biRNA: Fast RNA-RNA Binding Sites Prediction

Hamidreza Chitsaz[1], Rolf Backofen[2], and S. Cenk Sahinalp[1]

[1] School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada
{hrc4, cenk}@cs.sfu.ca
[2] Institut für Informatik, Albert-Ludwigs-Universität, Georges-Koehler-Allee,
Freiburg, Germany
backofen@informatik.uni-freiburg.de

Abstract. We present biRNA, a novel algorithm for prediction of binding sites between two RNAs based on minimization of binding free energy. Similar to RNAup approach [29], we assume the binding free energy is the sum of accessibility and the interaction free energies. Our algorithm maintains tractability and speed and also has two important advantages over previous similar approaches: 1) it is able to predict multiple simultaneous binding sites and 2) it computes a more accurate interaction free energy by considering both intramolecular and intermolecular base pairing. Moreover, biRNA can handle crossing interactions as well as hairpins interacting in a zigzag fashion. To deal with simultaneous accessibility of binding sites, our algorithm models their joint probability of being unpaired. Since computing the exact joint probability distribution is intractable, we approximate the joint probability by a polynomially representable graphical model namely a Chow-Liu tree-structured Markov Random Field. Experimental results show that biRNA outperforms RNAup and also support the accuracy of our approach. Our proposed Bayesian approximation of the Boltzmann joint probability distribution provides a powerful, novel framework that can also be utilized in other applications.

1 Introduction

RNA had been viewed as a simple working copy of the genomic DNA, simply transporting information from the genome into the proteins, until the discovery of ribozymes and the realization that the ribosome is in fact an RNA machine. Following the recent discovery of RNA interference (RNAi), the post transcriptional silencing of gene expression via interactions between mRNAs and their regulatory RNAs, RNA has moved from a side topic to a central research topic.

More recent studies have shown that a large fraction of the genome gives rise to RNA transcripts that do not code for proteins [39]. Several of these non-coding RNAs (ncRNAs) regulate gene expression post-transcriptionally through base pairing (and establishing a joint structure) with a target mRNA, as per the eukaryotic miRNAs and small interfering RNAs (siRNAs), antisense RNAs or bacterial small regulatory RNAs (sRNAs) [16]. In addition to such endogenous regulatory ncRNAs, antisense oligonucleotides have been used as exogenous inhibitors of gene expression; antisense technology is now commonly used as a research tool as well as for therapeutic purposes. Furthermore, synthetic nucleic acids systems have been engineered to self assemble into complex structures performing various dynamic mechanical motions [45].

A key tool for all the above advances is a fast, highly accurate computational method for predicting RNA-RNA interactions. Comprehensive methods for analyzing binding thermodynamics of nucleic acids are computationally expensive and prohibitively slow for real applications [1, 9]. Other existing methods suffer from a low specificity, possibly because several of these methods consider restricted versions of the problem (e.g. simplified energy functions or restricted types of interactions) - this is mostly for computational reasons.

In this paper we present an algorithm to predict the binding sites of two interacting RNA strands. Our most important goal in this work is tractability as well as high specificity. While

our algorithm considers the most general type of interactions, it is still practically tractable by making simplifying assumptions on the energy function. These assumptions are however natural and adopted by many other groups as well [1, 6, 22, 29, 40]. Our experiments also support these assumptions.

Our contribution

We give an algorithm to predict the binding sites of two interacting RNAs and also the interaction secondary structure constrained by the predicted binding sites. As opposed to previous approaches that are able to predict only one binding site [6, 29, 41], our algorithm predicts multiple simultaneous binding sites. We define a binding site to be a subsequence which interacts with exactly one binding site in the other strand. Crossing interactions (external pseudoknots) and zigzags (see [1] for exact definition) are particularly allowed. To the best of our knowledge, this allows for the most general type of interactions considered in the literature. Although intramolecular pseudoknots are not considered in the current work, they can be incorporated into our framework at the expense of additional computational complexity.

Following RNAup approach [29], we assume the total interaction free energy is the sum of two terms: 1) the free energy needed to make binding sites accessible in each molecule, and 2) the free energy released as a result of intermolecular bonds formed between the interacting binding site pairs. Based on that energy model, our algorithm is essentially composed of three consecutive steps: 1) building a tree-structured Markov Random Field (MRF) to approximate accessibility of a collection of potential binding sites, 2) computing pairwise interaction free energies between potential binding sites of one strand against those of the other strand, and 3) finding a minimum free energy matching of binding sites. Unlike RNAup that computes only the hybridization partition function for step 2, our algorithm computes the full interaction partition function [9]. Therefore, our algorithm not only considers multiple binding sites but also computes a more accurate free energy of binding.

The time complexity of the first two steps is $O(n^3r + m^3s + nmw^4)$ in which n and m denote the lengths of sequences, w denotes maximum binding site length, and $r \leq nw$ and $s \leq mw$ are the number of potential sites heuristically selected out of the $O(nw)$ and $O(mw)$ possible ones. More importantly, the space complexity of the first two steps is $O(r^2 + s^2 + nmw^2)$. The third step requires a nontrivial optimization namely minimum energy bipartite matching of two tree-structured Markov Random Fields, a topic on which we are currently working. In this paper, we implement an exhaustive search for the third step. Therefore, the running time of the third step is currently $O(r^\kappa s^\kappa)$ where κ is the maximum number of simultaneous binding sites. Since r and s are small in our experiments, an exhaustive search over single, pair, and triple sites is feasible. However, we are working on the matching problem and hope to either find an efficient algorithm for it or prove its hardness.

Related work

Since the initial works of Nussinov et al. [30] and Waterman and Smith [44] several computational methods have emerged to study the secondary structure thermodynamics of a single nucleic acid molecule. Those initial works laid the foundation of modern computational methods by adopting a divide and conquer strategy. That view, which originally exhibited itself in the form of a simple

base pair counting energy function, has evolved into Nearest Neighbor Thermodynamic model which has become the standard energy model for a nucleic acid secondary structure [26]. The standard energy model is based on the assumption that stacking base pairs and loop entropies contribute additively to the free energy of a nucleic acid secondary structure. Based on additivity of the energy, efficient dynamic programming algorithms for predicting the minimum free energy secondary structure [30, 36, 44, 46] and computing the partition function of a single strand [14, 27] have been developed.

Some previous attempts to analyze the thermodynamics of multiple interacting nucleic acids concatenate input sequences in some order and consider them as a single strand. For example, `pairfold` [2] and `RNAcofold` from Vienna package [4] concatenate the two input sequences into a single strand and predict its minimum free energy structure. Dirks et al. present a method, as a part of `NUPack`, that concatenates the input sequences in some order, carefully considering symmetry and sequence multiplicities, and computes the partition function for the whole ensemble of complex species [13]. However, concatenating the sequences is not accurate at all as even if pseudoknots are considered, some useful interactions are excluded while many physically impossible interactions are included. Several other methods, such as `RNAhybrid` [34], `UNAFold` [12, 24], and `RNAduplex` from Vienna package [4], avoid intramolecular base-pairing in either strand and compute minimum free energy hybridization secondary structure. Those approaches naturally work only for simple cases involving typically very short strands.

Alternatively, a number of studies aimed to take a more fundamental stance and compute the minimum free energy structure of two interacting strands under energy models with growing complexity. For instance, Pervouchine devised a dynamic programming algorithm to maximize the number of base pairs among interacting strands [32]. A followup work by Kato et al. proposed a grammar based approach to RNA-RNA interaction prediction [18]. More generally, Alkan et al. [1] studied the joint secondary structure prediction problem under three different models: 1) base pair counting, 2) stacked pair energy model, and 3) loop energy model. Alkan et al. proved that the general RNA-RNA interaction prediction under all three energy models is an NP-hard problem. Therefore, they suggested some natural constraints on the topology of possible joint secondary structures which are satisfied by all examples of complex RNA-RNA interactions in the literature. The resulting algorithms compute the minimum free energy secondary structure among all possible joint secondary structures that do not contain (internal) pseudoknots, crossing interactions (i.e. external pseudoknots), and *zigzags* (please see [1] for the exact definition). In our previous work [9], we gave an algorithm `piRNA` to compute the partition function, base pair probabilities, and minimum free energy structure over the type of interactions that Alkan et al. considered. We extended the standard energy model for a single RNA to an energy model for the joint secondary structure of interacting strands by considering new types of (joint) structural components. Although `piRNA` outperforms existing alternatives, it has $O(n^6)$ time and $O(n^4)$ space complexity which is prohibitive for many practical, particularly high-throughput, applications.

A third set of methods predict the secondary structure of each individual RNA independently, and predict the (most likely) hybridization between accessible regions of the two molecules. More sophisticated ones in this set view interaction as a multi step process [6, 29, 43]: 1) unfolding of the two molecules to expose bases needed for hybridization, 2) the hybridization at the binding site, and 3) restructuring of the complex to a new minimum free energy conformation. Some approaches in this set, such as `IntaRNA` [6] and `RNAup` [29], assume that binding happens at one location. Therefore, they are able to predict only one binding site, which is not the case for some

known interacting RNAs such as OxyS-fhlA and CopA-CopT [3, 20, 21]. In this paper, we consider multiple binding sites.

2 Preliminaries

Our algorithm is based on the assumption that binding is practically a stepwise process, a view that has been proposed by others as well [1, 6, 22, 29, 40]. In real world, each nucleic acid molecule has a secondary structure before taking part in any interaction. To form an interaction, as the first step the individual secondary structures are deformed so that the binding sites in both molecules become unpaired. As the second step, pairwise matching between the binding sites takes place. Each step is associated with an energy the sum of which gives the free energy of binding ΔG. Specifically, denote the energy difference that is needed for unpairing *all the binding sites* in **R** and **S** by ED_u^R and ED_u^S respectively, and denote by ΔG_b^{RS} the free energy that is released as a result of binding. Similar to [29],

$$\Delta G = ED_u^R + ED_u^S + \Delta G_b^{RS}. \tag{1}$$

This assumption is intuitively plausible because each molecule needs to reveal its interacting parts before the actual binding happens; moreover, these two steps are assumed to be independent from one another. Note that previous approaches such as `IntaRNA` [6], `RNAup` [29], and `RNAplex` [41] consider only one binding site in each molecule, which makes the problem easier, whereas we consider multiple binding sites. It is sometimes argued that nature does not usually favor too highly entangled structures [29]. Our algorithm easily accommodates an upper bound on the number of potential binding sites, which is another advantage of our approach.

To reduce the complexity, we assume that the length of a potential binding site is not more than a window size w in this work. This is a reasonable assumption, which has also been made in similar approaches [29], as most known RNA-RNA interactions such as OxyS-fhlA and CopA-CopT do not exhibit lengthy binding sites [3, 20, 21]. We call a subsequence of length not more than w a *site*.

3 Algorithm

Based on the assumption above, our program `biRNA` finds a combination of binding sites that minimizes ΔG. Let $\mathcal{V}^R$ and $\mathcal{V}^S$ denote the set of potential binding sites, which in our case is the collection of subsequences of length not more than w, of **R** and **S** respectively. `biRNA` is composed of five consecutive steps:

(I) For every site $W = [i, j]$ in $\mathcal{V}^R$ or $\mathcal{V}^S$, compute the probability $P_u^R(W)$ or $P_u^S(W)$ that W is unpaired.

(II) For every pair W_1 and W_2, compute the joint probabilities $P_u^R(W_1, W_2)$ and $P_u^S(W_1, W_2)$ that sites W_1 and W_2 are simultaneously unpaired.

(III) Build tree-structured Markov Random Fields (MRF) $\mathcal{T}^R = (\mathcal{V}^R, \mathcal{E}^R)$ and $\mathcal{T}^S = (\mathcal{V}^S, \mathcal{E}^S)$ to approximate the joint probability distribution of multiple unpaired sites. Denote the $\mathcal{T}$-approximated joint probability of unpaired sites $W_1, W_2, \ldots, W_k$ by $P_u^*(W_1, W_2, \ldots, W_k)$.

(IV) Compute $Q^I_{W^R W^S}$, the interaction partition function restricted to subsequences W^R and W^S, for every $W^R \in \mathcal{V}^R$ and $W^S \in \mathcal{V}^S$.

(V) Find a non-overlapping matching $M = \{(W_1^R, W_1^S), (W_2^R, W_2^S), \ldots, (W_k^R, W_k^S)\}$ that minimizes $\Delta G(M) = ED_u^R(M) + ED_u^S(M) + \Delta G_b^{RS}(M)$, in which

$$ED_u^R(M) = -RT \log P_u^{R*}(W_1^R, W_2^R, \ldots, W_k^R) \tag{2}$$

$$ED_u^S(M) = -RT \log P_u^{S*}(W_1^S, W_2^S, \ldots, W_k^S) \tag{3}$$

$$\Delta G_b^{RS}(M) = -RT \sum_{1 \le i \le k} \log(Q^I_{W_i^R W_i^S} - Q_{W_i^R} Q_{W_i^S}). \tag{4}$$

Above, R is the universal gas constant and T is temperature. To demonstrate (2) and (3), let for instance $P_u^R(W_1^R, W_2^R, \ldots, W_k^R)$ be the exact probability that the sites are unpaired. In that case, $ED_u^R(W_1^R, W_2^R, \ldots, W_k^R) = \Delta G^R(W_1^R, W_2^R, \ldots, W_k^R) - \Delta G^R$, and

$$\begin{aligned} &\Delta G^R(W_1^R, W_2^R, \ldots, W_k^R) - \Delta G^R = -RT \log Q_{\mathbf{R}}(W_1^R, W_2^R, \ldots, W_k^R) + RT \log Q_{\mathbf{R}} \\ &= -RT \log \frac{Q_{\mathbf{R}}(W_1^R, W_2^R, \ldots, W_k^R)}{Q_{\mathbf{R}}} = -RT \log P_u^R(W_1^R, W_2^R, \ldots, W_k^R), \end{aligned} \tag{5}$$

in which $Q_{\mathbf{R}}$ is the partition function of $\mathbf{R}$ and $Q_{\mathbf{R}}(W_1^R, W_2^R, \ldots, W_k^R)$ is the partition function of the structures in which $W_1^R, W_2^R, \ldots, W_k^R$ are unpaired.

In the following, we describe each step in more details. In Section 3.1 we explain (I) and (II) above. Section 3.2 is dedicated to (III) and also inference in tree-structured Markov Random Fields namely computing P_u^*. In Section 3.3 we describe (IV). Finally, (V) is presented in Section 3.4.

3.1 Accessibility of Site Pairs

As part of `RNAup`, Mückstein et al. present an efficient algorithm for computing the probability of an unpaired subsequence [29]. Their algorithm computes the probability of being unpaired for all subsequences in $O(n^3)$ time in which n is the sequence length. Based on `RNAup` algorithm, we present an $O(n^4 w)$ time and $O(n^2)$ space complexity algorithm to compute the joint probabilities of all unpaired site pairs. For every site, our algorithm uses constrained McCaskill's [27] and constrained `RNAup` algorithm to compute the conditional probabilities of all other unpaired subsequences. There are $O(nw)$ sites for each of which the algorithm takes $O(n^3)$ time. For triple joint probabilities, the same method is applicable but the running time will be multiplied by another factor of $O(nw)$. Therefore, we only compute pairwise probabilities and approximate the whole joint probability distribution by graphical models.

3.2 Simultaneous Accessibility of Multiple Sites

To deal with simultaneous accessibility of binding sites, we must model their joint probability of being unpaired. One way is to compute the exact joint probability distribution by using constrained McCaskill's algorithm [22]. For every collection of sites, the algorithm has polynomial time complexity, however, since there is an exponential number of different collections, this naïve approach is intractable. In this paper, we approximate the joint probability by a polynomially representable graphical model namely a Markov Random Field. Graphical models, including Bayesian Networks and Markov Random Fields, are powerful tools for approximating joint probabilities. They generally have enough expressive power which intuitively explains why the inference problem for

general graphical models is NP-hard [11]. Fortunately, there is an efficient inference algorithm for tree-structured models [31]. In this work, we build a Chow-Liu tree, which is a tree-structured Markov Random Field, to approximate the exact joint probability distribution [10].

To describe the Chow-Liu algorithm, let $\mathcal{G}$ be the complete weighted graph on $\mathcal{V}$, the set of potential binding sites, in which the weight of an edge between W_1 and W_2 is $I(W_1,W_2)$, the mutual information given by

$$I(W_1,W_2) = \sum_{\substack{x_1\in\{W_1,\sim W_1\}\\ x_2\in\{W_2,\sim W_2\}}} P(x_1,x_2)\log\left(\frac{P(x_1,x_2)}{P(x_1)P(x_2)}\right). \tag{6}$$

Above, $P(W_1,\sim W_2)$ is for instance the joint probability that W_1 is unpaired and W_2 is not unpaired. In Section 3.1, we explained how to compute the joint probabilities of all site pairs. The following equations use Bayes rule to calculate all the necessary terms from $P(W_1,W_2)$: $P(W_1,\sim W_2) = P(W_1) - P(W_1,W_2)$, $P(\sim W_1,W_2) = P(W_2) - P(W_1,W_2)$, $P(\sim W_1,\sim W_2) = 1 - P(W_1) - P(W_2) + P(W_1,W_2)$. The Chow-Liu tree $\mathcal{T}$ is the best tree-structured approximation for a joint probability distribution, in the sense that $\mathcal{T}$ has the maximum mutual information with the joint probability distribution [10]. Chow and Liu proved that $\mathcal{T}$ is the maximum spanning tree of $\mathcal{G}$. To compute $\mathcal{T}$, we use a standard maximum spanning tree algorithm such as Chazelle's algorithm [7]. We refer the reader to [17] or [31] for a detailed description of inference algorithm in $\mathcal{T}$. In summary, $P^*_u(W_1,W_2,\ldots,W_k)$ is computed by marginalizing over $\mathcal{V}\backslash\{W_1,W_2,\ldots,W_k\}$ the joint probability distribution defined by $\mathcal{T}$. There exists an efficient algorithm for inference which runs in $O(|\mathcal{V}|)$ time [17].

3.3 Free Energy of Interaction

The local free energy of interaction for a pair of sites W^R and W^S is $-RT\log(Q^I_{W^RW^S} - Q_{W^R}Q_{W^S})$ in which Q^I is the interaction partition function restricted to W^R and W^S [9] and Q is McCaskill's partition function restricted to W. Note that a simple version of Q^I would calculate only the hybridization partition function between W^R and W^S (see [29]); however, this would exclude any intramolecular structure in the binding sites. For that reason, we use our approach for Q^I which considers intermolecular as well as intramolecular structures. Our modified algorithm is the dynamic programming in [9] that starts with $l_R = 1, l_S = 1$ and incrementally computes all the recursive quantities up to $l_R = w, l_S = w$. Therefore, the windowed version of our interaction partition function algorithm has $O(nmw^4)$ time and $O(nmw^2)$ space complexity, in which n and m are the lengths of **R** and **S** respectively. Finally, for a non-overlapping matching $M = \{(W^R_1,W^S_1),(W^R_2,W^S_2),\ldots,(W^R_k,W^S_k)\}$ the free energy of interaction is $\Delta G^{RS}_b(M) = -RT\sum_{1\le i\le k}\log(Q^I_{W^R_iW^S_i} - Q_{W^R_i}Q_{W^S_i})$, where $Q^I_{W^R_iW^S_i} - Q_{W^R_i}Q_{W^S_i}$ is the partition function for those structures that constitute at least one intermolecular bond. Note that is based on the simplifying assumption in Section 3. A fundamental way to compute the free energy of interaction for a matching of binding sites is through the interaction partition function and the probabilities of Q^I tables. However, that approach is intractable for long RNA sequences mainly due to $O(n^2m^2)$ space requirement. In this paper, we replace the true free energy of interaction with local one, which is shown to be a reasonable approximation in practice (see the results in Section 4).

3.4 Binding Sites Matching

Having built the machinery to compute ΔG for a matching of binding sites, we would like to find a matching that minimizes ΔG. To clarify the importance and difficulty of the problem, suppose the binding sites were independent so that $P_u(W_1, W_2, \ldots, W_k) = P_u(W_1)P_u(W_2)\cdots P_u(W_k)$. In that case, the problem would reduce to finding a minimum weight bipartite matching with $\text{weight}(W_i^R, W_j^S) = -RT\log[P_u^R(W_i^R)P_u^S(W_j^S)(Q^I_{W_i^R W_j^S} - Q_{W_i^R}Q_{W_j^S})]$. There are efficient algorithms for minimum weight bipartite matching, but the issue is that the independence assumption is too crude of an approximation. Therefore, we propose the following problem, which has not been solved to our knowledge:

Minimum Weight Chow-Liu Trees Matching Problem

Given a pair of Chow-Liu trees $\mathcal{T}^R = (\mathcal{V}^R, \mathcal{E}^R)$ and $\mathcal{T}^S = (\mathcal{V}^S, \mathcal{E}^S)$, compute a non-perfect matching M between the nodes of $\mathcal{T}^R$ and $\mathcal{T}^S$ that minimizes $\Delta G(M)$.
Input: Chow-Liu trees $\mathcal{T}^R$ and $\mathcal{T}^S$.
Output: A matching $M = \{(W_1^R, W_1^S), (W_2^R, W_2^S), \ldots, (W_k^R, W_k^S)\} \subset \mathcal{V}^R \times \mathcal{V}^S$.

The complexity of minimum weight Chow-Liu trees matching problem is currently unknown. We are working on the problem, and we hope to either prove its hardness or give a polynomial algorithm; we incline toward the latter though. In this paper, we implemented an exhaustive search on the set of all collections of single, pair, and triple sites.

3.5 Complexity Analysis

Let n denote the length of **R**, m denote the length of **S**, and w denote the window length. Step (I) of the algorithm takes $O(n^3 + m^3)$ time and $O(n^2 + m^2)$ space. If we consider all site pairs, then (II) takes $O(n^4w + m^4w)$ time and $O(n^2w^2 + m^2w^2)$ space to store the joint probabilities. It is often reasonable to filter potential binding sites, for example based on the probability of being unpaired or the interaction partition function with another site in the other molecule. Suppose r potential sites out of nw possible ones and s sites out of mw ones are selected. In that case, (II) takes $O(n^3r + m^3s)$ time and $O(r^2 + s^2)$ space. Step (III) takes $O(r^2\alpha(r^2, r) + s^2\alpha(s^2, s))$ time where α is the classical functional inverse of the Ackermann function [7]. The function α grows extremely slowly, so that for all practical purposes it may be considered a constant. Step (IV) takes $O(nmw^4)$ time and $O(nmw^2)$ space. In this paper, we implement an exhaustive search for (V). Therefore, its running time is currently $O(r^\kappa s^\kappa)$ where κ is the maximum number of simultaneous binding sites. Therefore, the algorithm takes $O(n^3r + m^3s + nmw^4 + r^\kappa s^\kappa)$ time and $O(r^2 + s^2 + nmw^2)$ space. Note that $r \leq nw$ and $s \leq mw$, so that the algorithm has $O(n^4w + m^4w + n^2m^2w^4)$ time and $O(n^2w^2 + m^2w^2)$ space complexity without heuristic filtering, considering maximum two simultaneous binding sites. We are working on (V) and hope to either find an efficient algorithm for it or prove its hardness.

Interaction Structure Prediction Once the binding sites are predicted, a constrained minimum free energy structure prediction algorithm predicts the secondary structure of each RNA. For each binding site pair, a windowed version of our algorithm in [9] completes the interaction structure prediction.

4 Results

To evaluate the performance of biRNA, we used the program to predict the binding site(s) of 20 bacterial sRNA-mRNA interactions studied in the literature. Since all these sRNAs bind their target in close proximity to the ribosome binding site at the Shine-Dalgarno (SD) sequence, we restricted the program to a window of maximum 250 bases around the start codon of the gene. We compared our results with those obtained by RNAup on the same sequences with the same window size $w = 20$. The results are summarized in Table 1. RNAup and biRNA have generally very close performance for the cases with one binding site. However, biRNA outperforms RNAup in some single binding site cases such as GcvB-gltI. That is because biRNA uses the interaction partition function as opposed to the hybridization partition function used by RNAup. The interaction partition function is more accurate than the hybridization partition function as the interaction partition function accounts for both intermolecular and intramolecular base pairing [9]. Also, biRNA significantly outperforms RNAup for OxyS-fhlA and CopA-CopT which constitute more than one binding site. We noticed that our predicted energies are generally lower than those predicted by RNAup which may be due to different energy parameters. We used our piRNA energy parameters [9] which in turn are based on UNAFold v3.6 parameters [24].

We implemented biRNA in C++ and used OpenMP to parallelize it on Shared-Memory multiProcessor/core (SMP) platforms. Our experiments were run on a Sun Fire X4600 Server with 8 dual AMD Opteron CPUs and 64GB of RAM. The sequences were 71-253 nt long (see the supplementary materials for sequences) and the running time of biRNA with full features was from about 10 minutes to slightly more than one hour per sRNA-mRNA pair. The biRNA software will be available and also as a webserver at http://compbio.cs.sfu.ca/taverna/.

5 Conclusions and Future Work

In this paper, we presented biRNA, a new thermodynamic framework for prediction of binding sites between two RNAs based on minimization of binding free energy. Similar to RNAup approach, we assume the binding free energy is the sum of the energy needed to unpair all the binding sites and the interaction free energy released as a result of binding.

Our algorithm is able to predict multiple binding sites which is an important advantage over previous approaches. More importantly, our algorithm can handle crossing interactions as well as zigzags (hairpins interacting in a zigzag fashion, see [1]). To assess the performance of biRNA, we compared its predictions with those of RNAup for 20 bacterial sRNA-mRNA pairs studied in the literature. The results were presented in Table 1. As it was expected, biRNA outperforms RNAup for those RNA pairs that have multiple binding sites such as OxyS-fhlA and CopA-CopT. Moreover, biRNA performs slightly better than RNAup for those pairs that have only one binding site because biRNA accounts for intramolecular as well as intermolecular base pairing in the binding sites.

To deal with simultaneous accessibility of binding sites, our algorithm models their joint probability of being unpaired. Since computing the exact joint probability distribution is intractable, we approximate the joint probability by a polynomially representable graphical model namely a tree-structured Markov Random Field. Chow-Liu algorithm efficiently builds such tree model [10]. Computing a joint probability in the Chow-Liu tree is performed by efficient marginalization algorithms [31]. Eventually, two Chow-Liu trees, pertaining to the two input RNAs, are matched to find the minimum binding free energy matching. To the best of our knowledge, the complexity of

Pair		Binding Site(s) Literature		biRNA Site(s)		$-\Delta G$	RNAup Site(s)		$-\Delta G$	Ref.
GcvB	gltI	[66,77]	[44,31]	(64,81)	(44,26)	11.5	(75,93)	(38,19)	18.7	[38]
GcvB	argT	[75,91]	[104,89]	(71,90)	(108,90)	13.1	(72,91)	(107,89)	20.2	[38]
GcvB	dppA	[65,90]	[150,133]	(62,81)	(153,135)	14.7	(62,81)	(153,135)	23.5	[38]
GcvB	livJ	[63,87]	[82,59]	(66,84)	(73,54)	13.1	(71,90)	(67,49)	14.9	[38]
GcvB	livK	[68,77]	[177,165]	(67,86)	(175,156)	12.2	(67,86)	(175,157)	19.0	[38]
GcvB	oppA	[65,90]	[179,155]	(67,86)	(176,158)	9.3	(67,86)	(176,158)	15.3	[38]
GcvB	STM4351	[70,79]	[52,44]	(69,77)	(52,44)	9.6	(69,87)	(52,33)	17.7	[38]
MicA	lamB	[8,36]	[148,122]	(8,26)	(148,131)	6.1	(8,27)	(148,129)	12.9	[5]
MicA	ompA	[8,24]	[128,113]	(8,24)	(128,113)	14.0	(8,24)	(128,113)	19.4	[33]
DsrA	rpoS	[8,36]	[38,10]	(21,40)	(25,7)	9.4	(13,32)	(33,14)	16.3	[35]
RprA	rpoS	[33,62]	[39,16]	(40,51)	(32,22)	4.3	(33,51)	(39,22)	10.7	[23]
IstR	tisA	[65,87]	[79,57]	(66,85)	(78,59)	18.1	(66,85)	(78,59)	29.0	[42]
MicC	ompC	[1,30]	[139,93]	(1,16)	(119,104)	18.5	(1,16)	(119,104)	18.7	[8]
MicF	ompF	[1,33]	[125,100]	(14,30)	(118,99)	8.0	(17,33)	(116,100)	14.7	[37]
RyhB	sdhD	[9,50]	[128,89]	(22,41)	(116,98)	15.8	(22,41)	(116,98)	21.5	[25]
RyhB	sodB	[38,46]	[60,52]	(38,46)	(64,48)	9.7	(38,57)	(60,45)	10.3	[15]
SgrS	ptsG	[157,187]	[107,76]	(174,187)	(89,76)	14.5	(168,187)	(95,76)	22.9	[19]
Spot42	galK	[1,61]	[126,52]	(1,8) (25,37) (46,60)	(128,119) (86,73) (64,53)	20.5	(27,46)	(84,68)	14.6	[28]
OxyS	fhlA	[22,30] [98,104]	[95,87] [45,39]	(23,30) (96,104)	(94,87) (48,39)	7.9	- (96,104)	- (48,39)	10.3	[3]
CopA	CopT	[22,33] [48,56] [62,67]	[70,59] [44,36] [29,24]	(22,31) (49,57) (58,67)	(70,61) (43,35) (33,24)	25.9	- (49,67) -	- (43,24) -	23.9	[20]

Table 1. Binding sites reported in the literature and predicted by biRNA and RNAup with window size $w = 20$. ΔG is in kcal/m. Two RNAs interact in opposite direction, hence, sites in the second RNA are presented in reverse order. See the supplementary materials for sequences.

minimum weight Chow-Liu trees matching problem is currently unknown. We are working on the problem, and we hope to either prove its hardness or give a polynomial algorithm. In this paper, we implemented an exhaustive search on the set of all collections of single, pair, and triple sites.

Our proposed Bayesian approximation of the Boltzmann joint probability distribution provides a novel powerful framework which can also be utilized in individual and joint RNA secondary structure prediction algorithms. As graphical models allow for models with increasing complexity, our proposed Bayesian framework may inspire more accurate but tractable RNA-RNA interaction prediction algorithms in future work.

References

1. Can Alkan, Emre Karakoc, Joseph H. Nadeau, S. Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13(2):267–282, 2006.
2. M. Andronescu, Z.C. Zhang, and A. Condon. Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, 345:987–1001, Feb 2005.
3. L. Argaman and S. Altuvia. fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J. Mol. Biol.*, 300:1101–1112, Jul 2000.
4. S.H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P.F. Stadler, and I.L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1:3, 2006.
5. L. Bossi and N. Figueroa-Bossi. A small RNA downregulates LamB maltoporin in Salmonella. *Mol. Microbiol.*, 65:799–810, Aug 2007.

6. Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 2008.
7. Bernard Chazelle. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *J. ACM*, 47(6):1028–1047, 2000.
8. S. Chen, A. Zhang, L. B. Blyn, and G. Storz. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J. Bacteriol.*, 186:6689–6697, Oct 2004.
9. Hamidreza Chitsaz, Raheleh Salari, S.Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):in press, 2009.
10. C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.
11. Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell.*, 42(2-3):393–405, 1990.
12. Roumen A. Dimitrov and Michael Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophysical Journal*, 87:215–226, 2004.
13. Robert M. Dirks, Justin S. Bois, Joseph M. Schaeffer, Erik Winfree, and Niles A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review*, 49(1):65–88, 2007.
14. Robert M. Dirks and Niles A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.
15. T. A. Geissmann and D. Touati. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.*, 23:396–405, Jan 2004.
16. Susan Gottesman. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics*, 21(7):399–404, 2005.
17. M.I. Jordan and Y. Weiss. Graphical models: probabilistic inference. In M. Arbib, editor, *Handbook of Neural Networks and Brain Theory*. MIT Press, 2002.
18. Yuki Kato, Tatsuya Akutsu, and Hiroyuki Seki. A grammatical approach to RNA-RNA interaction prediction. *Pattern Recognition*, 42(4):531–538, 2009.
19. H. Kawamoto, Y. Koide, T. Morita, and H. Aiba. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol. Microbiol.*, 61:1013–1022, Aug 2006.
20. F. A. Kolb, H. M. Engdahl, J. G. Slagter-Jger, B. Ehresmann, C. Ehresmann, E. Westhof, E. G. Wagner, and P. Romby. Progression of a loop-loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. *EMBO J.*, 19:5905–5915, Nov 2000.
21. F. A. Kolb, C. Malmgren, E. Westhof, C. Ehresmann, B. Ehresmann, E. G. Wagner, and P. Romby. An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA*, 6:311–324, Mar 2000.
22. Z. J. Lu and D. H. Mathews. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.*, 36:640–647, Feb 2008.
23. N. Majdalani, D. Hernandez, and S. Gottesman. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol. Microbiol.*, 46:813–826, Nov 2002.
24. N.R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.
25. E. Massé and S. Gottesman. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 99:4620–4625, Apr 2002.
26. D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, May 1999.
27. J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
28. T. Møller, T. Franch, C. Udesen, K. Gerdes, and P. Valentin-Hansen. Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev.*, 16:1696–1706, Jul 2002.
29. U. Mückstein, H. Tafer, S. H. Bernhart, M. Hernandez-Rosales, J. Vogel, P. F. Stadler, and I. L. Hofacker. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In Mourad Elloumi, Josef Küng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *BIRD*, volume 13 of *Communications in Computer and Information Science*, pages 114–127. Springer, 2008.
30. R. Nussinov, G. Piecznik, J. R. Grigg, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.
31. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, September 1988.
32. D.D. Pervouchine. IRIS: intermolecular RNA interaction search. *Genome Inform*, 15:92–101, 2004.
33. A. A. Rasmussen, M. Eriksen, K. Gilany, C. Udesen, T. Franch, C. Petersen, and P. Valentin-Hansen. Regulation of ompA mRNA stability: the role of a small regulatory RNA in growth phase-dependent control. *Mol. Microbiol.*, 58:1421–1429, Dec 2005.

34. M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, Oct 2004.
35. F. Repoila, N. Majdalani, and S. Gottesman. Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Mol. Microbiol.*, 48:855–861, May 2003.
36. E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, Feb 1999.
37. M. Schmidt, P. Zheng, and N. Delihas. Secondary structures of Escherichia coli antisense micF RNA, the 5'-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry*, 34:3621–3631, Mar 1995.
38. C. M. Sharma, F. Darfeuille, T. H. Plantinga, and J. Vogel. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev.*, 21:2804–2817, Nov 2007.
39. Gisela Storz. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–3, 2002.
40. H. Tafer, S. L. Ameres, G. Obernosterer, C. A. Gebeshuber, R. Schroeder, J. Martinez, and I. L. Hofacker. The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, 26:578–583, May 2008.
41. Hakim Tafer and Ivo L. Hofacker. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–2663, 2008.
42. J. Vogel, L. Argaman, E. G. Wagner, and S. Altuvia. The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr. Biol.*, 14:2271–2276, Dec 2004.
43. S.P. Walton, G.N. Stephanopoulos, M.L. Yarmush, and C.M. Roth. Thermodynamic and kinetic characterization of antisense oligodeoxynucleotide binding to a structured mRNA. *Biophys. J.*, 82:366–377, Jan 2002.
44. M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc*, 42:257–266, 1978.
45. P. Yin, R.F. Hariadi, S. Sahu, H.M. Choi, S.H. Park, T.H. Labean, and J.H. Reif. Programming DNA tube circumferences. *Science*, 321:824–826, Aug 2008.
46. Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.