with Dynalign, an existing method for secondary structure prediction, the constraints significantly reduce computational requirements while simultaneously offering a small improvement in structural prediction accuracy. The speedup is particularly significant because it allows the algorithm to be deployed on larger sequences than was previously feasible.



(a) RD0260 (b) RD0500

Fig. 1. Common secondary structure for RD0260 and RD0500 tRNA molecules

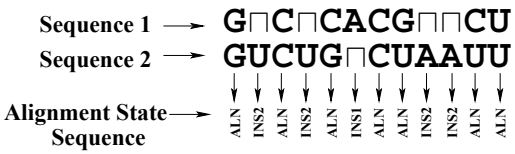


Fig. 2. A pairwise alignment and corresponding states

6. REFERENCES

[1] J. D. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick, *Molecular Biology of the Gene*, 5th ed. San Franciso, CA: Pearson Education, Benjamin Cummings, 2004.

[2] I. Tinoco, Jr. and C. Bustamante, "How RNA folds," *J Mol Biol*, vol. 293, no. 2, pp. 271–281, 1999.

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

[4] B. Masoumi and M. Turcotte, "Simultaneous alignment and structure prediction of three RNA sequences," *Int J Bioinformatics Research and Applications*, vol. 1, pp. 230–245, 2005.

[5] D. H. Mathews and D. H. Turner, "Dynalign: An algorithm for finding the secondary structure common to two RNA sequences," *Journal of Molecular Biology*, vol. 317, pp. 191–203, 2002.

[6] 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, p. 71, 2004.

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

[8] D. Sankoff, "Simultaneous solution of RNA folding, alignment and protosequence problems," *SIAM Journal of Applied Mathematics*, vol. 45, no. 5, pp. 810–825, Oct. 1985.

[9] A. V. Uzilov, J. M. Keegan, and D. H. Mathews, "Detection of noncoding RNAs on the basis of predicted secondary structure formation free energy change," *BMC Bioinformatics*, vol. 7, no. 1, p. 173, 2006.

[10] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *ASSPMAG*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[11] A. Harmanci, G. Sharma, and D. Mathews, "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign," *BMC Bioinformatics*, vol. 4, 2006, submitted for review Sept. 2006.

[12] D. H. Mathews, "Predicting a set of minimal free energy RNA secondary structures common to two sequences," *Bioinformatics*, vol. 21, no. 10, pp. 2246–2253, May 2005.

[13] M. Zuker, "Computer prediction of RNA structure," *Methods in Enzymology*, vol. 180, pp. 262–288, 1989.

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

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