PROBABILISTIC METHODS FOR IMPROVING EFFICIENCY OF RNA SECONDARY STRUCTURE PREDICTION ACROSS MULTIPLE SEQUENCES

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

[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

Prediction of common secondary structure across multiple RNA sequences is known to significantly increase accuracy in comparison with single-sequence based prediction methods. However, the computational requirements for joint prediction can often be daunting in comparison to single-sequence prediction. As a result, heuristic simplifications are often necessary for this joint estimation problem in order to perform computations on current hardware in reasonable times. In this paper, principled heuristics are presented for the purpose of computation reduction based on probabilistic methods. The methods presented eliminate the computations over extremely improbable alignments and structures, thereby reducing computation with little or no degradation in accuracy. Experimental results over databases of RNA families with known secondary structure validate our methods, demonstrating over a two-fold computational speed up in tests over the 5S rRNA family, without any compromise in accuracy.

Index Terms— RNA secondary structure, posterior base pairing probability, hidden Markov model

1. INTRODUCTION

One of the major developments in biology in recent years has been the discovery of new functions for ribonucleic acids (RNAs). It was once believed that RNA molecules were merely intermediate copies of parts of the genetic information residing in DNA (deoxy-ribonucleic acid) that were created for the purpose of protein synthesis. More recently, it has been realized that RNA is a central player in cellular biology and serves a number of direct functions in addition to the conventional roles of messenger RNAs and tRNAs in protein synthesis. In these direct roles, the RNAs are not "coding for proteins" and the corresponding RNAs are therefore referred to as *noncoding* RNAs (ncRNAs) [1, 2].

As is the case for most biomolecules, the three-dimensional structure of an ncRNA determines its function and therefore a determination of ncRNA structure is key problem in biology. The structure of RNA is determined by interactions among the atoms that form the molecule and also by interactions with other molecules that are in their vicinity in cellular physiological conditions. The interactions vary in strength and accordingly a hierarchy is seen in RNA structure [3] typically arranged in order of decreasing strength of the interactions. The *primary structure* of RNA is a linear chain of four different types of nucleotides that are joined together by covalent phosphodiester bonds [4]. The four types of nucleotides can be identified by their nitrogenous bases adenine (A), guanine (G), cytosine (C), and uracil (U), and accordingly, the primary structure can be specified as a sequence of these four bases. Within a RNA molecular chain, nucleotides pair through the formation of hydrogen bonds between (some of the) complementary nitrogenous bases: specifically, A can pair with U, G can pair with C, and G can pair with U. The set of the $A - U$, $G - C$, and $G - U$ pairings is referred to as the *secondary structure* of the RNA molecule. Additional interactions among the molecules in an RNA chain beyond

the secondary structure base pairings define the *tertiary structure* and interactions with other molecules such as proteins and other strands of RNA are classified as *quaternary structure*. The intramolecular[1] structure at a lower level of this hierarchy can be determined without involving the higher levels since the interactions become progressively weaker [5, 6, 7].

The primary structure of RNA is experimentally determined through sequencing [4]. In the coding role of RNA in protein synthesis, the function of the RNA molecule is determined largely by this primary structure, whereas for ncRNAs the complete three-dimensional structure is desired. Experimental determination of the complete three-dimensional structure is however rather difficult and expensive. Therefore computational methods for the prediction of the secondary, tertiary, and quaternary structure are of significant research interest. The methods for prediction follow the hierarchy of RNA structure and the first step in this process is the prediction of secondary structure from sequence data representing the primary structure, which is the focus of the current paper.

2. JOINT PREDICTION OF RNA STRUCTURE ACROSS MULTIPLE SEQUENCES

For ncRNAs, often multiple sequences exist with the same structure and function. These are referred to as *homologs*. At the secondary structure level, it can be seen that the replacement one of the base pairs $A - U$, $G - C$, or $G - U$ by another base pair from this set does not change the topology of the secondary structure. These base pair substitutions, referred to as *compensatory mutations*, are actually observed in nature in homologous RNA molecules. In addition, other minor changes in bases and base pairs may be encountered in homologous sequences without a change in secondary structure. Fig. 1 illustrates an example of this, where two homologous tRNA molecules are shown, which differ in their sequence data but have common (topology of) secondary structure.

The accuracy of RNA secondary structure prediction can be significantly improved by determining the (common) secondary structure across two or more homologous sequences [8]. The process can be viewed as simultaneously aligning the two sequences while constraining them to *folding* into a common secondary structure. This problem was mathematically formulated first by Sankoff [9] who proposed a dynamic programming solution for addressing this problem. Though the space of common secondary structures for two sequence is exponential in the shorter sequence length [10], dynamic programming makes polynomial time (in the length of the shorter of the two sequences) algorithms possible for determining the most likely common secondary structure for a pair of RNA sequences.

---

[1]Interactions governing tertiary and quaternary structure have similar strengths but the quaternary interactions are bimolecular and may need to overcome an entropic cost in order to change the underlying tertiary structures of the interacting molecules.
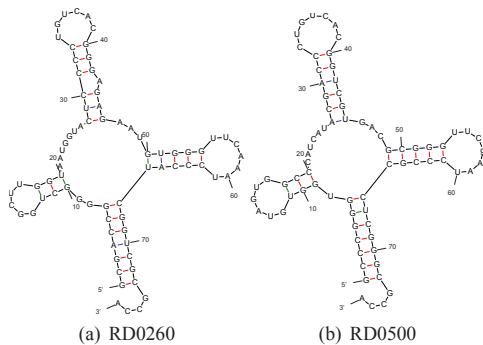
(a) RD0260   (b) RD0500

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

3. PRUNING OF SEARCH SPACE IN JOINT PREDICTION OF RNA STRUCTURE

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. As part of his original algorithm description, Sankoff [9] proposed limiting computations to a banded region motivated by a limitation on the maximum allowable insertion length. Such banded computations have been employed in practical implementations of the algorithm [11, 12]. One limitation with this methodology is that the maximum insertion lengths encountered can vary significantly from one ncRNA family to another requiring manual adjustment of this parameter. Additionally, for ncRNA families with longer sequences, often longer insertion lengths are encountered, which reduces the potential savings in computation for these cases even though there is greater need for computational saving.

The preceding comments indicate that data adaptive approaches to pruning the search space are likely to be more effective than methods based on a purely ad hoc heuristic. Along these lines, an alternate methodology for pruning computation has been based on determining, for some choice of positive integers k and l, the k most likely folds for the individual sequences and l most likely alignments for the sequences based on a simple sequence alignment model (that does not account for common secondary structure). These selected folds and alignments define corresponding fold and alignment *envelopes* as the regions they cover in the folding and alignment spaces, respectively, to which the joint computation can be constrained [13]. Though well-motivated and adaptive to the sequence data, the method has the disadvantage that in regions where the alignment or folding have low confidence, the choices of k and l determined from computational considerations may not be adequate. Yet another approach for pruning computation is employed in Consan [14], where, based on a hidden Markov model (HMM) for pairwise sequence alignment, nucleotide positions called "pins" are determined that are judged to be aligned with high confidence. These are then forced to be aligned for the joint alignment and secondary structure estimation. The method has the limitation that there are no constraints apart from the pins and no pins may be found in sequence pairs with low sequence conservation.

3.1. Search Space Pruning using *a posteriori* Probabilities

Models for single sequence folding based on thermodynamic stability and those for pairwise sequence alignment based on HMMs can provide not only estimates of the most likely (or k most likely) folds and alignments, but these models can also provide estimates of the posterior probabilities of base pairing and of nucleotide alignment, respectively. The algorithms for determining the posterior probabilities are closely

related to those for determining the most likely estimates and have the same order of computational complexity, just as the Viterbi [15] and BCJR [16] algorithms for error correction coding share strong similarities and have same order of complexity [2].

The above observation suggests an alternative for pruning of the search space. Posterior probabilities of fold and alignment events can be determined using relatively simple computational models for these individually [18, 19, 20]. The search space for the computationally demanding joint alignment and folding problem can then be restricted to regions over which these probabilities are higher than a pre-set threshold. If the threshold is set fairly low, these constraints exclude only the consideration of highly improbable base pairing and alignment states. In regions where the folding and alignment are known with high confidence, computation is restricted to narrowly constrained regions whereas in regions where the folding and alignment are poorly resolved, a wider range of possibilities are allowed (for the joint problem). This methodology constitutes a principled data adaptive heuristic that concentrates the joint computation in regions where it is required. The application of this idea is described next specifically with respect to the pruning of the allowed alignment space for the Dynalign algorithm. This work was recently reported in [21]. Constraints based on posterior probabilities of base pairing have also been developed, albeit for a different algorithm for joint alignment and secondary structure prediction [22].

4. ALIGNMENT CONSTRAINTS FROM POSTERIOR PROBABILITIES

Hidden Markov models (HMMs) provide effective probabilistic models for the alignment of protein and DNA sequences where sequence information is conserved for homologs [20]. For ncRNA sequences, as remarked earlier, homologs demonstrate conservation of secondary structure and can often show significant divergence in sequence data. The divergence in sequence, however, is often localized and conservation of sequence information is seen in large parts of homologous ncRNAs. For this reason, HMM based models are still effective in identifying the alignment of ncRNAs in regions with high sequence conservation and in identifying the regions in which sequence information alone is unreliable for alignment. This characteristic makes HMMs well suited to the task of pruning the alignment space: in regions where the sequence information alone restricts the alignment to a narrow band with high confidence, the alignment space may be pruned to the corresponding region and in regions where the sequence information alone does not provide a reliable estimate, a larger search space should be allowed for the structure based joint alignment and folding problem. As outlined next, the computation of posterior alignment probabilities allows this dynamic pruning of the search space.

4.1. Pairwise Alignment HMM

A pairwise alignment HMM models the alignment between two RNA sequences as a chain of states from a three state Markov chain, where the three states correspond to alignment (ALN) between the two sequences, nucleotide insertion in the first sequence (INS1), and nucleotide insertion in the second sequence (INS2). In the ALN state the model emits a nucleotide in each of the sequences, in the INS1 state it emits a nucleotide in the first sequence alone and in the INS2 state it emits a nucleotide in the second sequence alone. Each emitted nucleotide (in any of the states) takes on one of the values A, U, G, and C. The nucleotide sequences are observed, whereas the underlying alignment states are not and constitute the hidden part of the HMM. The state transition probabilities between the three alignment states, initial probabilities over the states, and the emission probabilities for the possible nucleotide emissions in each of the states constitute the parameters of the HMM alignment model. Given the values of these parameters and two sequences $\mathbf{s}_1$, $\mathbf{s}_2$, the *a posteriori* probability of an alignment $\mathbf{x}$ can be readily defined. As a specific example, consider the two sequences and postulated alignment between

---

[2]A recent paper [17] formalizes the basis for this similarity.

$$p(\mathbf{s}_1, \mathbf{s}_2, \mathbf{x}) = \pi_0(ALN)p_{ALN}^e(G,G)p^t(ALN, INS1)p_{INS1}^e(U)p^t(INS1, ALN)p_{ALN}^e(C,C)\cdots p^t(ALN, ALN)p_{ALN}^e(C,C) \quad (1)$$

these as shown in Fig. 2. The joint probability of the postulated alignment and the emission of the two sequences can be obtained for this case as shown in (1), where for $m, m_2 \in \{ALN, INS1, INS2\}$ and $x, y \in \{A, U, G, C\}$, $\pi_0(m)$ denotes the initial probability of the alignment state m, $p_{ALN}^e(x, y)$ denotes the probability of emitting nucleotide x in the first sequence and nucleotide y in the second sequence in an aligned state, $p^t(m, m_2)$ denotes the alignment Markov chain state transition probability that the next state is m_2 given than the current state is m, $p_{INS1}^e(x)$ is the probability of emitting the nucleotide x in the first sequence in the INS1 state, and $p_{INS2}^e(x)$ is the probability of emitting the nucleotide x in the second sequence in the INS2 state.

Fig. 2. Two RNA sequences and a hypothetical alignment.

The posterior probability of the alignment given the two sequences is then obtained from Bayes' rule and one can readily see that the maximum aposteriori probability (MAP) alignment can equivalently be obtained by maximizing $p(\mathbf{s}_1, \mathbf{s}_2, \mathbf{x})$. Additionally, using the model the posterior probability that the alignment state m occurs at nucleotide index n_1 along the first sequence and nucleotide index n_2 along the second sequence as

$$p(n_1 \overset{m}{\leftrightarrow} n_2 \mid \mathbf{s}_1, \mathbf{s}_2) = \frac{1}{p(\mathbf{s}_1, \mathbf{s}_2)} \sum_{\mathbf{x}:n_1 \overset{m}{\leftrightarrow} n_2 \in \mathbf{x}} p(\mathbf{s}_1, \mathbf{s}_2, \mathbf{x}) \quad (2)$$

Both the MAP alignment and the posterior probabilities can be efficiently computed in $O(n^2)$ time using dynamic programming algorithms, utilizing respectively, the Viterbi algorithm and the HMM forward-backward algorithm [23]. Specific details of the posterior probability computations for the pairwise alignment HMM can be found in [21].

4.2. Co-incidence Probabilities and Alignment Constraints

Observe that whenever a nucleotide is emitted in a sequence it contributes to an elongation of the corresponding sequence by exactly 1 nucleotide. Thus an alignment between the two sequences $\mathbf{s}_1$ and $\mathbf{s}_2$, having lengths N_1 and N_2, respectively, can be equivalently specified by a (connected) path in a 2-D $(N_1 + 1) \times (N_2 + 1)$ array. The path begins at $(0, 0)$ and ends at (N_1, N_2), at each successive position along the path the the first array index is incremented if the alignment state is ALN or $INS1$ and the second array index is incremented if the alignment state is ALN or $INS2$. Mathematically, the path is defined by a sequence of ordered-pairs of nucleotide indices $(i_1, k_1), (i_2, k_2), \ldots (i_L, k_L)$, where $0 = i_1 \le i_2 \le i_L = N_1$, $0 = k_1 \le k_2 \le k_L = N_2$, and $1 \le (i_l - i_{l-1}) + (k_l - k_{l-1}) \le 2$. The first value in each ordered-pair represents a nucleotide index in the the first sequence $\mathbf{s}_1$ and the second value in each ordered-pair represents a nucleotide index of the second sequence $\mathbf{s}_2$. The value index 0 represents the beginning of the sequences, allowing for insertions in either sequence before the start of the other. Under the specified alignment, for each $l = 1, 2, \ldots L$, the nucleotide position i_l in the first sequence is said to be *co-incident* with the nucleotide position k_l in the second sequence [21].

Dynamic programming algorithms for the joint prediction of RNA secondary structure across two sequences, can be thought of as jointly searching for the optimum solution over the combination of the folding spaces representing possible foldings for each of the two sequences and the alignment space for the inter-sequence alignments [13]. In particular, the connected path representation for an alignment indicated in the preceding paragraph for possible alignments forms the basis of the dynamic programming with respect to alignment: If any point (i, j)

lies on the path representing an alignment in the 2-D array representation, at least one of the "preceding" points $(i-1, j-1)$, $(i, j-1)$, or $(i-1, j)$ in the 2-D array must also be on the alignment path. This implies that when restricting the search space for alignment for reducing computation, a specific alignment will be allowable under the alignment constraints if all co-incident pairs of nucleotides in the alignment are allowed by the constraints. Thus alignment constraints are often implemented as boolean $(N_1 + 1) \times (N_2 + 1)$ arrays that indicate the nucleotide positions for which co-incidence is allowed. As a specific example, an implementation of Dynalign that preceded the present work [12] used the following banded constraint for the alignment space

$$\left| \frac{i \times N_2}{N_1} - k \right| \le M \quad (3)$$

where M specifies the width of the banded region.

Instead of the static band constraint, a data adaptive constraint on the alignment space can be obtained by utilizing the HMM based probabilistic model for sequence alignment presented in Section 4.1. Specifically, using the posterior probabilities for the alignment states in (2), the posterior probability that nucleotide position n_1 in $\mathbf{s}_1$ is coincident with nucleotide position n_2 in $\mathbf{s}_2$ is readily obtained as

$$p(n_1 \leftrightarrow n_2 \mid \mathbf{s}_1, \mathbf{s}_2) = \sum_m p(n_1 \overset{m}{\leftrightarrow} n_2 \mid \mathbf{s}_1, \mathbf{s}_2) \quad (4)$$

An example of computed posterior coincidence probabilities for two tRNA sequences is shown in Fig. 3, where the probabilities are plotted on a (natural) log scale. From the figure it is apparent that at the lower and upper diagonals of the matrix representing the alignment space, almost the entire the probability mass is contained in very narrow bands. This occurs because the sequences are highly conserved in these regions. In the middle region, however, there is an insertion in the first sequence and correspondingly the probability mass is dispersed over a wider region.
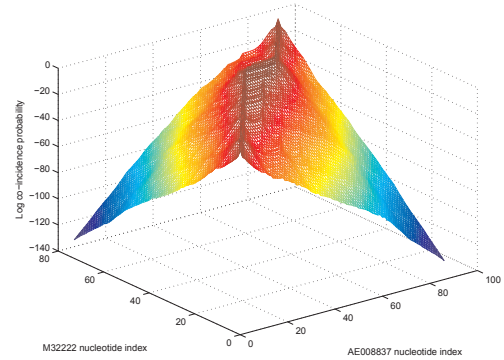

Fig. 3. Posterior co-incidence probabilities for two RNA sequences.

An alignment constraint that excludes only the nucleotide coincidence probabilities for which the probability is very small is obtained by selecting a sufficiently small threshold probability P_{thresh} and allowing all co-incidence possibilities $n_1 \leftrightarrow n_2$ for which $p(n_1 \leftrightarrow n_2 \mid \mathbf{s}_1, \mathbf{s}_2) > P_{thresh}$. The resulting set of allowed co-incidence possibilities is shown as the dark gray region in Fig. 4(a). In Fig. 4(b) the alignment path as manually determined by Biologists is overlaid on this constraint in black. It can be seen that the constraint allows the true alignment, which is desirable. A comparison of the probabilistically derived constraint set against the static banded constraint is instructive, the latter is shown in Fig. 4(c) for a typical value of $M = 6$. The difference between the two sets is shown in Fig. 4(d). From these

figures it can be seen that the probabilistically derived constraint set is data data adaptive: in regions of high sequence conservation, where the HMM narrows the alignment to a narrow region with high confidence, computation is restricted to a rather narrow band and in regions close to the insertion in the first sequence, a wider band of computation is allowed. Thus as compared to the banded computation the method concentrates the computation where it is required. Based on the fact that the light gray areas in Fig. 4(d) dominate the dark gray areas, it can also be conjectured that the method would provide an overall saving in computation.

5. EXPERIMENTAL RESULTS

The efficacy of the probabilistically derived alignment constraints as a method for pruning computations was evaluated by integrating these constraints in the Dynalign [11] algorithm for predicting the secondary structure common to two RNA sequences [21]. Dynalign with the probabilistically derived alignment constraints was also compared against the Dynalign with the banded constraint defined by (3) with $M = 7$, which was utilized in Dynalign prior to this work [12].

The two versions of Dynalign were compared with respect to the accuracy of the predicted structures and with respect to the execution times and memory requirements. For the evaluation of the accuracy of predicted structures, a dataset of 309 5S RNAs [24] and 484 tR-NAs [25] with known secondary structures is utilized. For each of the tRNA and 5S rRNA families, 2000 pairs of sequences were selected at random from the, respective, databases and the common secondary structures of the sequence pairs were predicted by each of the methods. The predictions were then evaluated against the known secondary structures in terms of their *sensitivity* and *positive predictive value* (PPV). The sensitivity is defined as the fraction of base pairings in the true secondary structure that are predicted (correctly) by the algorithm and PPV is defined as the fraction of the base pairings predicted by the algorithm that are present in the known structure. Execution time and memory estimates were obtained as averages over a randomly selected set of 100 tRNA and 5S RNA sequence pairs each selected at random from the RFAM database [26].

Table 1 compares the accuracy of the secondary structures predicted in terms of sensitivity and PPV. For the purpose of comparison, the accuracy of the predictions obtained using a single sequence based prediction of secondary structure [27] is also included in the Table. From the tabulated values it can be seen that Dynalign with the probabilistically derived constraint and Dynalign with the banded constraint perform comparably in terms of their sensitivity and PPV, with a minor (though not statistically significant) advantage for the version with the probabilistic constraint. Both versions significantly outperform the single sequence prediction.

Table 2 lists the average execution times and memory requirements of the methods for joint prediction of the common secondary structure of two sequences for the tRNA and 5S rRNA datasets. The tRNA data set had an average sequence length of approximately 77 nucleotides and the 5S rRNA dataset had an average sequence length of approximately 120 nucleotides. A comparison of the execution times for the 5S rRNA dataset demonstrates that the probabilistically derived alignment constraint sets cut computation time to less than half the value required by the banded constraint. This reduction is all the more remarkable since it comes without any reduction in accuracy (as already demonstrated in Table 1). For the tRNA dataset the methods are quite comparable with the banded constraint providing slightly faster execution. The gains for the 5S rRNA sequences are however more significant since the execution times are significantly larger for these sequences than for the tRNAs owing to their longer lengths. The results in Table 2 also indicate that the probabilistically derived alignment constraint sets also reduce the memory requirements, though only by a relatively modest amount.
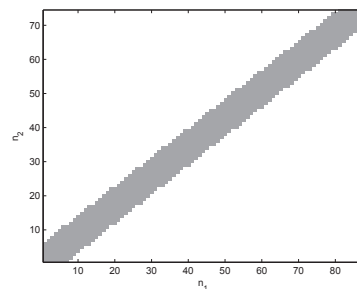
Often results for structure prediction accuracy are stratified by sequence percent identity. These can be found in [21], which also includes minimum and average statistics for execution times and mem-
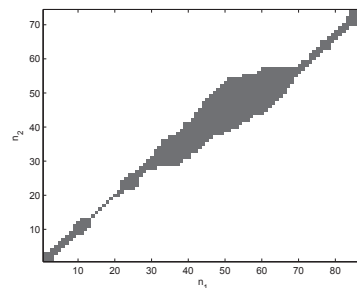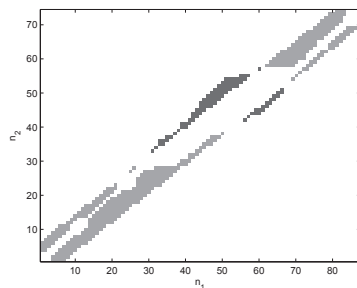

(a) Probabilistic alignment constraint
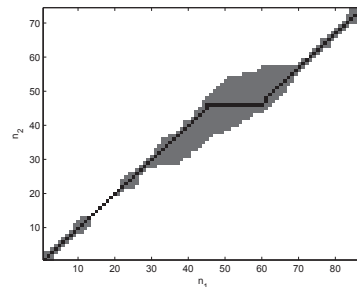
(b) Alignment constraint & true alignment

(c) Banded constraint

(d) Difference between probabilistic and banded alignment constraints

Fig. 4. Comparison of the probabilistic alignment constraint set against the true alignment and the banded computation constraint set for tRNA sequences: AE008837 and M32222. The abscissa and ordinates of the plots indicate nucleotide positions n_1 and n_2 along the sequences AE008837 and M32222, respectively. (a) Probabilistic alignment constraint set: permitted alignments shown in dark gray. (b) Alignment constraint set with true alignment super-imposed (in black). (c) Alignment constraint set for the prior banded constraint with $M = 6$ (shown in light gray). (d) Difference between the banded constraint and the probabilistic alignment constraint sets. Light gray regions indicate nucleotide position alignments permitted by the banded constraint and not by the probabilistic constraint and the situation is vice-versa for dark gray regions.

ory in an attempt to quantify the variability in these. In particular, as might be expected from the data dependency of the probabilistically derived alignment constraint sets, greater variation is seen with these than with the static banded constraint. Comparisons against a number of other methods for RNA secondary structure prediction using two RNA homologs are also included in [21].

		tRNA	5S rRNA
Dynalign Prob. constraint	Sens	0.861	0.871
	PPV	0.834	0.785
Dynalign banded constraint	Sens	0.855	0.870
	PPV	0.825	0.782
Single Prediction	Sens	0.748	0.709
	PPV	0.693	0.618

Table 1. Structural prediction accuracy for the different methods over 2000 random tRNA selections from[25] and 2000 5S rRNA selections from [24]. Dynalign Prob. constraint refers to Dynalign with probabilistic alignment constraints.

		Timing	Memory
Dynalign Prob. constraint	tRNA	9.98	10.988
	5S rRNA	34.38	12.277
Dynalign band constraint	tRNA	9.39	10.960
	5S rRNA	73.07	14.306

Table 2. Average execution times and memory requirements (in seconds and megabytes of main memory, respectively) of the structure prediction methods on 5S RNAs and tRNAs alignments from [24] and [25]. Based on a dual-core AMD Opteron®-270 2.0 GHz system with 4 GBytes of main memory running Linux Fedora Core 4.

6. CONCLUSIONS AND FUTURE WORK

Pruning of the computational search space for joint alignment and secondary structure prediction can be performed in a principled data adaptive fashion by computing posterior probabilities for alignment and folding from individual models and thresholding these to exclude highly improbable regions from the joint computation. Results demonstrate that methods for constraining alignment based on this idea offer a significant reduction in computation for Dynalign without compromising accuracy. Over a 5s RNA family dataset with an average sequence length of 120 nucleotides the method offers more than two-fold speed-up.

Joint ncRNA structure prediction across two or more sequences can be employed in a variety of applications. One application of particular interest is the scanning of genomes in order to search for novel ncRNA genes [12]. This is a computationally demanding task since genomes can be fairly large (the human genome, for instance, has over three billion base pairs). A very significant speed-up of the joint algorithms is therefore necessary in order to speed up searches for ncRNA genes. While pruning of the search space helps in this respect, alternative approaches that offer greater potential for speed-up are also worthy of exploration. In particular, turbo-decoding style iterative approaches for solving the joint problem by iterating over the individual problems with feedback may offer an attractive alternative in this respect [28].

7. REFERENCES

[1] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nat. Rev.*, vol. 2, pp. 919–929, Dec. 2001.

[2] ——, "Computational genomics of noncoding RNA genes," *Cell*, vol. 109, pp. 137–140, Apr. 2002.

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

[4] 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.

[5] D. M. Crothers, P. E. Cole, C. W. Hilbers, and R. G. Schulman, "The molecular mechanism of thermal unfolding of Escherichia coli formylmethionine transfer RNA," *J Mol Biol*, vol. 87, pp. 63–88, 1974.

[6] A. R. Banerjee, J. A. Jaeger, and D. H. Turner, "Thermal unfolding of a group i ribozyme: The low temperature transition is primarily a disruption of tertiary structure," *Biochemistry*, vol. 32, pp. 153–163, 1993.

[7] D. H. Mathews, A. R. Banerjee, D. D. Luan, T. H. Eickbush, and D. H. Turner, "Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposon element," *RNA*, vol. 3, pp. 1–16, 1997.

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

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

[10] R. Giegerich, B. Voß, and M. Rehmsmeier, "Abstract shapes of RNA," *Nucleic Acids Res*, vol. 32, pp. 4834–4851, 2004.

[11] D. H. Mathews and D. H. Turner, "Dynalign: An algorithm for finding the secondary structure common to two RNA sequences," *J Mol Biol*, vol. 317, pp. 191–203, 2002.

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

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

[14] 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.

[15] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Info. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[16] 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.

[17] S. Aji and R. McEliece, "The generalized distributive law," *IEEE Trans. Info. Theory*, vol. 46, no. 2, pp. 325–343, Mar. 2000.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] ——, "PARTS: Probabilistic alignment for RNA joinTs secondary structure prediction," *Nucleic Acids Res*, 2007, in preparation for submission to Nucleic Acids Research.

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

[24] 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.

[25] 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.

[26] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "RFAM: An RNA family database," *Nucleic Acids Res*, vol. 31, no. 1, pp. 439–441, September 2002.

[27] 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.

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