ALBERT-LUDWIGS-UNIVERSITY
FREIBURG
DEPARTMENT OF COMPUTER SCIENCE

Bioinformatics Group
Prof. Dr. Rolf Backofen

RNA Consensus Interaction Prediction

Master Thesis

Cameron Smith

June 2010 - January 2011

Declaration

I hereby declare, that I am the sole author and composer of my Thesis and that no other sources or learning aids, other than those listed, have been used.

Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.

I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

_____ _____

Date Signature

Acknowledgements

I would like to thank Prof. Dr. Rolf Backofen for the opportunity to undertake this thesis in the University of Freiburg Bioinformatics Department. Prof. Backofen has engendered a friendly and supportive atmosphere in the group which has made the completion of my thesis within the Bioinformatics Department an enjoyable and positive experience. As such, I would also like to thank the members of the group for their encouragement, support, good cake and powerful coffee. In particular, my supervisor Andreas Richter has surpassed himself in the performance of his task. Vielen Dank Andreas for your wise guidance, endless patience and friendly disposition. I'd like to thank my fellow students Phillip Chlap and Hadi Kheyruri, with whom I've learnt, suffered, rejoiced and eaten more meals at the mensa than I'd care to remember, for making the last three years especially memorable. Finally, I'd like to thank my family for their selfless support that has made this possible.

Summary

RNA-RNA interaction is a subject of considerable biological relevance as the binding of ncRNA to mRNA can affect both the transcription and translation of the bound mRNA and hence regulate gene expression. The accuracy and reliability of single sequence RNA structure prediction has been shown to increase significantly when the structure of an aligned set of RNA homologs is computed. As such, it is posited that by augmenting an existing RNA-RNA interaction prediction algorithm, that determines an interaction structure based only on thermodynamics, with a phylogenetic component a structure prediction of improved quality can be obtained. This thesis presents the theory, implementation and evaluation of an algorithm that combines thermodynamic and phylogenetic information to predict a consensus interaction structure on a set of aligned mRNAs and ncRNAs.

Zusammenfassung

Interaktionen zwischen zwei RNA-Molekülen sind von großer biologischer Bedeutung. Beispielsweise kann die Bindung einer nicht-kodierenden RNA (ncRNA) an eine mRNA sowohl Transkription als auch Translation der gebundenen mRNA beeinflussen und damit die Expression des kodierten Gens regulieren. Die Genauigkeit und Verlässlichkeit von RNA-Struktur-Vorhersage auf einzelnen Sequenzen wird deutlich verbessert, wenn die Struktur nicht für eine Einzelsequenz, sondern für ein Alignment von Homologen der entsprechenden RNA berechnet wird. Von daher wird postuliert, dass durch die Erweiterung eines bestehenden RNA-RNA-Interaktionsvorhersage-Algorithmus um eine phylogenetische Komponente die Vorhersagequalität verbessert werden kann im Vergleich zur Interaktionsvorhersage nur anhand von Thermodynamik. Diese Arbeit präsentiert die Theorie, Implementierung und Evaluation von einem Algorithmus, der thermodynamische und phylogenetische Information kombiniert, um eine Konsensus-Interaktions-Struktur auf einer Menge von alignierten mRNA- und ncRNA-Sequenzen vorherzusagen.

Contents

1. Introduction

The focus of this thesis is the prediction of inter-molecular binding of RNA strands. Specifically, between a(n) non-coding RNA (ncRNA) and messenger a(n) RNA (mRNA). This problem is commonly termed **R**NA-RNA **I**nteraction **P**rediction and hence is also known as the **RIP** problem. The RIP problem deals specifically with complementary RNA-RNA interactions, i.e. base pair formation of canonical (Watson-Crick) and G-U (wobble) base pairs. RNA-RNA interaction is of considerable biological relevance as the binding of ncRNA to mRNA can affect both the transcription and translation of the bound mRNA and hence regulate gene expression. Consequently the RIP problem for single pairs of interacting RNA sequences has been extensively studied and a wide range of methods and tools exist for working with this problem.
The goal of this thesis is to improve the prediction of RNA-RNA interaction by adapting a technique from RNA secondary structure prediction. Predicting the structure of an RNA sequence was originally computed by maximizing base pairs [33], minimizing free energy [52] or predicting base pair probabilities [28] of a *single* sequence. The accuracy and reliability of RNA structure prediction has been shown to increase significantly when the structure of an aligned set of RNA homologs is computed [19]. This *consensus* structure incorporates both a thermodynamic contribution, usually free energy minimization or partition function based, and an evolutionary component such that only base pairs that a majority of the aligned sequences allow will be contained in the consensus structure.
As interaction prediction is essentially the prediction of structure formed by two molecules, it follows that a shift to a consensus model should improve the quality of inter-molecular interactions predicted. As little work has, up to now, gone into Consensus Interaction Prediction this thesis aims to explicate this topic and develop an algorithm for the fast, accurate and reliable prediction of RNA-RNA interaction.

1.1. Motivation

The RNA world theory, popularized by Walter Gilbert in [14], promotes the possibility that the functional arrangement of DNA, RNA and protein that comprises all living organisms today, evolved from an RNA only system. Although RNA is now only one part in the system of molecules that support life, it plays a role of fundamental importance. RNA was earlier thought to only be an intermediate stage in the synthesis of proteins from information stored in DNA, its functional role in the ribosome was seen as an exception. Recently a large number of RNAs that are not, or may not be, translated into protein have been discovered and the importance of the functional role RNA plays within the cell has been highlighted.

It has been shown that eukaryotic miRNAs and siRNAs and bacterial sRNAs post transcriptionally regulate the expression of target genes by binding to mRNAs [37] [3] [17] [49]

[15]. Non-coding RNAs in bacteria (sRNAs) have also been shown to activate genes by a variety of mechanisms [12]. Meyer [29] provides the following list of functions non-coding (or functional) RNAs can perform, RNA: cleavage, editing, modification, splicing, translation, suppression of translation and degradation. Meyer also states that pre-messenger RNAs, through alternative-splicing, can play a functional role meaning a single RNA can act both functionally and non-functionally. Interactions of microRNA with the 3'-UTR or coding sequence (CDS) of targets mostly causes repression of the encoded gene through a variety of mechanisms such as translational control, induced mRNA cleavage and deadenylation [12] [48].
Small non-coding RNAs have predominantly been shown to repress bacterial mRNAs by masking the Shine-Dalgarno (SD) or AUG start codon sequence, thereby preventing 30S ribisome entry and, consequently, translation initiation [7] [12] [47]. The sRNA-mRNA duplex is then frequently subject to degradation by RNase E [47]. As an example, the ncRNA MicA is shown to be an antisense regulator of *ompA* and blocks ribosome binding at the translation start site and facilitates RNase E cleavage which leads to mRNA decay [43].
Animal miRNAs target transcripts through imperfect base-pairing to multiple sites in 3' untranslated regions, and Watson-Crick base pairing to the 5' end of miRNAs, especially to the seed region, is crucial for targeting [8]. MiRNAs use base-pairing to guide RNA-induced silencing complexes (RISCs) to repress targeted messages through mechanisms such as translational inhibition, accelerated exonucleolytic mRNA decay or site-specific endonucleolytic cleavage [8].
According to Waters and Storz [47], the suitability of RNA as a regulatory medium for gene expression can be attributed to the fact that RNA regulators are less costly to the cell and can be faster to produce, they also do not require the extra step of translation. Also when input signals are large and persistent, sRNAs are hypothesized to be better than transcription factors at strongly and reliably repressing protein levels, as well as filtering noise [47].

Translation of the *Salmonella ompN* mRNA is repressed by base pairing of the RybB sRNA. Although this base pairing does not sequester the SD or AUG sequences, it sufficiently disrupts formation of the ternary complex to inhibit translation [7]. *Staphylococcus aureus*, the Gram-positive pathogen most commonly responsible for staph infections, yielded the first sRNA discovered to activate gene expression [12] [32]. Investigation has suggested that an upstream anti-Shine-Dalgarno sequence folds back to pair with the SD region of the *hla* mRNA, and that by binding to the anti-SD and competing with the formation of a hairpin structure the 514nt RNAIII releases the SD of *hla* and promotes its translation [12]. $\mathrm{DsrA_{Bb}}$ acts through the same anti-antisense mechanism to activate translation of the *rpoS* mRNA in *Borrelia burgdorferi*, an agent of Lyme disease and non-Hodgkin lymphomas [12].

An increasing number of bacterial sRNAs have recently been shown to repress multiple if not large sets of mRNAs [7]. The fact that a base pairing sRNA often regulates multiple targets means that a single sRNA can globally modulate a particular physiological response [47]. MiRNAs are thought to regulate a large part of the protein-coding transcriptome and play a vital role in development, stress adaption and hormone signalling [8]. Furthermore, non-coding bacterial RNAs have been implicated in the regulation of stress responses and virulence traits [44]. Examples cited by [43] [47], of functionally characterized sRNAs of *Escherichia coli* involved in stress response and and adaptive change are listed in table 1.1.

Oxidative stress	OxyS
SOS response	IstR
Cold shock	DsrA
Low iron	RhyB
Osmotic stress	MicF
Outer membrane stress	MicA, RybB
Elevated glycine	GcvB
Glucose concentration change	Spot42, CyaR
Elevated glucose-phosphate levels	SgrS

Table 1.1.: Stress response in bacteria

Functional features tend to be conserved in evolution, and in the context of RNA-RNA interactions this means nucleotides whose evolution along a phylogenetic tree is coupled so that the functional features of the RNA-RNA interaction are conserved [29]. This motivates the effort to combine thermodynamic and evolutionary information in the prediction of RNA-RNA structure. Comparative methods can distinguish, given input alignments of sufficient quality, spurious base-pairs from evolutionarily conserved bp's [29]. In the context of RNA-RNA interactions, a comparative method may prefer a highly conserved inter-molecular bond over a more thermodynamically stable intra-molecular bond.
An example taken from [43] of the relevance of evolutionary information to structure prediction follows. MicA sequences differ substantially between bacteria. However, base changes are often located in single stranded regions/loops, or occur as compensatory changes when in stem regions. In two bacterial species *Erwinia carotovara* and *Klebsiella pneumoniae*, compensatory changes in both MicA and the *ompA* target maintain base-pairing. Chemical and enzymatic probing results were essentially consistent with the conformation predicted by MFold [51] and additional support was obtained from comparative analyses of the MicA-homologous sequences in *Klebsiella, Shigella, Salmonella, Yersinia, Enterobacter* and *Serratia* species [43].

The role of non-coding RNAs in development, virulence traits, physiological response to stress and hormone signalling highlights the importance of understanding how RNA-RNA interactions function. The prevalence of regulatory RNAs and the variety of mechanisms through which they act require improved prediction techniques. The fact that functional features tend to be conserved in evolution provides strong motivation for the incorporation of evolutionary information into the model for predicting RNA-RNA interactions.

1.2. Related Work

The work related to this thesis falls into three categories. RNA secondary structure prediction based on a multiple alignment of RNA sequences, RNA-RNA interaction prediction and the combination of these two approaches.

RNA secondary structure prediction (folding) using the Zuker algorithm [52] provides the commonly used method of predicting secondary structure from a sequence. A structure is

computed by finding the unique decomposition of a sequence, into secondary structure elements, that minimizes the free energy of the system. Such secondary structure elements, as described in the following chapter, have experimentally determined energies [42], [25], [26]. Two extensions of this basic method of structure prediction are important for this thesis. Consensus structure prediction, i.e. the simultaneous structure prediction of a set of aligned RNA sequences, provides a way to combine thermodynamic and phylogenetic information to predict a more accurate and reliable secondary structure. The specific approach to consensus structure prediction that will be utilized in this thesis has been developed and implemented in RNAalifold [5].
The second important extension to secondary structure prediction is the partition function approach developed by McCaskill [28]. This work provides a method to access the base pair and structure probabilities of an RNA sequence. Specifically it provides base-pairing probabilities over the ensemble of structures an RNA sequence may form. A useful adaption of the partition function to the task of predicting RNA-RNA interaction is found in the work done on RNAup [30].

RNA-RNA interaction prediction forms the second category of related work. One approach to predicting the interaction structure of an mRNA and ncRNA is to concatenate the sequences and apply a variant of Zuker and Stiegler's algorithm as in PairFold [1] and RNAcofold [4]. The limitations of this approach, such as the inability to predict important motifs such as kissing-hairpin loops [20] [29], suggest the alternative; treat the mRNa and ncRNA sequences separately and predict the interaction complex they form.
Three components conventionally comprise the quality of an RNA-RNA interaction, the central being the energy contribution of inter-molecular bonds also termed hybridization energy. An intra-molecular binding energy (accessibility) is used to more accurately model competing forces that form the RNA duplex. A region of perfect complementarity (seed region) is often observed in base pairing RNA interactions and can be applied as a constraint on prediction to increase the specificity of a computed interaction.
RNAhybrid is an extension of Zuker's algorithm that predicts multiple potential binding sites of ncRNAs (miRNA) in large target RNAs [37] using energy parameters from [25]. Bulge and internal loops are restricted to a constant maximum length (c), branching structures (multi-loops) and intra-molecular bonds are forbidden. RNAhybrid enforces the existence of a seed region and finds the minimal free energy structure with time and space requirements $O(c^2mn)$, $O(mn)$ respectively, where m and n are the lengths of input sequences.
RNAup implements an extension of the standard partition function that computes probabilities of unpaired sequence intervals [31]. As such, Mückstein et al. incorporate both the hybridization energy (energy gained from inter-molecular bonding) and interaction site accessibility (or energetic unfolding cost) of the individual sequences. RNAup's computation of the MFE structure requires $O(n^3)$ time and $O(n^2)$ space where n is the length of the mRNA.
IntaRNA integrates both the enforcement of a seed region and the accessibility of the individual sequences with the energy gained from hybridization to predict an MFE structure. The complete approach (restricting bulge and loop sizes) requires $O(m^2n^2)$ time and $O(mn)$ space, with m and n as the lengths of input sequences. A heuristic simplification of the complete approach provides space complexity $O(mn)$ and time complexity $O(m\overline{n})$ where $\overline{n} = \max n, L^3$ and L is the size of the sequence window in which both mRNA and ncRNA are folded. This thesis will extend IntaRNA to include a phylogenetic component in order

that prediction of consensus interaction is supported.

The third category of related work concerns methods that already include both thermodynamic and evolutionary components. Currently, the only known approach is implemented in PETcofold [40] which utilizes covariance information to hierarchically fold concatenated sequences to predict a joint secondary structure. PETcofold considers both hybridization and accessibility in computing an MFE structure and allows pseudknots between intra and inter-molecular base pairs. An extended version of PETfold [39] along with an hierarchical folding strategy form the basis of this approach.

1.3. Contribution

Presented in this work is an algorithm for predicting RNA-RNA interactions based on thermodynamic and evolutionary information. The algorithm implemented by IntaRNA [9] has been extended to predict a structure on a set of aligned mRNAs and ncRNAs. The Consensus Interaction Prediction has been implemented and evaluated against a dataset of experimentally validated structures, and compared against the quality of interactions predicted by IntaRNA.

1.4. Overview

The content and organization of the rest of this thesis are as follows. Chapter 2 provides the concepts and formal definitions that are necessary in the design of an algorithm for Consensus Interaction Prediction. Chapter 3 develops the prediction algorithm, formally defines its function and explains through example the detail of its working. Chapter 4 gives detail on the implementation of the algorithm and outlines how the algorithm is evaluated. The final chapter presents results and an analysis of the evaluation.

2. Fundamental Concepts and Definitions

The focus of this chapter is to develop the concept of computational RNA-RNA Interaction Prediction and formally define the components necessary to successfully implement an algorithm for Consensus Interaction Prediction.

2.1. Ribonucleic Acid (RNA)

RNA is a biologically fundamental macro-molecule present in eukaryotes, prokaryotes and archaea. RNA is composed of a series of nucleotides. It forms the bridge between information storage (in the form of DNA) and function as enacted by protein. RNA can be coarsely divided into two classes: Messenger RNA (mRNA) that as it's primary function acts as a template for protein synthesis and Non-coding RNA (ncRNA). From ribosomal RNA (rRNA) to long non-coding RNAs such as *Xist*, non-coding RNA plays an important **functional** role in living organisms although it is generally not translated into protein. As this thesis focuses on the prediction of interaction between an mRNA and ncRNA, the biological relevance is in how ncRNA can affect, and specifically regulate, cell function. Explain **trans**.

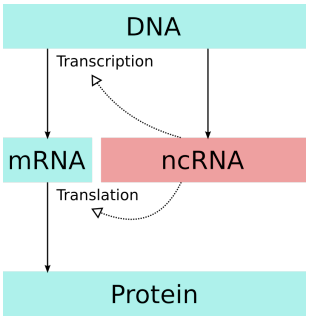

Figure 2.1.: mRNA and ncRNA

Primary structure

A sequence consisting of four nucleotides; Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). This can be formally defined as follows.

$$\text{Let } S \in \{A, C, G, U\}^n \text{ be an RNA sequence of length } n = |S|. \tag{2.1}$$

For access to individual nucleotides in S the subscript notation is used. The following example demonstrates.

$$S = ACGUCGUCGUACUGACGU, S_1 = A, S_4 = U, S_n = U$$

Secondary structure of a single stranded RNA

The secondary structure of an RNA molecule is primarily determined by hydrogen bonds formed between Watson-Crick nucleotide pairs and G-U wobble pairs. As such, the ability

of two nucleotides to pair is defined by the following.

Define $\mathcal{B} = \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}$ as the set of canonical base pairs. (2.2)

$$\pi(a,b) = \begin{cases} true & \text{if } (a,b) \in \mathcal{B} \\ false & \text{otherwise} \end{cases} \tag{2.3}$$

A secondary structure of an RNA sequence can be defined by a set of base-pairs.

$$\text{Let } P \subseteq \{(i,j)|1 \leq i < j \leq n, \pi(S_i, S_j)\} \tag{2.4}$$

The following constraints must be satisfied for a secondary structure to be valid.

Each nucleotide may pair with at most 1 other nucleotide. Such that,

$$\forall (i,j),(i',j') \in P : (i = i' \iff j = j') \wedge i \neq j. \tag{2.5}$$

A structure **may not** contain pseudoknots (crossing base pairs). Given $(i,j),(i',j') \in P, P$ is a non-nested (crossing) structure iff

$$i < i' < j < j' \vee i' < i < j' < j. \tag{2.6}$$

Secondary structure of an RNA hybridization

The following formally defines the structure of a hybridization between two RNA sequences. Intra-molecular base pairs are not yet considered, only the definition and constraints on inter-molecular structure are presented.

> Let M be an mRNA of length m and N be an ncRNA of length n.
>
> Let $P \subseteq \{(i,j)|1 \leq i \leq m, 1 \leq j \leq n, \pi(M_i, N_j)\}$ describe a duplex structure formed by M and N such that $(i,k),(j,l)$ are constrained by (2.5). The constraint on non-crossing base-pairs (2.6) needs to be adapted for a hybrid structure.

$$\forall (i,j),(k,l) \in P \Rightarrow i < k \Leftrightarrow j < l \tag{2.7}$$

Shown below is the simplest example of a hybridization structure violating the non-crossing constraint.

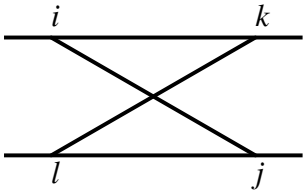

Figure 2.2.: Invalid hybridization structure

Secondary structure elements

RNA secondary structure is composed of sub-structures of type: hairpin, bulge, internal loop, stacking, multi-loop and dangling ends. An RNA-RNA interaction structure (duplex) consists of two components. The independent structures both mRNA and ncRNA form is used to account for intra-molecular base pairing. All types of sub-structure are admitted. The second component considered is the structure of the duplex formed by both sequences, which accounts for inter-molecular bonds. Stacking, internal loops, bulges and dangling ends sub-structure types are considered.

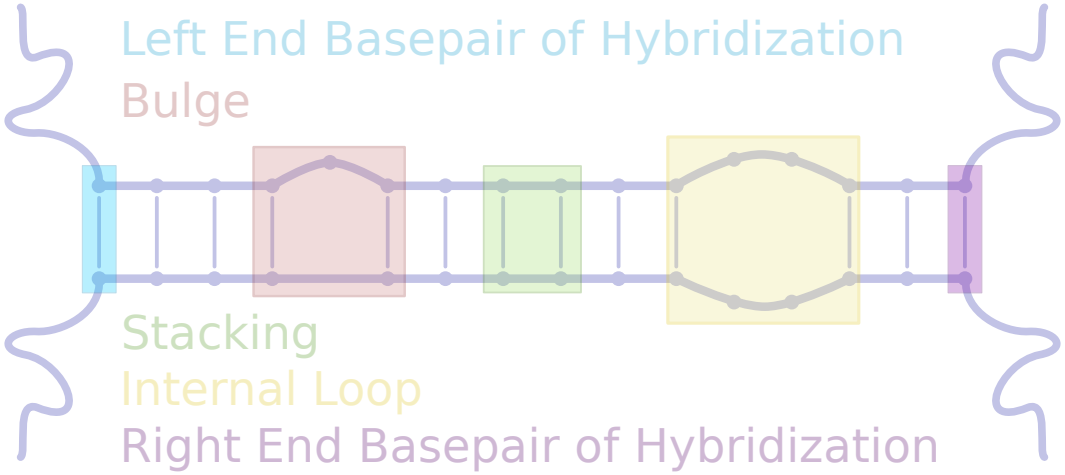

Figure 2.3.: Structural elements of an RNA-RNA interaction

The following formally defines the sub-structures comprising hybridization.

Stacking: consists of two *consecutive* base pairs. As such, $(i,j),(k,l) \in P$, where $1 \leq i < k \leq m$ and $1 \leq j < l \leq n$, comprise a stacking if and only if $j = i+1$ and $l = k+1$.

Bulge: is closed by two base pairs where *exactly one* set of involved nucleotides on the same molecule are consecutive. Formally $(i,j),(k,l) \in P$ form a bulge in the mRNA molecule iff $l = k+1$ and $j - i > 1$. In the case that $j = i+1$ and $l - k > 1$, $(i,j),(k,l)$ form an ncRNA bulge. All nucleotides between the bases forming the bulge must be *unpaired*, the non-crossing constraint (2.7) is sufficient to ensure this.

Internal Loop: is closed by two base pairs $(i,j),(k,l) : k - i > 1, l - j > 1$. Additionally, all nucleotides between i,k and j,l must be unpaired. Formally, $\neg\exists(p,q) \in P : i < p < k, j < q < l$ and the non-crossing constraint (2.7) suffice to describe a valid internal loop sub-structure.

Dangling Ends: are included at both ends of the hybridization region. They consist of the *outermost base pairs* formed by the mRNA and ncRNA. Additionally, 0-2 unpaired nucleotides outside the hybridization region can also contribute to each dangling end. Thus, base pair $(i,j) \in P : \neg\exists(i',j') \in P \wedge i' < i$ (which through the non-crossing constraint (2.7) ensures $j' > j$) denotes the left-end terminal base pair and for $i > 1, j > 1$ the dangling end sub-structure includes nucleotides $(i-1)$ and $(j-1)$.

Free energy minimization

Free energy minimization ([52], [42], [50]) is based upon a large number of measurements performed on small RNAs and the and the assumption that stacking base pairs and loop entropies contribute additively to the free energy of an RNA secondary structure [31] [25] [27] [26]. An RNA molecule is folded according to the combination of RNA secondary sub-structures that minimize the total free energy of the system.

2.2. Partition Function

Although the performance of MFE methods decreases for long structured regions of RNA, the performance for short (e.g. RNA-RNA interactions) can be increased by considering the entire statistical ensemble of configurations as described by the Boltzmann distribution [29].
Let $\mathcal{P}$ be the ensemble of structures that sequence S may form.
The equilibrium partition function is defined as follows.

$$Z_{\mathcal{P}} = \sum_{P \in \mathcal{P}} e^{-\frac{E(P)}{RT}}$$

And can be alternatively expressed.

$$E(\mathcal{P}) = -RT \ln(Z_{\mathcal{P}}).$$

If we take $\mathcal{P}_{i,k}$ as the set of all structures with $i \dots k$ unpaired, the accessibility (energy required to unfold) a region i, k is:

$$ED(i,k) = E(\mathcal{P}_{i,k}) - E(\mathcal{P}).$$

2.3. RNA Interaction

RNA interaction prediction can be achieved similarly to the prediction of single sequence RNA secondary structure. The set of base pairs formed between two RNA molecules can be divided into a set of RNA secondary sub-structures. Presented below is the core of the hybridization recursion defined for RNAhybrid[37] with the notation adapted from the original to fit the unified notation used in this document.

$$H_{i,j} = \begin{cases} \min \left\{ \begin{array}{l} stacking(i,j,H_{i+1,j+1}) \\ \min\limits_{i+2 \le k \le \min(i+16,m-1)} \left\{ bulge^m(i,j,k,H_{k,j+1}) \right\} \\ \min\limits_{j+2 \le l \le \min(j+16,n-1)} \left\{ bulge^n(i,j,l,H_{i+1,l}) \right\} \\ \min\limits_{\substack{i+2 \le k \le \min(i+16,m-1) \\ j+2 \le l \le \min(j+16,n-1)}} \left\{ internalLoop(i,j,k,l,H_{k,l}) \right\} \\ openEnd(i,j,m-1,n-1) \end{array} \right\} & \text{if } \pi(x_i, y_j) \\ \infty & \text{otherwise} \end{cases}$$

An optimal structure is calculated by minimizing over a score comprised of the thermodynamic contributions [42] of a set of RNA secondary sub-structures.

2.4. RNA Consensus Structure

RNAalifold [5] computes an RNA consensus secondary structure by folding an alignment of homologous sequences. Secondary structure elements able to be formed by each sequence in the alignment contribute to sum total energy.
A base pair conservation score is introduced, that penalizes gaps in sequences when alignment columns are paired. The score also provides a bonus to alignment columns where consistent and compensating mutations occur in paired bases.

3. Consensus Interaction Prediction

Let $\mathbb{M}$ and $\mathbb{N}$ be alignments of mRNA and ncRNA sequences. The number of sequences in alignment $\mathbb{M}$ is defined as $|\mathbb{M}|$. $\mathbb{M}^x$ denotes the x^{th} row, or sequence, of the mRNA alignment. $\mathbb{M}^x$ and $\mathbb{N}^x$ form a sequence pair. As such,

$$|\mathbb{M}| = |\mathbb{N}|$$

The following constants are used in the calculation of a Consensus Interaction Score, and although mentioned where relevant are provided here for reference.

DI	Duplex initiation penalty
NGC	Non GC pair penalty for the right-most base pair of an interaction
ILL	Maximum allowable internal loop length size
ϕ_γ	Co-variance bonus weighting
ϕ_r	0
ϕ_f	0

3.1. Mapping

To deal with gaps in the alignments the following mapping function has been introduced.

$$\overline{\mathbb{M}}^x = \mathbb{M}_z : z = 1 \ldots m, \mathbb{M}_z \in \{A, C, G, U\}$$

$\overline{\mathbb{M}}^x$ is the raw, or un-gapped sequence in the x^{th} alignment row. $\overline{\mathbb{N}}$ is similarly defined. These raw sequence mappings are required primarily for the calculation of accessibility, however they are also used to access nucleotides. $\overline{\mathbb{M}}^x_i$, where i is an alignment column index. $\dot{M}^x_i$ provides access to the index of i in $\overline{\mathbb{M}}^x$.

The following function is useful for determining whether an RNA secondary sub-structure should be a stacking, bulge or internal loop.

$$\dot{M}^x_{i\Delta p} = |\dot{M}^x_i - \dot{M}^x_p|$$

3.2. Utility

The functions in this section support the calculation of a Consensus Interaction Structure. The hamming distance of two alignment positions (nucleotides or gaps) is used in the calculation of the co-variance bonus.

$$h(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

Sequence-pair specific left-most base pair of optimal interaction with left-end i, j.

$$R_{ip;jq} = \biguplus_{0 \le x < |\mathbb{M}|} R^x_{ip;jq}$$

$$R^x_{ip;jq} = \begin{cases} r^x_{i;j} :=< i, j > & \text{if } \pi^x_{i;j} \\ r^x_{i;j} := r^x_{p;q} & \text{if } r^x_{p;q} \neq \emptyset \\ r^x_{i;j} := \emptyset & \text{otherwise} \end{cases}$$

Right-end base pair of optimal interaction with left-end i, j.

$$F_{ip;jq} = \begin{cases} f_{i;j} :=< i, j > & \text{if } (i = p) \wedge (j = q) \\ f_{i;j} := f_{p;q} & \text{otherwise} \end{cases}$$

3.3. Accessibility

Intra-molecular energy contributions to the RNA-RNA complex are calculated using the partition function, as explained in the previous chapter. In implementation, RNAup [30] is used to calculate these values. Accessibility calculation is unchanged from how it is specified in IntaRNA [9], except that the accessibility of an alignment region is calculated as the average of the single sequence accessibilities for the same region, as below. The notation has also been adapted for clarity.

$$ED^{\mathbb{M}}_{ik} = \frac{1}{|\mathbb{M}|} \sum_{0 \le x < |\mathbb{M}|} ED(\overline{\mathbb{M}}^x_{ik})$$

$$ED_{ik;jl} = ED^{\mathbb{M}}_{ik} + ED^{\mathbb{N}}_{jl}$$

$$ED^{\Delta}_{ipk;jql} = ED_{ik;jl} - ED_{pk;ql}$$

3.4. Hybridization

The hybridization score of an interaction is the inter-molecular energy contribution of base pairing between RNA sequences. The following recursions have been adapted from IntaRNA [9] to incorporate evolutionary information in the same manner as RNAalifold [5]. By averaging the energy contributions of RNA secondary sub-structures closed by indices (i, j), (p, q) a score is obtained that incorporates both thermodynamic and evolutionary information.

The following recursions pertain to the right end of an interaction (i.e. dangling ends).

$$L^d_{ip;jq} = \frac{1}{|\mathbb{M}|} \sum_{0 \le x < |\mathbb{M}|} L^{dx}_{ip;jq}$$

$$L^{dx}_{ip;jq} = \begin{cases} dl^x_{ip;jq} + NGC_{i;j} & \text{if } \pi^x_{i;j} \\ 0 & \text{otherwise} \end{cases}$$

The following provides access to the Turner [42] free energies of RNA sub-structures.

$$dl^x_{ip;jq} = \begin{cases} 0 & \text{if } (p = i) \wedge (q = j) \\ ed5(i,j,q) & \text{if } (p = i) \wedge (q = j+1) \\ ed3(i,j,p) & \text{if } (p = i+1) \wedge (q = j) \\ ed5(i,j,q) + ed3(i,j,p) & \text{if } (p = i+1) \wedge (q = j+1) \end{cases}$$

For clarity of exposition, the mapping from alignment indices to raw sequence nucleotides has been omitted.

Recursion pertaining to hybridization energies for the purposes of extending an interaction.

$$L_{ip;jq} = \frac{1}{|\mathbb{M}|} \sum_{0 \le x < |\mathbb{M}|} L^x_{ip;jq}$$

The significant difference to RNAalifold [5], RNAhybrid [37] and IntaRNA [9] is the addition of cases 2, 3, 4 below.

$$L^x_{ip;jq} = \begin{cases} el^x_{ip;jq} & \text{if } \pi^x_{i;j} \wedge \pi^x_{p;q} \\ el^x_{ik;jl} & \text{if } \pi^x_{i;j} \wedge \neg\pi^x_{p;q} \wedge r^x_{p;q} \neq \emptyset \\ Q^x_{i;j} & \text{if } \pi^x_{i;j} \wedge \neg\pi^x_{p;q} \wedge r^x_{p;q} = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } r^x_{p;q} =< k, l >$$

$$Q^x_{i;j} = \min_{\substack{i \le p \le min(i+1,m) \\ j \le q \le min(j+1,n)}} \{L^{dx}_{ip;jq} + \frac{DI}{|\mathbb{M}|}\}$$

The first important difference is when (i, j) cannot pair, 0 is returned instead of ∞. Non consensus interaction prediction methods [9] [37] would disallow the pairing, and RNAalifold [5] penalizes the mismatch with an explicit ad hoc penalty. The purpose in not penalizing a mismatch can be seen in case 2.

If $r^x_{p;q} \neq \emptyset$, it contains the left-most base-pair in the interaction we are testing for extension, and as such (i, j) can form an RNA secondary sub-structure with $(k, l) = r^x_{p;q}$. It is important to note that because (p, q) in x cannot pair, the region $\mathbb{M}^x_{p,k}, \mathbb{N}^x_{q,l}$ has not yet contributed to the score of the interaction. The intention is that consensus will be scored more naturally than with the introduction of an ad-hoc penalty.

Case 3 covers the case where a sequence in an interaction has not formed a right end base pair. As such, $Q^x_{i;j}$ incorporates the interaction start structure.

Access to RNA secondary sub-structures with type of structure dependent on mapped raw

sequence index differences.

$$el^x_{ip;jq} = \begin{cases} stack(i,p,j,q) & \text{if } \dot{M}^x_{i\Delta p} = 1 \wedge \dot{N}^x_{j\Delta q} = 1 \\ iloop(i,p,j,q) & \text{if } (\dot{M}^x_{i\Delta p} > 1 \wedge \dot{N}^x_{j\Delta q} > 1) \\ bulge(i,p,j,q) & \text{if } (\dot{M}^x_{i\Delta p} = 1 \wedge \dot{N}^x_{j\Delta q} > 1) \vee (\dot{M}^x_{i\Delta p} > 1 \wedge \dot{N}^x_{j\Delta q} = 1) \end{cases}$$

Again, mapping of alignment indices to raw sequence nucleotides has been omitted for clarity.

3.5. Covariance Scoring

In order to integrate evolutionary information, and calculate a *Consensus* Interaction Score, a bonus is added for consistent and compensatory mutations of pairing bases. This method has been adapted from RNAalifold [5] to removed ad hoc gap penalties and provide a normalized score. Interaction partners (mRNA-ncRNA sequence pairs) are compared pair-wise (with other mRNA-ncRNA sequence pairs, or rows in the alignments) and a bonus is added for any shared base pairs that are comprised of different nucleotides.

$$\Gamma_{i;j} = \frac{1}{\binom{|\mathbb{M}|}{2}} \sum_{0 \le x < y < |\mathbb{M}|} \begin{cases} h(\mathbb{M}^x_i, \mathbb{M}^y_i) + h(\mathbb{N}^x_j, \mathbb{N}^y_j) & \text{if } \pi^x_{i;j} \wedge \ \pi^y_{i;j} \\ 0 & \text{otherwise} \end{cases}$$

3.6. Combined Consensus Interaction Score

The Combined Consensus Interaction Score integrates the inter-molecular energy contribution (hybridization), intra-molecular energy contribution (accessibility) and the evolutionary information (co-variance bonus) to calculate a score for each possible (i, j) forming a left-end base pair to an interaction region.
The optimal consensus score for an alignment of mRNAs and ncRNAs can be found by a minimization over all (i, j). From this, the interaction producing such a score can be found by dynamic programming traceback.

$$C(\mathbb{M}, \mathbb{N}) = \min_{\substack{0 \le i < m \\ 0 \le j < n}} C^{\mathbb{M};\mathbb{N}}_{i;j}$$

In its simplest form, the score for left-end base pair (i, j) can be calculated as the co-variance bonus of the alignment positions and the energetically most favourable of two choices. The first possibility is to start a new interaction region, with (i, j) as the right most base pair. Alternatively, (i, j) can form the left-most base pair of an existing interaction region.

$$C^{\mathbb{M};\mathbb{N}}_{i;j} = \min\{C^{start}_{i;j}, C^{grow}_{i;j}\} + \phi_\gamma \Gamma_{i;j}$$

To calculate a score for both possibilities and determine the optimal choice, the following additions are needed. It is important to note that, while F and R appear alongside C^{start} and C^{grow} they do not directly contribute to the score of a structure. That is, $\phi_r = 0$ and

$\phi_f = 0$.

$$C_{i;j}^{\mathbb{M};\mathbb{N}} = \min \left\{ \begin{array}{l} C_{i;j}^{start} + \phi_r R_{ii;jj} + \phi_f F_{ii;jj} \\ C_{i;j}^{grow} + \phi_r R_{ip;jq} + \phi_f F_{ip;jq} \end{array} \right\} + \phi_\gamma \Gamma_{i;j},$$

where p and q are obtained by applying an argmin in C^{grow}

Both F and R update matrices that support the calculation of hybridization energy and accessibility. They are integrated here to make explicit the point at which the support matrices are updated, and with what values. Note, the inclusion of these functions assumes only one case succeeds, and therefore F and R are only updated once. That is, if C^{start} is minimum, then F and R are updated with the above values and vice-versa.

3.6.1. Starting a new interaction

To start a new interaction at (i, j) all four possible cases of dangling ends must be evaluated and the accessibility relevant to each case included in the search for a minimal energy structure. A duplex initiation penalty is also included.

$$C_{i;j}^{start} = \min_{\substack{i \le p \le min(i+1,m) \\ j \le q \le min(j+1,n)}} \{L_{ip;jq}^{d} + ED_{ip;jq} \ + DI\}$$

3.6.2. Extending an existing interaction

Extending an existing interaction to a new left-most base pair (i, j) involves finding a minimum score over all the possible left-most base pairs $((p, q))$ to extend from. The range of previous base pairs to extend from has been limited by the ILL, internal loop length. The score derived from extending an interaction region is comprised of a hybridization component L, accessibility component ED and the score of the previous interaction region ending at (p, q). The support matrix f contains the right-end base pair of the interaction with left-end base pair (p, q).

$$C_{i;j}^{grow} = \min_{\substack{i < p \le min(i+ILL,m) \\ j < q \le min(j+ILL,n)}} \{L_{ip;jq} + ED_{ipk;jql}^{\Delta} + C_{p;q}\}, \text{ where } (k, l) = f_{p;q}$$

4. Implementation and Evaluation of the Algorithm

This chapter details how the Consensus Interaction Prediction algorithm is implemented. The data and methods used to evaluate the quality of predicted interaction structures are also explained.

4.1. Implementation

The Consensus Interaction Prediction algorithm is implemented in C++. The implementation makes use of the Vienna RNA package [18], the University of Freiburg Bioinformatics Department BIU library(www.bioinf.uni-freiburg.de/SW/BIU/), C++ Boost library(www.boost.org) and re-uses some code from the IntaRNA [9] program.

The program takes the following as input:

- A co-variance weighting (ϕ_γ)
- An input file in fasta format containing $|\mathbb{M}|$ aligned mRNA sequences
- An input file in fasta format containing $|\mathbb{N}|$ aligned ncRNA sequences

As the tool operates on sequences pairs, it is necessary that $|\mathbb{M}| = |\mathbb{N}|$. As output, the tool provides a mapping of the predicted consensus structure to each input sequence pair. Dot-bracket notation, IntaRNA style format and a listing of base pair indices are available as output options for hybridization visualization.
Source code for the recursions relating to the algorithm is provided in Appendix C.

4.2. Evaluation

The performance of the algorithm, in terms of accuracy of interaction prediction was evaluated in the following way. Data sets were prepared as for [40]. The Consensus Interaction Prediction tool was on orthologs of the validated interactions in a number of organisms. A total of 29 interactions were predicted
IntaRNA single sequence pair predictions were created for each interaction with a validated structure using mRNA and ncRNA genes from the reference organisms (AE006468, U00096). The predicted structures were then compared with the experimentally validated structures using the measures described in the following section. The results of the evaluation are contained in Appendix B, and discussed in the following chapter.

4.2.1. Measures

The following measures are used to evaluate the quality of interactions. Both the Consensus Interaction structure and the structure predicted by IntaRNA are compared with an experimentally validated structure.

Sensitivity of a prediction measures the proportion of base pairs, according to the experimentally validated interaction, correctly identified by the algorithm.

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Specifically, true positives are base pairs the prediction shares with the validated structure. False negatives are base pairs in the validated structure, not contained in the prediction.

The positive predictive value (PPV) of a structure is the proportion of predicted base pairs which are contained in the experimentally validated structure.

$$\text{PPV} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

True positives are as in the sensitivity measure, false positives are base pairs predicted by the algorithm but not contained in the validated structure.

F-measure combines both sensitivity and PPV. As such, it is primarily used to compare structures predicted by IntaRNA and the Consensus Interaction method. It is also used to calculate the optimal value of ϕ_γ.

$$\text{F-measure} = 2 \times \frac{\text{sensitivity} \times \text{PPV}}{\text{sensitivity} + \text{PPV}}$$

4.2.2. Datasets

The datasets used in the evaluation of the Consensus Interaction Prediction algorithm are a subset of those used in the evaluation of PETcofold [40]. The datasets with references are in Appendix D.

5. Results and Discussion

Table 5.1 lists the F-measure score for IntaRNA and the Consensus Interaction Prediction algorithm with the co-variance weighting set at the global optimum and a per sequence optimum. The following table lists how many cases each tool performs as well as or better than the others.

CIP $\phi_\gamma = 0.00$	CIP $argmax\phi_\gamma$	IntaRNA
13	16	17

The average F-measure for both data sets was measured with 10 different values of ϕ_γ.

ϕ_γ	U00096	AE006468	average
0.00	0.635	0.658	0.647
0.02	0.628	0.657	0.643
0.04	0.625	0.657	0.641
0.06	0.626	0.551	0.589
0.08	0.628	0.542	0.585
0.10	0.627	0.542	0.585
0.12	0.619	0.542	0.581
0.14	0.610	0.542	0.576
0.16	0.576	0.542	0.559
0.18	0.573	0.444	0.509

IntaRNA scored the following average F-measure scores.

U00096	AE006468	average
0.680	0.602	0.641

5.1. Discussion

The Consensus Interaction Prediction algorithm performs favourably when compared with IntaRNA, however the optimal weighting of $\phi_\gamma = 0.00$ suggests that the co-variance bonus detracts from the quality of prediction. This is likely a product of the chosen evaluation method. The consensus interaction structure was predicted using an alignment of sequences from different organisms. The mapping back to the reference organism and comparison to a single sequence pair prediction in the same organism is likely to be an unfair comparison. A genome wide scan, and comparison with another consensus prediction tool, such as PETcofold [40] would better demonstrate the capability of the algorithm to predict a conserved and energetically favourable interaction structure. Overall, the method looks promising, however more analysis is necessary to determine the true benefit of the Consensus Interaction Prediction algorithm and how the method can be improved.

ncRNA, mRNA, Organism	CIP with $\phi_\gamma = 0.00$	CIP with best ϕ_γ	IntaRNA
CyaR,luxs,U00096	0.889	0.889	0.889
CyaR,yqaE,U00096	0.643	0.643	0.857
DsrA,hns,U00096	0.621	0.621	0.690
DsrA,rpoS,U00096	0.692	0.717	0.792
GcvB,sstT,U00096	0.000	0.000	0.000
GlmZ,glmS,U00096	0.636	0.636	1.000
MicA,ompA,U00096	0.968	1.000	0.897
MicC,ompC,U00096	0.800	0.800	0.842
MicF,ompF,U00096	0.941	0.941	0.941
OmrA,cirA,U00096	0.621	0.621	0.462
OmrA,ompR,U00096	0.750	0.750	0.750
OmrA,ompT,U00096	0.514	0.632	0.514
OxyS,fhlA,U00096	0.667	0.667	0.545
RprA,rpoS,U00096	0.000	0.000	0.286
RyhB,fur,U00096	0.000	0.000	0.353
RyhB,sodB,U00096	1.000	1.000	0.900
SgrS,ptsG,U00096	0.850	0.850	0.850
CyaR,ompX,AE006468	0.478	0.478	0.478
GcvB,argT,AE006468	0.750	0.750	0.848
GcvB,gltI,AE006468	0.316	0.316	0.000
GcvB,livJ,AE006468	0.955	0.955	0.000
GcvB,livK,AE006468	0.743	0.743	0.722
GcvB,oppA,AE006468	0.905	0.905	0.978
GcvB,STM4351,AE006468	0.588	0.588	0.529
MicA,lamB,AE006468	0.286	0.286	0.902
MicC,nmpC,AE006468	0.917	0.960	0.960

Table 5.1.: Comparison of F-measure, Consensus Interaction Prediction and IntaRNA

A. Sensitivity, PPV, F-measure of Consensus Interaction Prediction

A.1. CyaR,luxs,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.800	1.000	0.889
0.02	0.800	1.000	0.889
0.04	0.800	1.000	0.889
0.06	0.800	1.000	0.889
0.08	0.800	1.000	0.889
0.10	0.800	1.000	0.889
0.12	0.800	1.000	0.889
0.14	0.800	1.000	0.889
0.16	0.800	1.000	0.889
0.18	0.800	1.000	0.889

A.2. CyaR,nadE,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.909	1.000	0.952
0.02	0.909	0.625	0.741
0.04	0.909	0.625	0.741
0.06	0.909	0.625	0.741
0.08	0.909	0.625	0.741
0.10	0.909	0.625	0.741
0.12	0.909	0.556	0.690
0.14	0.909	0.556	0.690
0.16	0.909	0.556	0.690
0.18	0.909	0.556	0.690

A.3. CyaR,ompX,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.900	0.375	0.529
0.02	0.900	0.375	0.529
0.04	0.900	0.375	0.529
0.06	0.900	0.375	0.529
0.08	0.900	0.375	0.529
0.10	0.900	0.375	0.529
0.12	0.900	0.375	0.529
0.14	0.900	0.375	0.529
0.16	0.900	0.375	0.529
0.18	0.900	0.375	0.529

A.4. CyaR,yqaE,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.750	0.562	0.643
0.02	0.750	0.562	0.643
0.04	0.750	0.562	0.643
0.06	0.750	0.562	0.643
0.08	0.750	0.562	0.643
0.10	0.750	0.562	0.643
0.12	0.750	0.562	0.643
0.14	0.750	0.562	0.643
0.16	0.750	0.562	0.643
0.18	0.750	0.562	0.643

A.5. DsrA,hns,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.692	0.562	0.621
0.02	0.692	0.562	0.621
0.04	0.692	0.562	0.621
0.06	0.692	0.562	0.621
0.08	0.692	0.562	0.621
0.10	0.692	0.562	0.621
0.12	0.692	0.562	0.621
0.14	0.692	0.562	0.621
0.16	0.692	0.562	0.621
0.18	0.692	0.562	0.621

A.6. DsrA,rpoS,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.692	0.692	0.692
0.02	0.692	0.692	0.692
0.04	0.692	0.692	0.692
0.06	0.731	0.704	0.717
0.08	0.731	0.704	0.717
0.10	0.731	0.704	0.717
0.12	0.731	0.704	0.717
0.14	0.731	0.704	0.717
0.16	0.731	0.704	0.717
0.18	0.731	0.704	0.717

A.7. GcvB,sstT,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.000	0.000	0.000
0.02	0.000	0.000	0.000
0.04	0.000	0.000	0.000
0.06	0.000	0.000	0.000
0.08	0.000	0.000	0.000
0.10	0.000	0.000	0.000
0.12	0.000	0.000	0.000
0.14	0.000	0.000	0.000
0.16	0.000	0.000	0.000
0.18	0.000	0.000	0.000

A.8. GlmZ,glmS,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.467	1.000	0.636
0.02	0.467	1.000	0.636
0.04	0.467	1.000	0.636
0.06	0.467	1.000	0.636
0.08	0.467	1.000	0.636
0.10	0.467	1.000	0.636
0.12	0.467	1.000	0.636
0.14	0.467	1.000	0.636
0.16	0.000	0.000	0.000
0.18	0.000	0.000	0.000

A.9. MicA,ompA,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.938	1.000	0.968
0.02	0.938	1.000	0.968
0.04	0.938	1.000	0.968
0.06	0.938	1.000	0.968
0.08	1.000	1.000	1.000
0.10	1.000	1.000	1.000
0.12	1.000	1.000	1.000
0.14	1.000	1.000	1.000
0.16	1.000	1.000	1.000
0.18	1.000	1.000	1.000

A.10. MicC,ompC,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.727	0.889	0.800
0.02	0.727	0.889	0.800
0.04	0.727	0.889	0.800
0.06	0.727	0.889	0.800
0.08	0.727	0.889	0.800
0.10	0.727	0.842	0.780
0.12	0.682	0.682	0.682
0.14	0.682	0.682	0.682
0.16	0.682	0.682	0.682
0.18	0.682	0.682	0.682

A.11. MicF,ompF,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.960	0.923	0.941
0.02	0.960	0.923	0.941
0.04	0.960	0.923	0.941
0.06	0.960	0.923	0.941
0.08	0.960	0.923	0.941
0.10	0.960	0.923	0.941
0.12	0.960	0.923	0.941
0.14	0.960	0.632	0.762
0.16	0.960	0.632	0.762
0.18	0.960	0.632	0.762

A.12. OmrA,cirA,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.450	1.000	0.621
0.02	0.450	1.000	0.621
0.04	0.450	1.000	0.621
0.06	0.450	1.000	0.621
0.08	0.450	1.000	0.621
0.10	0.450	1.000	0.621
0.12	0.450	1.000	0.621
0.14	0.450	1.000	0.621
0.16	0.450	1.000	0.621
0.18	0.450	0.750	0.563

A.13. OmrA,ompR,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.600	1.000	0.750
0.02	0.600	1.000	0.750
0.04	0.600	1.000	0.750
0.06	0.600	1.000	0.750
0.08	0.600	1.000	0.750
0.10	0.600	1.000	0.750
0.12	0.600	1.000	0.750
0.14	0.600	1.000	0.750
0.16	0.600	1.000	0.750
0.18	0.600	1.000	0.750

A.14. OmrA,ompT,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.346	1.000	0.514
0.02	0.462	1.000	0.632
0.04	0.462	1.000	0.632
0.06	0.462	1.000	0.632
0.08	0.462	1.000	0.632
0.10	0.462	1.000	0.632
0.12	0.462	1.000	0.632
0.14	0.462	1.000	0.632
0.16	0.462	1.000	0.632
0.18	0.462	1.000	0.632

A.15. OxyS,fhlA,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.500	1.000	0.667
0.02	0.500	1.000	0.667
0.04	0.500	0.800	0.615
0.06	0.500	0.800	0.615
0.08	0.500	0.800	0.615
0.10	0.500	0.800	0.615
0.12	0.500	0.800	0.615
0.14	0.500	0.800	0.615
0.16	0.500	0.800	0.615
0.18	0.500	0.800	0.615

A.16. RprA,rpoS,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.000	0.000	0.000
0.02	0.000	0.000	0.000
0.04	0.000	0.000	0.000
0.06	0.000	0.000	0.000
0.08	0.000	0.000	0.000
0.10	0.000	0.000	0.000
0.12	0.000	0.000	0.000
0.14	0.000	0.000	0.000
0.16	0.000	0.000	0.000
0.18	0.000	0.000	0.000

A.17. RyhB,fur,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.000	0.000	0.000
0.02	0.000	0.000	0.000
0.04	0.000	0.000	0.000
0.06	0.000	0.000	0.000
0.08	0.000	0.000	0.000
0.10	0.000	0.000	0.000
0.12	0.000	0.000	0.000
0.14	0.000	0.000	0.000
0.16	0.000	0.000	0.000
0.18	0.000	0.000	0.000

A.18. RyhB,sodB,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	1.000	1.000	1.000
0.02	1.000	0.900	0.947
0.04	1.000	0.900	0.947
0.06	1.000	0.900	0.947
0.08	1.000	0.900	0.947
0.10	1.000	0.900	0.947
0.12	1.000	0.900	0.947
0.14	1.000	0.900	0.947
0.16	1.000	0.900	0.947
0.18	1.000	0.900	0.947

A.19. SgrS,ptsG,U00096

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.739	1.000	0.850
0.02	0.739	1.000	0.850
0.04	0.739	1.000	0.850
0.06	0.739	1.000	0.850
0.08	0.739	1.000	0.850
0.10	0.739	1.000	0.850
0.12	0.739	1.000	0.850
0.14	0.739	1.000	0.850
0.16	0.739	1.000	0.850
0.18	0.739	1.000	0.850

A.20. CyaR,ompX,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.500	0.458	0.478
0.02	0.500	0.458	0.478
0.04	0.500	0.458	0.478
0.06	0.500	0.458	0.478
0.08	0.500	0.458	0.478
0.10	0.500	0.458	0.478
0.12	0.500	0.458	0.478
0.14	0.500	0.458	0.478
0.16	0.500	0.458	0.478
0.18	0.500	0.458	0.478

A.21. GcvB,argT,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.938	0.625	0.750
0.02	0.938	0.625	0.750
0.04	0.938	0.625	0.750
0.06	0.938	0.625	0.750
0.08	0.938	0.536	0.682
0.10	0.938	0.536	0.682
0.12	0.938	0.536	0.682
0.14	0.938	0.536	0.682
0.16	0.938	0.536	0.682
0.18	0.938	0.536	0.682

A.22. GcvB,dppA,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.941	0.485	0.640
0.02	0.941	0.485	0.640
0.04	0.941	0.485	0.640
0.06	0.941	0.485	0.640
0.08	0.941	0.485	0.640
0.10	0.941	0.485	0.640
0.12	0.941	0.485	0.640
0.14	0.941	0.485	0.640
0.16	0.941	0.485	0.640
0.18	0.941	0.485	0.640

A.23. GcvB,gltI,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.500	0.231	0.316
0.02	0.500	0.231	0.316
0.04	0.500	0.231	0.316
0.06	0.500	0.231	0.316
0.08	0.500	0.231	0.316
0.10	0.500	0.231	0.316
0.12	0.500	0.231	0.316
0.14	0.500	0.231	0.316
0.16	0.500	0.231	0.316
0.18	0.500	0.231	0.316

A.24. GcvB,livJ,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.955	0.955	0.955
0.02	0.955	0.955	0.955
0.04	0.955	0.955	0.955
0.06	0.955	0.328	0.488
0.08	0.955	0.328	0.488
0.10	0.955	0.328	0.488
0.12	0.955	0.328	0.488
0.14	0.955	0.328	0.488
0.16	1.000	0.344	0.512
0.18	0.955	0.328	0.488

A.25. GcvB,livK,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	1.000	0.591	0.743
0.02	1.000	0.591	0.743
0.04	1.000	0.591	0.743
0.06	1.000	0.591	0.743
0.08	1.000	0.591	0.743
0.10	1.000	0.591	0.743
0.12	1.000	0.591	0.743
0.14	1.000	0.591	0.743
0.16	1.000	0.591	0.743
0.18	1.000	0.591	0.743

A.26. GcvB,oppA,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.864	0.950	0.905
0.02	0.864	0.950	0.905
0.04	0.864	0.950	0.905
0.06	0.864	0.950	0.905
0.08	0.864	0.950	0.905
0.10	0.864	0.950	0.905
0.12	0.864	0.950	0.905
0.14	0.864	0.950	0.905
0.16	0.864	0.950	0.905
0.18	0.864	0.950	0.905

A.27. GcvB,STM4351,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.909	0.435	0.588
0.02	0.909	0.435	0.588
0.04	0.909	0.435	0.588
0.06	0.000	0.000	0.000
0.08	0.000	0.000	0.000
0.10	0.000	0.000	0.000
0.12	0.000	0.000	0.000
0.14	0.000	0.000	0.000
0.16	0.000	0.000	0.000
0.18	0.000	0.000	0.000

A.28. MicA,lamB,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.217	0.417	0.286
0.02	0.217	0.385	0.278
0.04	0.217	0.385	0.278
0.06	0.217	0.385	0.278
0.08	0.217	0.208	0.213
0.10	0.217	0.208	0.213
0.12	0.217	0.208	0.213
0.14	0.217	0.208	0.213
0.16	0.217	0.167	0.189
0.18	0.217	0.167	0.189

A.29. MicC,nmpC,AE006468

ϕ_γ	Sensitivity	PPV	F-measure
0.00	0.917	0.917	0.917
0.02	0.917	0.917	0.917
0.04	0.917	0.917	0.917
0.06	0.917	0.917	0.917
0.08	1.000	0.923	0.960
0.10	1.000	0.923	0.960
0.12	1.000	0.923	0.960
0.14	1.000	0.923	0.960
0.16	1.000	0.923	0.960
0.18	0.000	0.000	0.000

B. Sensitivity, PPV, F-measure of IntaRNA Interaction Prediction

ncRNA, mRNA, Organism	Sensitivity	PPV	F-measure
CyaR,luxs,U00096	0.800	1.000	0.889
CyaR,nadE,U00096	-	-	-
CyaR,ompX,U00096	-	-	-
CyaR,yqaE,U00096	0.750	1.000	0.857
DsrA,hns,U00096	0.769	0.625	0.690
DsrA,rpoS,U00096	0.808	0.778	0.792
GcvB,sstT,U00096	0.000	0.000	0.000
GlmZ,glmS,U00096	1.000	1.000	1.000
MicA,ompA,U00096	0.812	1.000	0.897
MicC,ompC,U00096	0.727	1.000	0.842
MicF,ompF,U00096	0.960	0.923	0.941
OmrA,cirA,U00096	0.300	1.000	0.462
OmrA,ompR,U00096	0.600	1.000	0.750
OmrA,ompT,U00096	0.346	1.000	0.514
OxyS,fhlA,U00096	0.375	1.000	0.545
RprA,rpoS,U00096	0.316	0.261	0.286
RyhB,fur,U00096	0.214	1.000	0.353
RyhB,sodB,U00096	1.000	0.818	0.900
SgrS,ptsG,U00096	0.739	1.000	0.850
CyaR,ompX,AE006468	0.500	0.458	0.478
GcvB,argT,AE006468	0.875	0.824	0.848
GcvB,dppA,AE006468	-	-	-
GcvB,gltI,AE006468	0.000	0.000	0.000
GcvB,livJ,AE006468	0.000	0.000	0.000
GcvB,livK,AE006468	1.000	0.565	0.722
GcvB,oppA,AE006468	1.000	0.957	0.978
GcvB,STM4351,AE006468	0.818	0.391	0.529
MicA,lamB,AE006468	1.000	0.821	0.902
MicC,nmpC,AE006468	1.000	0.923	0.960

C. Source Code for Algorithm Recursions

C.1. $C_{i;j}$

```
static double C(int i, int j,
PairMatrix& f, DoubleMatrix& gamma, BoolMatrix3D& pi, DoubleMatrix& avg_pi,
DoubleMatrix& avg_ED_mRNA, DoubleMatrix& avg_ED_ncRNA,
DoubleMatrix& C_ij,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA,
vector<char>& mode_base_mRNA, vector<char>& mode_base_ncRNA,
PairMatrix3D& r, StructMatrix& u)
{
double grow = MAX_DOUBLE;
double start = MAX_DOUBLE;
double ded = MAX_DOUBLE;
double oldc = MAX_DOUBLE;
double newl = MAX_DOUBLE;
Pair argmin_pq = undef;

if ((unsigned)i != rawmap_mRNA.numColumns()-1 &&
(unsigned)j != rawmap_ncRNA.numColumns()-1)
{
grow = Cgrow(i, j, f,
pi, avg_pi, avg_ED_mRNA, avg_ED_ncRNA, C_ij,
raw_mRNA, raw_ncRNA,
rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA,
r, argmin_pq, ded, oldc, newl);
}

/* ## start a new interaction at i,j ## */
start = C_start(i, j,
pi, avg_pi, avg_ED_mRNA, avg_ED_ncRNA, C_ij,
raw_mRNA, raw_ncRNA,
rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA);
```

```
if (grow == MAX_DOUBLE && start == MAX_DOUBLE)
{
//return MAX_DOUBLE;
}

/* ## test which score is better */
if (grow < start && (unsigned)i != rawmap_mRNA.numColumns()-1 &&
 (unsigned)j != rawmap_ncRNA.numColumns()-1)
{
Pair current(i,j);
f[i][j] = f[argmin_pq.get<0>()][argmin_pq.get<1>()];
u[i][j] = u[argmin_pq.get<0>()][argmin_pq.get<1>()];
u[i][j].insert(current);

for (unsigned x = 0; x < r.size(); ++x)
{
if (pi[x][current.get<0>()][current.get<1>()] == 1)//error?
{
r[x][i][j] = current;
}
else if (r[x][argmin_pq.get<0>()][argmin_pq.get<1>()] != undef)
{
r[x][i][j] = r[x][argmin_pq.get<0>()][argmin_pq.get<1>()];
}
}
}
else
{
Pair start(i,j);
u[i][j].insert(start);
f[i][j] = start; //set the current bp to be the final bp
for (unsigned x = 0; x < r.size(); ++x)
{
if (pi[x][i][j] == 1)
{
r[x][i][j] = start;
}
}
}

double min_score = min(start, grow);
return PHI_GAMMA*gamma[i][j] + min_score;
}
```

C.2. C^{start}

```
static double dl(int i, int p, int j, int q, int x,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{
if (p == i && q == j)
{
return 0.0;
}
else if (p == i && q == j + 1) //try for 5' dangle
{
int nextj = map_right(rawmap_ncRNA, rightmap_ncRNA, x, j);
if (nextj != -1)
{
return ed5(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]), char2int_base(raw_ncRNA[x][nextj]));
}
else
{
return 0.0;
}

}
else if (p == i + 1 && q == j) //try for 3' dangle
{
int nexti = map_right(rawmap_mRNA, rightmap_mRNA, x, i);
if (nexti != -1)
{
return ed3(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]), char2int_base(raw_mRNA[x][nexti]));
}
else
{
return 0.0;
}
}
else //if (p == i+1 && q == j+1)
{
int nexti = map_right(rawmap_mRNA, rightmap_mRNA, x, i);
int nextj = map_right(rawmap_ncRNA, rightmap_ncRNA, x, j);

if (nexti == -1 && nextj == -1)
{
return 0.0;
```

```
}
else if (nexti == -1 && nextj != -1)
{
return ed5(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
char2int_base(raw_ncRNA[x][nextj]));
}
else if (nexti != -1 && nextj == -1)
{
return ed3(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
char2int_base(raw_mRNA[x][nexti]));
}
else
{
return ed5(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
char2int_base(raw_ncRNA[x][nextj])) +
ed3(char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
 char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
 char2int_base(raw_mRNA[x][nexti]));
}
}
}

static double L_dx(int i, int p, int j, int q, int x,
BoolMatrix3D& pi,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{
if (pi[x][i][j] == 1)
{
return  (
dl(i, p, j, q, x,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA) +
nonGC_penalty(raw_mRNA[x][rawmap_mRNA[x][i]],
 raw_ncRNA[x][rawmap_ncRNA[x][j]])
);
}
else
{
return 0.0;
}
}
```

```
static double L_d(int i, int p, int j, int q,
BoolMatrix3D& pi,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{
double sum = 0;

for (unsigned x = 0; x < raw_mRNA.size(); ++x)
{
sum += L_dx(i, p, j, q, x, pi,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA);
}
return sum/raw_mRNA.size();
}

static double C_start(int i, int j,
BoolMatrix3D& pi, DoubleMatrix& avg_pi,
DoubleMatrix& avg_ED_mRNA, DoubleMatrix& avg_ED_ncRNA,
DoubleMatrix& C_ij,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{

double min_score = MAX_DOUBLE;
double temp_score = min_score;

for (int p = i; p <= min(i+1, (int)rawmap_mRNA.numColumns()-1); ++p)
{
for (int q = j; q <= min(j+1, (int)rawmap_ncRNA.numColumns()-1); ++q)
{
temp_score =
(
L_d(i, p, j, q, pi, raw_mRNA, raw_ncRNA,
rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA) +
ED(avg_ED_mRNA, avg_ED_ncRNA, i, p, j, q) +
duplex_init
);
if (temp_score < min_score)
{
min_score = temp_score;
}
```

```
}
}
return min_score;
}
```

C.3. C^{grow}

```
static double el(int i, int p, int j, int q, int x,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{
unsigned pd_M = pos_diff(x, i, p, rawmap_mRNA);
unsigned pd_NC = pos_diff(x, j, q,  rawmap_ncRNA);

if (pd_M == 1 && pd_NC == 1) //stacking
{
return StackingEnergy
(
char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
char2int_base(raw_mRNA[x][rawmap_mRNA[x][p]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][q]])
);
}
else if ((pd_M == 1 && pd_NC > 1)
|| (pd_M > 1 && pd_NC == 1)) //ncRNA bulge OR mRNA bulge
{
return BulgeEnergy
(
pd_NC+pd_M-2,
char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
char2int_base(raw_mRNA[x][rawmap_mRNA[x][p]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][q]])
);
}
else if (pd_M > 1 && pd_NC > 1) //interior loop
{
return InteriorLoopEnergy
(
pd_M-1,
pd_NC-1,
char2int_base(raw_mRNA[x][rawmap_mRNA[x][i]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][j]]),
```

```
char2int_base(raw_mRNA[x][rawmap_mRNA[x][p]]),
char2int_base(raw_ncRNA[x][rawmap_ncRNA[x][q]]),
char2int_base(raw_mRNA[x][map_right(rawmap_mRNA, rightmap_mRNA, x, i)]),
char2int_base(raw_ncRNA[x][map_right(rawmap_ncRNA, rightmap_ncRNA, x, j)]),
char2int_base(raw_mRNA[x][map_left(rawmap_mRNA, leftmap_mRNA, x, p)]),
char2int_base(raw_ncRNA[x][map_left(rawmap_ncRNA, leftmap_ncRNA, x, q)])
);
}
}

static double Rstart(int i, int j, int x,
BoolMatrix3D& pi,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA)
{
double min_score = MAX_DOUBLE;
double temp_score = min_score;

for (int p = i; p <= min(i+1, (int)rawmap_mRNA.numColumns()-1); ++p)
{
for (int q = j; q <= min(j+1, (int)rawmap_ncRNA.numColumns()-1); ++q)
{
temp_score =
(
L_dx(i, p, j, q, x, pi,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA) +
duplex_init/rawmap_mRNA.numRows()

);
if (temp_score < min_score)
{
min_score = temp_score;
}
}
}
return min_score;
}

static double L_x(int i, int p, int j, int q, int x,
BoolMatrix3D& pi,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA,
PairMatrix3D& r)
```

```
{
if (pi[x][i][j] == 1 && pi[x][p][q] == 1)
{
return el(i, p, j, q, x,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA);
}
else if (pi[x][i][j] == 1 && pi[x][p][q] == 0 && r[x][p][q] != undef)
{
return el(i, r[x][p][q].get<0>(), j, r[x][p][q].get<1>(),
x, raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA);
}
else if (pi[x][i][j] == 1 && pi[x][p][q] == 0)
{
return Rstart(i, j, x, pi, raw_mRNA, raw_ncRNA,
rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA);
}

return 0.0;
}

static double L(int i, int p, int j, int q,
BoolMatrix3D& pi,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA,
PairMatrix3D& r)
{
double sum = 0;

for (unsigned x = 0; x < raw_mRNA.size(); ++x)
{
sum += L_x(i, p, j, q, x, pi,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA, r);
}
return sum/raw_mRNA.size();
}

static double Cgrow(int i, int j, PairMatrix& f,
BoolMatrix3D& pi, DoubleMatrix& avg_pi,
DoubleMatrix& avg_ED_mRNA, DoubleMatrix& avg_ED_ncRNA,
DoubleMatrix& C_ij,
vector<string>& raw_mRNA, vector<string>& raw_ncRNA,
```

```
IntMatrix& rawmap_mRNA, IntMatrix& leftmap_mRNA, IntMatrix& rightmap_mRNA,
IntMatrix& rawmap_ncRNA, IntMatrix& leftmap_ncRNA, IntMatrix& rightmap_ncRNA,
PairMatrix3D& r, Pair& argmin_pq, double& ded, double& oldc, double& newl)
{
double min_score = MAX_DOUBLE;
double temp_score = min_score;
int argmin_p = -1;
int argmin_q = -1;
int k = -1;
int l = -1;

for (int p = i+1; p <= min(i+ILOOP_SIZE, (int)rawmap_mRNA.numColumns()-1); ++p)
{
for (int q = j+1; q <= min(j+ILOOP_SIZE, (int)rawmap_ncRNA.numColumns()-1); ++q)
{
k = f[p][q].get<0>();
l = f[p][q].get<1>();

if (p <= k && q <= l)
{

temp_score = (
L(i, p, j, q, pi,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA, r) + //hybridization
delta_ED(avg_ED_mRNA, avg_ED_ncRNA, i, p, k, j, q, l) + //accessibility
C_ij[p][q] //combined score to the right of current loop (DP recursion)
 );
if (temp_score < min_score)
{
min_score = temp_score;
argmin_p = p;
argmin_q = q;
ded = delta_ED(avg_ED_mRNA, avg_ED_ncRNA, i, p, k, j, q, l);
newl = L(i, p, j, q, pi,
raw_mRNA, raw_ncRNA, rawmap_mRNA, leftmap_mRNA, rightmap_mRNA,
rawmap_ncRNA, leftmap_ncRNA, rightmap_ncRNA, r);
oldc = C_ij[p][q];
}
}
}
}
argmin_pq.get<0>() = argmin_p;
argmin_pq.get<1>() = argmin_q;
return min_score;
}
```

D. Data sets of validated interactions

D.1. AE006468

ncRNA	mRNA	acc.id	reference
CyaR	ompX	AE006468	[34]
GcvB	argT	AE006468	[41]
GcvB	dppA	AE006468	[41]
GcvB	gltI	AE006468	[41]
GcvB	livJ	AE006468	[41]
GcvB	livK	AE006468	[41]
GcvB	oppA	AE006468	[41]
GcvB	STM4351	AE006468	[41]
MicA	lamB	AE006468	[6]
MicC	nmpC	AE006468	[35]

D.2. U00096

ncRNA	mRNA	acc.id	reference
CyaR	luxs	U00096	[11]
CyaR	nadE	U00096	[11]
CyaR	ompX	U00096	[11]
CyaR	yqaE	U00096	[11]
DsrA	hns	U00096	[22]
DsrA	rpoS	U00096	[23]
GcvB	sstT	U00096	[36]
GlmZ	glmS	U00096	[45]
MicA	ompA	U00096	[43]
MicC	ompC	U00096	[10]
MicF	ompF	U00096	[38]
OmrA	cirA	U00096	[16]
OmrA	ompR	U00096	[16]
OmrA	ompT	U00096	[16]
OxyS	fhlA	U00096	[2]
RprA	rpoS	U00096	[24]
RyhB	fur	U00096	[46]
RyhB	sodB	U00096	[13]
SgrS	ptsG	U00096	[21]

Bibliography

[1] Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon. Secondary structure prediction of interacting rna molecules. *Journal of Molecular Biology*, 345(5):987 – 1001, 2005.

[2] Liron Argaman and Shoshy Altuvia. fhla repression by oxys rna: kissing complex formation at two sites results in a stable antisense-target rna complex. *Journal of Molecular Biology*, 300(5):1101 – 1112, 2000.

[3] David P. Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[4] Stephan Bernhart, Hakim Tafer, Ulrike Muckstein, Christoph Flamm, Peter Stadler, and Ivo Hofacker. Partition function and base pairing probabilities of rna heterodimers. *Algorithms for Molecular Biology*, 1(1):3, 2006.

[5] Stephan H. F. Bernhart, Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9, 2008.

[6] Lionello Bossi and Nara Figueroa-Bossi. A small rna downregulates lamb maltoporin in salmonella. *Molecular Microbiology*, 65(3):799–810, 2007.

[7] Marie Bouvier, Cynthia M. Sharma, Franziska Mika, Knud H. Nierhaus, and Jörg Vogel. *Molecular cell*, volume 32, chapter Small RNA Binding to 52 mRNA Coding Region Inhibits Translational Initiation, pages 827–837. Cell Press, Dec 2008.

[8] Peter Brodersen and Olivier Voinnet. Revisiting the principles of microrna target recognition and mode of action. *Nat Rev Mol Cell Biol*, 10(2):141–148, Feb 2009.

[9] 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):2849–56, 2008.

[10] Shuo Chen, Aixia Zhang, Lawrence B. Blyn, and Gisela Storz. Micc, a second small-rna regulator of omp protein expression in escherichia coli. *J. Bacteriol.*, 186(20):6689–6697, 2004.

[11] Nicholas De Lay and Susan Gottesman. The crp-activated small noncoding regulatory rna cyar (ryee) links nutritional status to group behavior. *J. Bacteriol.*, 191(2):461–476, 2009.

[12] Kathrin S. Fröhlich and Jörg Vogel. Activation of gene expression by small rna. *Current Opinion in Microbiology*, 12(6):674 – 682, 2009. Growth and development: eukaryotes/prokaryotes.

[13] Thomas A. Geissmann and Daniele Touati. Hfq, a new chaperoning role: binding to messenger rna determines access for small rna regulator. *EMBO J*, 23(2):396–405, Jan 2004.

[14] Walter Gilbert. Origin of life: The rna world. *Nature*, 319(6055):618–618, Feb 1986.

[15] Susan Gottesman. Micros for microbes: non-coding regulatory rnas in bacteria. *Trends in Genetics*, 21(7):399 – 404, 2005.

[16] Maude Guillier and Susan Gottesman. The 5 end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator. *Nucleic Acids Research*, 36(21):6781–6794, 2008.

[17] Gregory J. Hannon. Rna interference. *Nature*, 418(6894):244–251, Jul 2002.

[18] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, Feb 1994.

[19] Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned rna sequences. *Journal of Molecular Biology*, 319(5):1059 – 1066, 2002.

[20] Fenix W. D. Huang, Jing Qin, Christian M. Reidys, and Peter F. Stadler. Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, 25(20):2646–2654, 2009.

[21] Hiroshi Kawamoto, Yukari Koide, Teppei Morita, and Hiroji Aiba. Base-pairing requirement for rna silencing by a bacterial small rna and acceleration of duplex formation by hfq. *Molecular Microbiology*, 61(4):1013–1022, 2006.

[22] Richard A. Lease, Michael E. Cusick, and Marlene Belfort. Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21):12456–12461, 1998.

[23] Nadim Majdalani, Christofer Cunning, Darren Sledjeski, Tom Elliott, and Susan Gottesman. DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21):12462–12467, 1998.

[24] Nadim Majdalani, David Hernandez, and Susan Gottesman. Regulation and mode of action of the second small rna activator of rpos translation, rpra. *Molecular Microbiology*, 46(3):813–826, 2002.

[25] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of Molecular Biology*, 288(5):911 – 940, 1999.

[26] David H Mathews and Douglas H Turner. Prediction of rna secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16(3):270 – 278, 2006. Nucleic acids/Sequences and topology - Anna Marie Pyle and Jonathan Widom/Nick V Grishin and Sarah A Teichmann.

[27] David H. Matthews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.

[28] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[29] Irmtraud M Meyer. Predicting novel rna-rna interactions. *Current Opinion in Structural Biology*, 18(3):387 – 393, 2008. Nucleic acids / Sequences and topology.

[30] Ulrike Mückstein, Hakim Tafer, Stephan H. Bernhart, Maribel Hernandez-Rosales, Jörg Vogel, Peter F. Stadler, and Ivo L. Hofacker. *Translational Control by RNA-RNA Interaction: Improved Computation of RNA-RNA Binding Thermodynamics*, chapter Translational Control by RNA-RNA Interaction: Improved Computation of RNA-RNA Binding Thermodynamics, pages 114–127. Springer Berlin Heidelberg, 2008.

[31] Ulrike Muckstein, Hakim Tafer, Jorg Hackermuller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

[32] Richard P. Novick and Edward Geisinger. Quorum sensing in staphylococci. *Annual Review of Genetics*, 42(1):541–564, 2008.

[33] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[34] Kai Papenfort, Verena Pfeiffer, Sacha Lucchini, Avinash Sonawane, Jay C. D. Hinton, and Jörg Vogel. Systematic deletion of salmonella small rna genes identifies cyar, a conserved crp-dependent riboregulator of ompx synthesis. *Molecular Microbiology*, 68(4):890–906, 2008.

[35] Verena Pfeiffer, Kai Papenfort, Sacha Lucchini, Jay C. D. Hinton, and Jorg Vogel. Coding sequence targeting by micc rna reveals bacterial mrna silencing downstream of translational initiation. *Nat Struct Mol Biol*, 16(8):840–846, Aug 2009.

[36] Sarah C. Pulvermacher, Lorraine T. Stauffer, and George V. Stauffer. The small rna gcvb regulates sstt mrna expression in escherichia coli. *J. Bacteriol.*, 191(1):238–248, 2009.

[37] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[38] Matthew Schmidt, Ping Zheng, and Nicholas Delihas. Secondary structures of escherichia coli antisense micf rna, the 5'-end of the target ompf mrna, and the rna/rna duplex. *Biochemistry*, 34(11):3621–3631, 1995. PMID: 7534474.

[39] Stefan E. Seemann, Jan Gorodkin, and Rolf Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucl. Acids Res.*, 36(20):6355–6362, 2008.

[40] Stefan E. Seemann, Andreas S. Richter, Tanja Gesell, Rolf Backofen, and Jan Gorodkin. PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*.

[41] Cynthia M. Sharma, Fabien Darfeuille, Titia H. Plantinga, and Jörg Vogel. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Development*, 21(21):2804–2817, 2007.

[42] D H Turner, N Sugimoto, and S M Freier. Rna structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):167–192, 1988.

[43] Klas I. Udekwu, Fabien Darfeuille, Jörg Vogel, Johan Reimegård, Erik Holmqvist, and E. Gerhart H. Wagner. Hfq-dependent regulation of ompa synthesis is mediated by an antisense rna. *Genes Development*, 19(19):2355–2366, 2005.

[44] Klas I. Udekwu and E. Gerhart H. Wagner. Sigma E controls biogenesis of the antisense RNA MicA. *Nucl. Acids Res.*, 35(4):1279–1288, 2007.

[45] Johannes H Urban and Jörg Vogel. Two seemingly homologous noncoding rnas act hierarchically to activate <italic>glms</italic> mrna translation. *PLoS Biol*, 6(3):e64, 03 2008.

[46] Branislav Vecerek, Isabella Moll, and Udo Blasi. Control of fur synthesis by the non-coding rna ryhb and iron-responsive decoding. *EMBO J*, 26(4):965–975, Feb 2007.

[47] Lauren S. Waters and Gisela Storz. *Cell*, volume 136, chapter Regulatory RNAs in Bacteria, pages 615–628. Cell Press, Feb 2009.

[48] Ligang Wu and Joel G. Belasco. *Molecular cell*, volume 29, chapter Let Me Count the Ways: Mechanisms of Gene Regulation by miRNAs and siRNAs, pages 1–7. Cell Press, Jan 2008.

[49] Phillip D. Zamore and Benjamin Haley. Ribo-gnome: The Big World of Small RNAs. *Science*, 309(5740):1519–1524, 2005.

[50] Michael Zuker. Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology*, 10(3):303 – 310, 2000.

[51] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

[52] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9(1):133–148, 1981.