IRIS: Intermolecular RNA Interaction Search

Dmitri D. Pervouchine[1]
dp@bu.edu

[1] Center for BioDynamics, Boston University, Boston MA 02215, USA

Abstract

Here we present IRIS, a method for prediction of RNA-RNA interactions that is based on dynamic programming and extends current RNA secondary structure prediction approaches. Using this method we have found a number of interesting refinements to the structures of RNA-RNA complexes that have been studied previously and predicted novel targets for several known regulatory RNAs in *E. coli.* The computational time and memory usage of IRIS are $O(n^3m^3)$ and $O(n^2m^2)$, respectively, where n and m are the lengths of the input sequences. IRIS can be used for analysis of antisense regulatory systems in sequenced organisms and for the design of artificial riboregulators such as antisense drugs.

Keywords: rioboregulator, riboswitch, micro RNA, small interfering RNA, secondary structure, dynamic programming, pseudoknot.

1 Introduction

In recent years, the class of non-coding RNAs has acquired many new members in prokaryotes, eukaryotes and archaea [1, 2, 3]. Non-coding RNAs are involved in catalysis and metabolite-sensing, but more typically they are employed by the cell to guide sequence-specific recognition and processing of other RNA molecules [4]. In particular, they can silence or repress genes at the posttranscriptional level by using base complementarity to hybridize with mRNAs. In animals and plants they are known as small interfering RNAs (siRNAs) and microRNAs (miRNAs), whereas in bacteria the commonly used term is riboregulator [5]. While siRNAs are often fully complementary to their targets, miRNAs and riboregulators interact with mRNAs in a more intricate manner, one which does not involve perfect duplexing [6]. Riboregulators in *E. coli* demonstrate that this interaction can be rather complex, such that intermolecular helices alternate with unpaired regions and intramolecular secondary structure. For instance, the small RNA *oxyS*, which is involved in the oxidative stress response system, interacts with its target, *fhlA* mRNA at two sites that reside in the loops of two stem-loop structures [7, 8]. The *dsrA* RNA, which activates and represses, respectively, two other transcriptional regulators hn-S and *rpoS*, pairs to their mRNAs by different parts of its three stem-loops [9, 10]. The antisense RNA *copA* binds to the leader region of *copT* mRNA and forms an asymmetrical X-shaped conformation with a four-way junction [11]. The small RNA *ryhB* interferes with translation and transcription of the polycystronic gene *sdhCDAB* by hybridization with the ribosome binding site [12]. In *C. elegans*, the small temporal RNAs *lin-4* and *let-7* pair to the 3' untranslated regions of their target genes in multiple copies, forming bulged double-stranded structures [13, 14, 15].

The wealth of genomic information that was brought by high throughput sequencing poses a challenging problem to systematically search for new targets of known riboregulators and miRNAs. The first step in this direction is to develop an approach for predicting RNA-RNA interactions that extends beyond usage of standard sequence alignment tools. The most straightforward way to find out whether two RNA sequences are able to hybridize with each other is to search for a reverse complement of one of them in another using, for instance, nucleotide BLAST [16]. This method is applicable for locating long stretches of complementarity and is useful in, for instance, siRNA target search [17, 18]. Another approach, which seems to be more suitable for finding miRNA targets, consists in adopting the RNA structure prediction programs to treat two input sequences [19]. One could concatenate the two RNA sequences, input the result to MFOLD, and obtain the intra and intermolecular pairings from the MFOLD output [20]. However, this approach has a serious problem because MFOLD can only predict pseudoknot-free structures, that is, it will erroneously miss the optimal intermolecular interaction in favor of intramolecular secondary structure, if the individual molecules are highly structured [21]. The secondary structure prediction method that is tolerant to pseudoknots is potentially applicable but not practicable because of high computational complexity [22].

Here we introduce a method for Intermolecular RNA Interaction Search (IRIS) [23]. Essentially, it is a product of sequence alignment and two MFOLD-type secondary structure prediction algorithms, implemented as four-dimensional dynamic programming. It shares many common features with secondary structure prediction method with pseudoknots, but is less computationally intensive. The input consists of two RNA sequences. Each of the sequences is allowed to form its own nested secondary structure and to hybridize to the other molecule. The computational time and storage are $O(n^3m^3)$ and $O(n^2m^2)$, respectively, where n and m are the lengths of the input sequences. Although the total degree of the algorithm is six (as in [22]), it is more practical than the pseudoknotted algorithm when one of the sequences is much shorter than the other. This, indeed, is the case for riboregulators and miRNAs. The computational time needed to obtain the optimal secondary structure with pseudoknots by [22] would be $(n+m)^6$, whereas with IRIS it only takes n^3m^3. This facilitates more than 10^6-fold increase in speed for $n = 200$ and $m = 20$.

2 Methods

The principle of secondary structure prediction methods that are based on the dynamic programming consists of the recursive derivation of the secondary structure for all segments of the sequence, proceeding progressively from short segments to the entire molecule [24]. A fitness function (typically, the lowest equilibrium free energy) is optimized on each step of the recursion. The fitness function used in MFOLD is the sum of energy impacts from stacking interactions, dangling bases, hairpins, bulges, internal loops and multiloops [25]. It is clear that the set of thermodynamic parameters that is used for RNA secondary structure prediction can be also used for structure prediction of RNA-RNA complexes, as they are based on energies of the same structural elements. In order to provide clear explanation and simple of notation, here we use trivial fitness function, the maximum number of base pairs, although the actual algorithm is based on thermodynamic parameters [26].

First, we recall the formal language of secondary structures. A secondary structure on a sequence $(x_1, \dots, x_n)$ is a set of pairs (i, j) such that for every two pairs (i, j) and (i', j') the condition $i = i'$ implies $j = j'$ and vice versa. In other words, each position in can be paired to at most one other position. A secondary structure is said to be nested, if for every two pairs (i, j) and (i', j'), either $i \le i' \le j' \le j$ or $i' \le i \le j \le j'$. A nested secondary structure can be represented by an outer-planar graph, that is, a graph that can be embedded in the plane such that all vertices lie on the boundary of its exterior region [27]. Nested secondary structures admit an equivalent representation by regular expressions. In this notation, balanced parentheses are used to indicate paired positions, while dots are used to indicate ones that are unpaired. There is a one-to-one correspondence between nested secondary structures and their parenthesis representations [27]. In order to set up the recursion, Nussinov denotes by M_{ij} the maximum number of pairs in the segment $(x_i, \dot{;} x_j)$. Then, M_{ij} can be calculated by formula

$$M_{i\,j} = \max\left\{ M_{i+1\,j-1} + \delta_{ij},\ \max_{i<k<j}\{M_{i\,k} + M_{k+1\,j}\} \right\} \tag{1}$$

with the initial conditions $M_{i\,i} = M_{i\,i+1} = 0$, $i = 1 \dots n$. Here $\delta_{ij} = 2$, if x_i and x_j are complementary nucleotides, and $\delta_{ij} = 0$ otherwise.

Now we extend this formalism to the case of two sequences. Denote them by $X = (x^1, \dots, x^n)$ and $Y = (y_1, \dots, y_m)$. In what follows, we assume that X is written in 5'-3' direction, Y is written in 3'-5' direction, the superscripts always refer to X, and the subscripts always refer to Y. Consider the following sets of pairings: (1) pairings of x^i with x^j, (2) pairings of y_k with y_l, and (3) cross-pairings of x^i with y_k. We require that

1. Every position in X and Y participates in at most one pairing.
2. The pairings of x^i with x^j form nested secondary structure.
3. The pairings of y_k with y_l form nested secondary structure.
4. If x^i is paired to y_k, and x^j is paired to y_l, then $i < j$ implies $k < l$ and vice versa.

If the conditions 1-3 are met then the set of pairings is said to be the *joint secondary structure*. If, in addition, the condition 4 is met then the joint secondary structure is said to contain *no generalized pseudoknots*. Geometrically, the conditions 1-4 mean that the first and the second set of pairings can be represented by two outer-planar graphs, whose nodes are connected by the third set of pairings such that the connecting edges don't intersect. Absence or presence of generalized pseudoknots is not critical for the algorithm's performance. However, the joint secondary structure admits an analog of parenthesis representation when generalized pseudoknots are absent. This representation may contain gaps, and, therefore, one structure can have more than one parenthesis representation, but the inverse relation in unambiguous: one representation corresponds to only one joint secondary structure.

From now on we consider the structures that are free of generalized pseudoknots. A pair of segments $(x^i, \dots, x^{j-1})$ and $(y_k, \dots, y_{l-1})$ is called the *bisegment* and is denoted by XY^{ij}_{kl}. Note that the last positions x^j and y_l and are not included. The way of indexing has been changed because we now have to deal with segments of zero length. Let M^{ij}_{kl} be the maximum number of base pairs in XY^{ij}_{kl}. Similarly to (1), we can calculate M^{ij}_{kl} recursively by formula

$$M^{ij}_{kl} = \max\left\{M^{i+1j-1}_{kl}+\delta^{ij-1}, M^{ij}_{k+1l-1}+\delta_{kl-1}, M^{ij-1}_{kl-1}+\delta^{j-1}_{l-1}, M^{i+1j}_{k+1l}+\delta^{i}_{l}, \max_{s,t}\{M^{is}_{kt}+M^{s+1j}_{t+1l}\}\right\} \tag{2}$$

In other words, either two terminal bases of the first sequence, or two terminal bases of the second sequence, or a terminal base of the first and a terminal base of the second sequences pair, or the bisegment XY^{ij}_{kl} is spited into two previously processed bisegments. The last term in equation (2), $M^{is}_{kt}+M^{s+1j}_{t+1l}$ does not account for generalized pseudoknots. In order to treat generalized pseudoknots, one should replace it with $\max\{M^{is}_{kt}+M^{s+1j}_{t+1l}, M^{is}_{tl}+M^{s+1j}_{k\,t+1}\}$.

The recursion (2) stops when $i = j$ or $k = l$, that is, when one of the sequences in the bisegment becomes empty. Therefore, the initial conditions for (2) are $M^{ij}_{kk} = M^{i\,j-1}$ and $M^{ii}_{kl} = M_{k,l-1}$ for all i, j, k, and l, where M^{ij} and M_{kl} are the corresponding Nussinov matrices for X and Y, respectively. They are calculated recursively by equation (1):

$$M^{i\,j} = \max\left\{M^{i+1\,j-1} + \delta^{ij}, \max_{i<s<j}\{M^{i\,s} + M^{s+1\,j}\}\right\}, \tag{3}$$

$$M_{k\,l} = \max\left\{M^{k+1\,l-1} + \delta^{kl}, \max_{k<t<l}\{M^{k\,t} + M^{t+1\,l}\}\right\}. \tag{4}$$

with initial conditions $M_{kk} = M_{k\,k+1} = M^{ii} = M^{i\,i+1} = 0$ for all i and k. As always, the superscripts refer to the sequence X, and the subscripts refer to the sequence Y.

Calculation of M^{ij}_{kl} is organized as follows. First, we initialize the matrices M^{ij} and M_{kl}, and then compute M^{ij} and M_{kl} by equations (3) and (4). Next, we initialize the four-dimensional matrix M^{ij}_{kl} using M^{ij} and M_{kl}, and then compute M^{ij}_{kl} by equation (2). The number $M^{1\,n+1}_{1\,m+1}$ is the desired

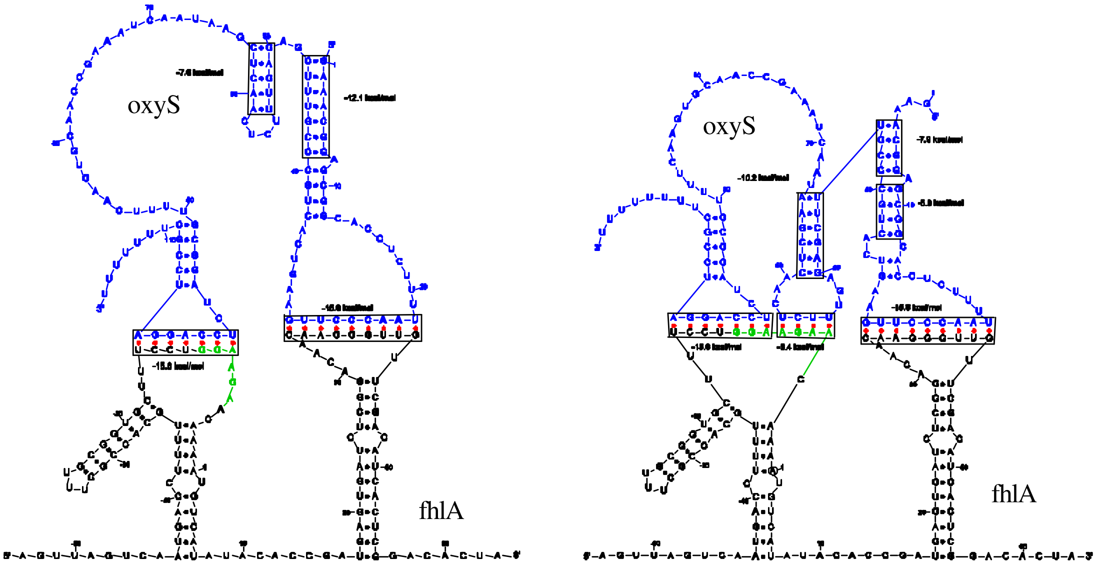

Figure 1: The structure of *oxyS*(blue)-*fhlA*(black) complex proposed by Argaman *et al* [8] (left) and the structure predicted by IRIS (right). The Shine-Dalgarno sequence is shown in green.

maximum number of base pairs, and the optimal joint secondary structure is obtained from the matrix M^{ij}_{kl} by traceback. In this setup, the computational time and space are $O(n^3m^3)$ and $O(n^2m^2)$, respectively.

The algorithm described in this section can be modified and extended to a more realistic schema, one which is based on thermodynamic parameters rather than on scoring matrix. The reader is referred to the supplementary material for the desciption of the complete algorithm.

3 Results

In this section we IRIS to several RNA-RNA complexes that have been described in the literature. The annotated genomic sequences of *E. coli* K12 (NC_000913), *S. flexneri* 2457T (NC_004741), *S. typhi* Ty2 (NC_004631), and *S. typhimurium* LT-2 (NC_003197) were obtained from NCBI. The calculations were performed with a temperature parameter setting of 37C for all sequences. Comparative sequence analysis was performed using CLUSTALW [28] (alignments not shown). The following pairs of regulatory RNA vs. target mRNA were analysed: *oxyS* and *fhlA*, *dsrA* and *rpoS/hns* (results not shown), *gcvB* and *dppA/oppA*, *dicF* and *ftsZ/ftsA*, and *ryhB* and *sdhC/bfr/sodB* (results not shown). The predicted structures have been translated from the parenthesis notation to a more friendly, pictorial representation (figures 1-2) using JAVA-based software [29].

4 Discussion

oxyS and fhlA. The *oxyS* RNA is expressed in *E. coli* in response to oxidative stress and is known to repress the translation of *fhlA* gene by blocking ribosome binding. It was found in [7] that *oxyS* operates by pairing with a short sequence overlapping the Shine-Dalgarno sequence. Later on, deletion and mutation studies performed by the same group revealed that *oxyS-fhlA* interaction involves

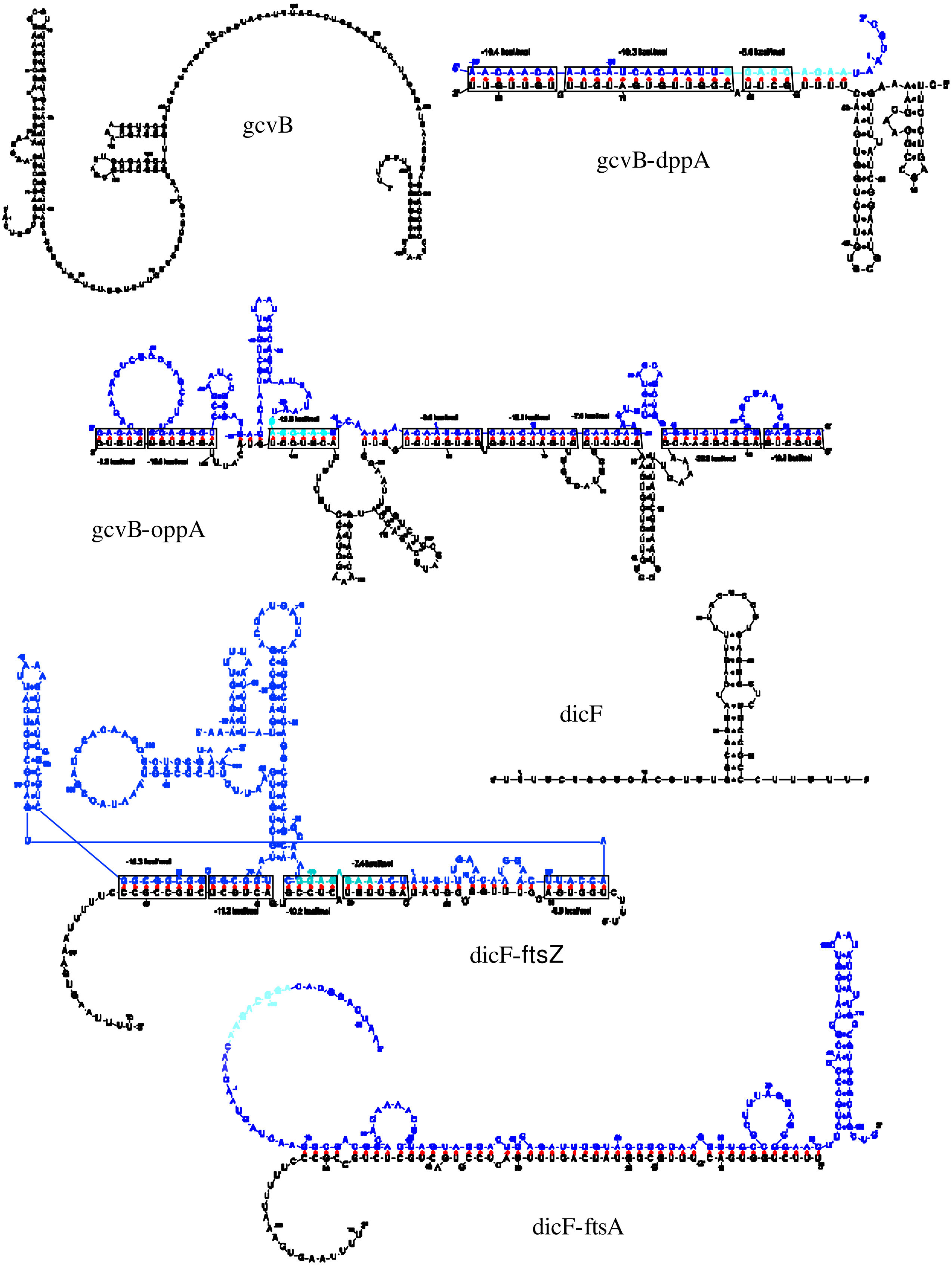

Figure 2: The predicted structures for gcvB RNA, *gcvB*(black)-*dppA*(blue), *gcvB*(black)-*oppA*(blue) complexes, *dicF* RNA, *dicF*(black)-*ftsZ*(blue), and *dicF*(black)-*ftsA*(blue) complexes. The ribosome binding site is shown in light blue.

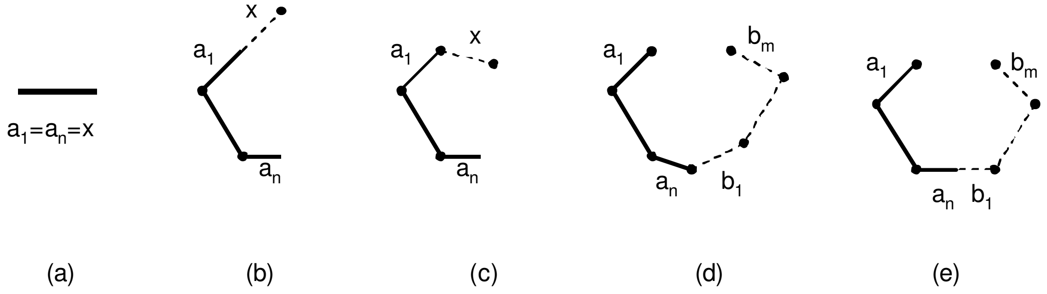

Figure 3: Transformations of polygons (see text).

a second site residing further downstream, within the coding region of fhlA [8]. The structure of oxyS-fhlA complex that was proposed in [8] consists of four adjacent stem-loops, two in each of the interacting molecules, which form stable kissing complex (figure 1a). We examine this complex and find that, in fact, it is not the minimum free energy structure. Moreover, the Shine-Dalgarno sequence (shown in green) is only partially obstructed. The optimal structure predicted by IRIS is shown in figure 1b and has slightly different arrangement of hairpins. According to IRIS prediction, there exist the third site of interaction that is located between two other sites, which is able to sequester the Shine-Dalgarno sequence completely. Comparative sequence analysis in other enteric bacteria shows that this structure is conserved in *S. flexneri*, *S. typhi*, and *S. typhimurium.*

gcvB and dppA/oppA. Transcription of *gcvB* RNA in *E. coli* is controlled by transcriptional regulators of the gcvTHP operon encoding the enzymes of the glycine cleavage system [30]. It has been shown to repress the translation of OppA and DppA genes, the periplasmic-binding protein components of the two major peptide transport systems. We propose the following structures that could be responsible for inhibition of translation of OppA and DppA by gcvB, although the detailed mechanism involving gcvB in the repression of these two genes has not been studied yet (figure 2). The structure of the gcvB RNA (200-bp long) contains three distinguished stem-loops, two of which are folowed by polyuridine track ($n = 5 - 7$) and, therefore, may act as terminators. The structure of the *gcvB-oppA* complex involves intermolecular helices that precede nd follow the putative terminator, while the interaction between *gcvB* and *dppA* mRNA only contains helices that precede the putative terminator. This difference might impose another level of control or differential influence on regulation of these two genes. The Shine-Dalgarno sequence in *gcvB-oppA* complex is mostly obstructed, while the most of the structure in *gcvB-oppA* complex is in the upstream region. This correlates very well with the fact that *oppA* regulation appears to be at the translational level, whereas *dppA* regulation occurs at the mRNA level [30]. The structure is conserved in *S. flexneri*, *S. typhi*, and *S. typhimurium.*

dicF and ftsZ. A 190 nt RNA *dicF* is processed from a polycistronic transcript (*dicB* operon) by RNase III and RNase E [31]. A cell-division gene, *ftsZ*, has been identified as the target of *dicF* RNA by a genetic screen for suppressors of *dicF*-dependent inhibition of cell division [32]. We confirm that *dicF* RNA has significant complementarity to the *ftsZ* mRNA in the region surrounding the ShineDalgarno sequence, which is consistent with the result that *dicF* regulates *ftsZ* by interfering with ribosome binding. We also find that *dicF-ftsZ* complex admits another region of complementarity (figure 2), which gives rise to a generalized pseudoknot. The structure of the *dicF-ftsZ* complex is similar to the one of *dsrA-hns*, in which *dsrA* RNA interacts with both 5' and 3' ends of the *hns* mRNA [9]. We also report that *ftsA*, another gene that is involved in cell division, has significant complementarity to *dicF* RNA (figure 2). Interestingly, the *dicF-ftsA* interaction appear to be downstream of the start codon. As in the case of *gcvB-dppA*, it suggests that *dicF*-mediated regulation of *ftsA* occurs at the mRNA level, while the regulation of *ftsA* occurs at the translational level.

The structures of *oxyS-fhlA* complex shown in figure 1, have an interesting property. The problem is that the rightmost intermolecular helix is surrounded by AA and UU nucleotides, which could have been paired. However, the helix has to be "straight" because it makes a full turn over the distance of 10 base pairs. Thus, the distance between two nucleotides at the ends of the helix is approximately 3.4 nm [33]. On the other hand, the phosophodiester bond is approximately 0.7 nm long. Therefore, the bases A and U that are adjacent to the helix cannot reach each to form a base pair because of backbone constrains. In the case of one RNA molecule we didn't have this problem because on each step of the recursion the paired nucleotides were independent of previously built helices. Now we need to keep track of the geometry of the growing structure, as the multiloop energy is not a simple function of the number of interior bases. The same sort of artifacts is peculiar to the RNA structure prediction algorithm with pseudoknots [22].

Geometrically, this problem comes down to arranging a number of straight segments in 3D space and connecting them by freely-joint chains such that obvious spatial constraints are met. The straight segments and the freely-joint chains are helices and unpaired fragments, respectively. In particular, the triangle inequality must hold, that is, for any closed chain of segments in the structure, the length of each segment is not greater than the sum of lengths of the other segments. Although the triangle inequality filters out the configurations that *a priori* contradict Euclidian geometry, it is necessary but not sufficient condition for the given arrangement of helices to exist. In general, this problem can be as hard as the problem of graph embeddings [34].

We use the triangle inequality to detect structures that violate backbone constraints as follows. For each bisegment we consider four chains of straight segments that correspond to its four faces: top, bottom, left, and right. They account for four possibilities of pairing one of the two 5' ends with one of the two 3' ends. Here by $a_1, \ldots, a_n$ we denote the lengths of the segments, including both helices and single-stranded regions. Then, the ends of the chain can touch each other if and only if all of the

$$\lambda_j = a_j - \sum_{i \neq j} a_i \tag{5}$$

are negative, or, equivalently, $\lambda = \max\{\lambda_j\}$ is negative. We call λ_j the *discrepancy of the j-th segment.* Denote by π the perimeter of the chain $a_1, \ldots, a_n$. Heuristically, λ and π are the minimum and the maximum distance between the ends of the freely-joint chain. On each step of the dynamic programming we either make a base pair at one of the ends of the bisegment or split it into smaller parts. The criterion $\lambda < 0$ can be used to determine whether a base pair is possible. However, calculation and evaluation of λ requires checking all the subsegments of the given segment and, therefore, results in non-polynomial computational time.

However, it is not necessary to keep track of λ_j for each segment, if we only extend but not change the structures that have been built before. These extensions are summarized in figure 3. They start with a dinucleotide (a). A new base pair can elongate the existing straight segment (b) or be connected to it freely-jointly (c). Similarly, when the bisegment is split into parts, the corresponding polygons are concatenated. The concatenation can be soft (d), if the terminal segments are connected freely-jointly, or rigid (e), if they merge into a new straight segment. It turns out that we only need λ, λ_1, λ_n, and π in order to determine whether the chain can close to form a polygon. These parameters are recalculated on every step of the dynamic programming for each of the four chains according to table 1. By α and β in table 1 we denote the discrepancies of the initial and the terminal segments, that is λ_1 and λ_n, respectively. The subscripts refer to the first and the second polygon, which are being connected, concatenated, or closed.

Although the triangle inequality doesn't give a sufficient condition for structure to exist, it allows to get rid of simple artifacts. In principle, the theory of graph embeddings can be used here to obtain not only necessary but also the sufficient conditions, but it would give rise to an NP-complete algorithm.

Table 1: Transformations of polygons (see text).

Transformation		α	β	λ	π
Initiation		x	x	x	x
Elongation	left	$\alpha + x$	$\beta - x$	$\max\{\lambda - x, \alpha + x\}$	$\pi + x$
	right	$\alpha - x$	$\beta + x$	$\max\{\lambda - x, \beta + x\}$	$\pi + x$
Connection	left	$x - \pi$	$\beta - x$	$\max\{\lambda - x, x - \pi\}$	$\pi + x$
	right	$\alpha - x$	$x - \pi$	$\max\{\lambda - x, x - \pi\}$	$\pi + x$
Concatenation	soft	$\alpha_1 - \pi_2$	$\beta_2 - \pi_1$	$\max\{\lambda_1 - \pi_2, \lambda_2 - \pi_1\}$	$\pi_1 + \pi_2$
	rigid	$\alpha_1 - \pi_2$	$\beta_2 - \pi_1$	$\max\{\beta_1 + \alpha_2, \lambda_1 - \pi_2, \lambda_2 - \pi_1\}$	$\pi_1 + \pi_2$
Closing	soft/soft	$\max\{\lambda_1 - \pi_2, \lambda_2 - \pi_1\}$			
	soft/rigid	$\max\{\beta_1 + \alpha_2, \lambda_1 - \pi_2, \lambda_2 - \pi_1\}$			
	rigid/rigid	$\max\{\alpha_1 + \beta_2, \beta_1 + \alpha_2, \lambda_1 - \pi_2, \lambda_2 - \pi_1\}$			

5 Conclusion

Riboregulators and micro RNAs represent one of those few cases in molecular biology where functional relationship between genes can indeed be established from sequence data. Prediction of targets of regulatory non-coding RNAs is a logical and pertinent task at the current state of art. The method developed in this work provides a generic framework for this problem. It has been shown to agree with several known examples of RNA regulation, yielded a number of interesting refinements to their structures, and allowed to predict novel targets. Although the inherent computational complexity precludes applications of this method on genome-wide scale, it still can be used for the analysis of antisense regulatory systems in sequenced organisms and for the design of artificial riboregulators such as antisense drugs.

6 Acknoledgements

I thank James Collins, Farren Isaacs, Daniel Dwyer for their support and inspiring discussions, and Maxim Frank-Kamenetskii for his valuable remarks on the thermodynamic model. This work was supported by NIH grant 1 R01 NS46058.

References

[1] A. Vitreschak, D. Rodionov, A. Mironov, and M. Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *TRENDS in Genetics*, 20(1):44–50, 2004.

[2] N. Sudarsan, J. Barrick, and R. Breaker. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, 9:644–647, 2003.

[3] C. Gaspin, J. Cavaillé, G. Erauso, and J.-P. Bachellerie. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. *J. Mol. Biol.*, 297:895–906, 2000.

[4] T. Tuschl. Functional genomics: RNA sets the standard. *Nature*, 421:220–221, 2003.

[5] E. Massé, N. Majdalani, and S. Gottesman. Regulatory roles for small RNAs in bacteria. *Curr Opin Microbiol.*, 6(2):120–4, 2003.

[6] S. Altuvia, E. Gerhart, and H. Wagner. Switching on and off with RNA. *Proc. Natl. Acad. Sci. USA*, 97(18):9824–9826, 2000.

[7] S. Altuvia, A. Zhang, L. Argaman, A. Tiwari, and G. Storz. The Escherichia coli OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *The EMBO Journal*, 17(20):6069–6075, 1998.

[8] 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(5):1101–12, 2000.

[9] R. A. Lease and M. Belfort. A trans-acting RNA as a control switch in Escherichia coli: DsrA modulates function by forming alternative structures. *Proc. Natl. Acad. Sci. USA*, 97(18):9919–9924, 2000.

[10] F. Repoila, N. Majdalani, and S. Gottesman. Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Molecular Microbiology*, 48(4):855, 2003.

[11] F. A. Kolb, C. Malmgren, E. Westhof, C. Ehresmann, B. Ehresmann, E. G. H. 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, 2000.

[12] E. Masse and S. Gottesman. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl. Acad. Sci. USA*, 99(7):4620–4625, 2002.

[13] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294:853–857, 2001.

[14] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science*, 294:858–862, 2001.

[15] R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294:862–864, 2001.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[17] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel. Prediction of plant microRNA targets. *Cell*, 110(4):513–20, 2002.

[18] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen. Identification of Drosophila MicroRNA Targets. *PLoS Biol.*, 3:E60, 2003.

[19] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome Biol.*, 5(1):R1. Epub, 2003.

[20] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker. Coaxial Stacking of Helixes Enhances Binding of Oligoribonucleotides and Improves Predictions of RNA Folding. *Proceedings of National Academy of Sciences*, 91:91, 1994.

[21] D. Pervouchine, J. Graber, and S. Kasif. On the normalization of RNA equilibrium free energy to the length of the sequence. *Nuc. Ac. Res.*, 31(9):e49, 2003.

[22] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[23] http://math.bu.edu/people/dp/prj/iris.

[24] M. Zucker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:46, 1984.

[25] D. Mathews, J. Sabina, M. Zucker, and H. Turner. Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure. *Journal of Molecular Biology*, 288:911–940, 1999.

[26] R. Nussinov, G. Pieczenik, J. Griggs, and D. J. Kleitman. Algorithms for Loop Matching. *Journal of Applied Mathematics*, 35(1):68–82, 1978.

[27] J. Leydold and P. F. Stadler. Minimal Cycle Bases of Outerplanar Graphs. *Elec. J. Comb.*, 5:209–222, 1998.

[28] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[29] D. Pervouchine. Drawing RNA structures with pseudoknots (currently in preparation), http://math.bu.edu/people/dp/prj/rnadraw.

[30] M. L. Urbanowski, L. T. Stauffer, and G. V. Stauffer. The gcvB gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in Escherichia coli. *Molecular Microbiology*, 37(4):856, 2000.

[31] M. Faubladier, K. Cam, and J. P. Bouche. Escherichia coli cell division inhibitor DicF-RNA of the dicB operon. Evidence for its generation in vivo by transcription termination and by RNase III and RNase E-dependent processing. *J Mol Biol*, 212(3):461–71, 1990.

[32] F. Tetart and J. P. Bouche. Regulation of the expression of the cell-cycle gene ftsZ by DicF antisense RNA. Division does not require a fixed number of FtsZ molecules. *Mol Microbiol.*, 6((5):615–20, 1992.

[33] J. D. Watson and F. H. C. Crick. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737, 1953.

[34] R. J. Wilson. *Introduction to Graph Theory.* London: Longman, 1975.