Genome scale search of noncoding RNAs Bacteria to Vertebrates

Zizhen Yao

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2008

Program Authorized to Offer Degree:
Department of Computer Science and Engineering

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Zizhen Yao

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Walter L. Ruzzo

Reading Committee:

_____

Walter L. Ruzzo

_____

Martin Tompa

_____

Phil Green

_____

William Noble

Date: _____

Signature_____

Date_____

University of Washington

Abstract

Genome scale search of noncoding RNAs
Bacteria to Vertebrates

Zizhen Yao

Chair of the Supervisory Committee:
Professor Walter L. Ruzzo
Department of Computer Science and Engineering

Traditionally scientists believed that, with a few key exceptions, RNAs played a secondary role in the cell. Recent discoveries have sharply revised this simple picture, revealing widespread and surprisingly sophisticated functional roles of RNAs.

Discovery of new functional RNA elements remains a very challenging task, both computationally and experimentally. It is computationally difficult largely because of the importance of an RNA molecule's 3-D structure, and the fact that molecules with very different nucleotide sequences can fold into the same shape.

In this thesis, we describe a computational tool called CMfinder that addresses the RNA motif discovery problem. It is one of the most effective tools for constructing multiple local structural alignments. It can extract an RNA motif from unaligned sequences with long extraneous flanking regions, and in cases when the motif is only present in a subset of sequences. On the basis of the original CMfinder, we propose several speedup techniques, which make this tool scalable to large datasets.

Another important problem regarding ncRNA discovery is to evaluate the "significance" of a predicted RNA motif, which is critical to sift high quality ncRNA candidates from an enormous number of predictions produced in a genome scale scan. We have designed two ranking schemes to address this problem in different application settings. The first is a heuristic method that is generally applicable, and the second is a probabilistic method

based on the evolution theory. While we have effectively rediscovered known ncRNAs and obtained promising candidates using the first method, we found that the second behaves more robustly and has better statistical properties. The second scheme, however, requires a phylogeny of input sequences, which can be difficult to be obtained in some applications.

We have great success in applying CMfinder in genome scale discovery of noncoding RNAs. In particular, we applied a CMfinder centered computational pipeline to all bacteria, and found 22 novel putative RNA motifs. Six are high quality riboswitches candidates, and five have been confirmed as novel riboswitches in separate studies. We have also tested CMfinder in vertebrate ENCODE regions. This study produced thousands of candidates, most of which are not covered by any previous studies. Closer examination of these candidates suggests that CMfinder revised the alignment significantly compared to the multiple alignment based on the sequence only, and consequently, strongly argues for taking RNA structure directly into account in any searches for such structural elements. We have experimentally validated eleven top ranking candidates, and found transcription activities and tissue specificities for most of them. We are now in the process of applying CMfinder to search the whole human genome.

Our experiences have demonstrated that CMfinder can accelerate significantly the discovery of novel ncRNAs, with promises of many more discoveries to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ARGUMENT: replacement text which customizes a LaTeX macro for each particular usage.

ACKNOWLEDGMENTS

First of all, I want to thank my advisor, Larry Ruzzo, for bearing with me through all these years, for giving me much needed guidance, patience and understanding, and for setting me a role model as a committed scientist. I would also like to thank my committee members, Martin Tompa, Phil Green, Bill Noble and Adrian Raftery, for constructive suggestions and insightful comments, and Joe Felsenstein for numerous informative discussions. I have learned and benefited significantly from my collaborators Zasha Weinberg, Jeffery Barrick, Ronald Breaker, Elfar Torarinsson, Jan Gorodkin, and many others, who helped elevate the impact of my research by pushing its application to the frontier of noncoding RNA research. I am also very grateful to my friends and colleagues at University of Washington, too many to name here, who have given me great research advice and precious friendship. Together we share the "good old days" when we look back a few years from now.

Finally, I would love to thank my parents and my uncle who taught me to love science as a kid, and had faith in me even when I disappointed them. I want to save the last thanks for my dear husband, for his unlimited support and love.

Chapter 1

INTRODUCTION

In this chapter, we will first introduce the basic biological background on RNAs and current status on noncoding RNA research, then present several important computational problems and the methods that address these problems in various phase of noncoding RNAs prediction.

1.1 RNA Background

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic information. DNA can be abstracted as a string from the four letter alphabet A(adenine),C(cytosine), G(guanine) and T(thymine). It usually exists as a tightly-associated pair of molecules in the shape of a double helix. The two strands are held together by hydrogen bonds between the Watson-Crick pairs in which A pairs with T, and G pairs with C.

Ribonucleic acid (RNA) is similar to DNA in that it is also a nucleic acid consisting of nucleotides, with one letter difference from the DNA alphabets: instead of thymine (T), it uses uracil (U). Unlike DNA which is double-stranded, RNA is single stranded but highly structured. The three-dimensional (tertiary) structure of an RNA can be in part captured by its secondary structure, defined by the base pairs held together by hydrogen bonds. Besides the canonical Watson-Crick pairs that are common in DNA double helices, RNAs often use wobble G-U pairs, and other non-canonical base pairs, such as A-C reverse Hoogsteen pair, AC wobble pair, etc.

An RNA secondary structure can be partitioned into domains: A double stranded region formed by base pairs is referred to as a *stem*, or *helix*. The single stranded loops can be further categorized as *hairpin loop* - the end loop that joins two strands of a stem, *bulge loop* - the loop that occurs on one strand of a stem, *interior loop* - the loop that forms in

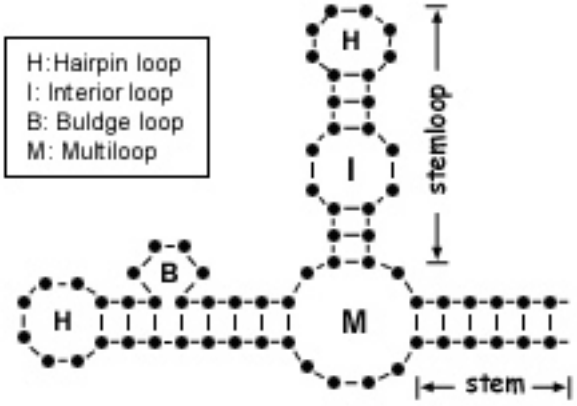


Figure 1.1: RNA Secondary structure domains. Different types of loops are annotated.

the middle of a stem, and *multiloop* - the loop that connects to three or more stems. These domains are illustrated in figure 1.1. A stemloop, also known as hairpin, refers to a single branch structure with only one hairpin loop.

RNAs are usually viewed as tree-like structures, as base pairs are either nested or within disjoint regions. However, this simplified abstraction can be complicated by presence of pseudoknots and base triples. A pseudoknot is formed by the base pairing of the loop region of one stem with the loop region of another stem, while a base-triple consist a base pair interacting with a distant single-stranded base.

RNAs play critical roles in the central dogma in biology: in order to produce proteins that perform many important functions in the cell, DNA genes with the genetic information need to be transcribed into messenger RNAs (mRNAs), which are then translated into proteins by the ribosomes. Besides mRNAs that code for proteins, some RNAs do not code for proteins, and are referred to as noncoding RNAs (ncRNAs). These include classical examples such as ribosomal RNA (rRNA), the catalytic component of the protein synthesis factories - ribosomes, and transfer RNA (tRNA), which transfers a specific amino acid to a growing polypeptide chain in translation.

1.2 Advances in noncoding RNA research

Before the 1990s, RNAs received relatively little attention compared to DNA genes and proteins, because mRNAs are only intermediate and transient products, while very few ncRNAs had been discovered. During the last decade, however, the discoveries of a large number of new ncRNAs suggest that they play much more important, sophisticated and diversified functions than previously thought. Noncoding RNAs have been classified by many different criteria, e.g. by size into small RNAs and long RNAs, by transcription unit into ncRNA genes and cis-regulatory ncRNAs, and by others such as functional roles, regulatory mechanisms, etc. Next, we will introduce briefly a few most well known types of ncRNAs.

Most prevalent ncRNAs in bacteria are the cis-regulatory elements known as riboswitches—parts of mRNA molecules that can directly bind a small target molecule to regulate their own activity (89; 162; 90). A riboswitch typically has two components: an aptamer and an expression platform. The aptamer recognizes and binds to the target molecule, while the expression platform changes structure in response to the binding. This structural change acts like an on/off switch for gene regulation. The first riboswitch was experimentally validated only recently in 2002 (101; 99). While the metabolic pathways that most riboswitches are involved in have been studied for decades, these elements remained undetected for a long time due to the long held assumption that all genes are regulated by proteins. The discoveries of riboswitches add support to the RNA world hypothesis (49), which proposes that life originated from RNA, and proteins came later.

In parallel, a large number of new noncoding RNAs have been discovered in higher organisms. The term microRNA was introduced in 2001 (123), which refers to the short RNAs of 21-23 nucleotides in length and forming partially complementary base-pairs with target messenger RNAs to regulate gene expression. Since the discovery of the first microRNA lin-4 in 1993 (84), thousands more have been found in plants, worms, flies and mammals.

Another interesting group of ncRNAs are small nucleolar RNAs (snoRNAs), small RNA molecules that guide chemical modifications of ribosomal RNAs (rRNAs) and other RNA

genes. Besides miRNAs and snoRNAs, other small RNAs such as small interfering (siRNAs) and piwi interacting RNAs (piRNAs) also exist. A siRNA is a short (usually 21-nt) double-stranded RNA (dsRNA). It can be introduced into cells to target a gene of interest for knockdown, therefore serving as a powerful tool for gene function and drug target validation studies. PiRNAs are found uniquely in mammalian testes and form RNA-protein complexes with Piwi proteins. They have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells.

Long ncRNAs in vertebrates, such as Xist, Avf and Air (31; 39; 135), can be over thousands or hundreds of thousands of bases long. For example, Xist is a major effector of X-inactivation, a process in which one of the two copies of the X chromosome present in female mammals is inactivated. Expression of Xist leads to X chromosome silencing. Structure analysis of these ncRNAs is difficult due to their large sizes. On the other hand, these molecules may contain short functional elements that are more critical than other regions, and it is possible to use computational tools to probe such elements.

A large number of RNA *cis*–regulatory elements exist in vertebrates. Different from DNA *cis*–regulatory elements in the promoters (regions that host functional elements modulating transcription initiation), these elements are located within the $5'$ or $3'$ untranslated region (UTR) of a transcript, and many of them are involved in regulation of transcription, translation, localization, degradation, modification and replication (5; 21; 65; 50; 86). For example, the Iron-responsive element (IRE) is a short stem-loop bound by iron response proteins (IRPs), and found in UTRs of many genes involved in iron metabolism. IRE has two distinctive functional roles: IRPs binding to the ferritin mRNA IRE leads to translation repression when iron concentration is low; on the other hand, IRPs binding to the transferrin receptor IRE leads to increased mRNA stability. The structures of these RNA *cis*–regulatory elements are typically functionally important, as disruption of these structures usually leads to loss or deterioration of function.

Recent genome wide tiling microarray studies suggest that at least half of the mammalian transcripts do not code for proteins (19; 17). Yet, the functions of many of these newly identified ncRNAs are still unknown. Understanding the functions of ncRNAs is a critical component in deciphering genome structure, one of the most challenging tasks of modern

biology.

1.3 Computational advances in ncRNA prediction

In this section, we focus on the most fundamental theories and methods for ncRNA prediction that make many application tools such as CMfinder possible.

RNA secondary structure prediction

The secondary structure of RNA molecules can be predicted computationally. Fresco, Alberts and Doty (42) first approached this problem by maximizing base pairing for optimal secondary structure. Ruth Nussinov and colleagues (107) later proposed a dynamic programming algorithm to compute the structure with maximum number of base pairs. Tinoco *et al.* (141) suggested that the secondary structure of an RNA molecule at equilibrium should minimize its free energy, and proposed a very simple model to estimate the free energy of a given secondary structure. More sophisticated energy models and the corresponding secondary structure prediction algorithms have been proposed (106; 172). These models follow the *nearest neighbor* energy rules, i.e. free energies are assigned to loops rather than base pairs. Based on these rules, one can decompose a secondary structure uniquely into loops (two adjacent base pairs also form a loop), and compute the total free energy as the sum of the energies of each loop. Pseudoknots and base triples are ignored, which otherwise break down the loop decomposition. Using a dynamic algorithm that is in principle highly similar to the one that maximizes the number of base pairs, one can compute the optimal secondary structure with the minimal free energy (MFE) with time complexity of $O(n^3)$, where n is the sequence length (106; 172; 88). Some prediction methods take pseudoknots into account (118), but they are less commonly used due to their computational expense. The parameters of the energy model can be derived from calorimetric experiments, and they have been revised numerous times (40; 96; 95; 130).

An RNA molecule can switch between a set of alternative structures, and the MFE structure does not necessarily approximate well the native structure. To address this issue, Michael Zuker proposed an algorithm that finds distinctive suboptimal foldings of an RNA (170). In this line, McCaskill and colleagues suggested a method to calculate the full

equilibrium partition function, which specifies the probabilities of all possible substructures (97).

Many important algorithms for energy-based secondary structure prediction have been implemented in Mfold (171) and the Vienna software package (68). Both are widely used for routine RNA analysis.

1.3.1 ncRNA probabilistic modeling

Due to the stochastic natures of ncRNA structure folding and ncRNA changes during evolution, we need probabilistic models to capture the inherent variability on sequences and structures.

SCFG

Besides the energy model described above, RNA secondary structure can also be modeled by stochastic context-free grammars (SCFGs), a probabilistic augmentation of context-free grammar (CFG) (134). I will first give a brief introduction of context-free grammars.

Based on the formal language theory, a computer language can be described by a grammar which consists of terminal symbols Σ, non-terminal symbols V and a set of production rules R for transforming a string of symbols to another. From the start symbol S, one can enumerate all words of the language by repeatedly applying the production rules until all letters in the string are terminal symbols. On the other hand, given a string, one can determine whether it belongs to the language based on the grammar using a parsing algorithm. Different families of grammars have been defined based on different constraints over the production rules, and corresponding parsing algorithms have different complexities. In particular, for a context-free grammar (CFG), every production rule has the form

$$A \rightarrow w$$

in which A is a non-terminal symbol and w is a string of symbols. A simple example of CFG is given below:

$$S \rightarrow (S)|SS|\epsilon$$

where | is used to delimit alternatives, and ϵ denotes the empty string. This grammar generates a string with balanced and nested parentheses, and here is an example of derivation:

$$S \rightarrow (S) \rightarrow (SS) \rightarrow ((S)(S)) \rightarrow (()())$$

The Cocke-Younger-Kasami (CYK) algorithm (134) can determine whether a string can be generated by a CFG. This algorithm has time complexity of $O(n^3)$, where n is the string length.

CFGs have wide applications in programming language design, compilers and natural language processing. Note that the RNA secondary structure prediction problem draws analogy to the balanced parentheses example above, as base pairs can be modeled as paired terminals similar to a pair of parentheses. In fact, one can define a CFG for all pseudoknot-free RNA structures using the following simple CFG:

$$S \rightarrow aS\hat{a}|aS|Sa|SS|\epsilon$$

where a represents a single nucleotide, and $aS\hat{a}$ implies that a pairs with $\hat{a}$. A parse of a sequence based on this grammar determines uniquely its secondary structure. On the other hand many different parses of a sequence exist, so how can we decide which is the most probable one? This is a problem that can be addressed by SCFGs. An SCFG is just like a CFG, except that each of its production rules has a probability. The probability of a parse is the product of the probabilities of all the production rules used in the derivation. In the context of RNA secondary structure prediction, the structural parse with the highest probability can be computed by a variant of the CYK algorithm (30), and is usually selected as the optimal solution. Alternatively, one can use the so-called inside-outside algorithm (30) to compute the overall probability of a sequence by summing over all possible structures, and compute the probability of a given base pair.

In some sense, SCFGs are analogous to energy models, both of which specify the rules that govern RNA folding. On the other hand, the parameters of SCFGs can be estimated computationally from known RNAs. SCFGs also have a potential advantage that they can be readily extended to incorporate additional sources of information, such as evolutionary history. Dowell and Eddy compared several SCFGS to evaluate the tradeoff between model complexity and prediction performance (28).

Covariance model

The SCFG is used to describe the general secondary structure folding principle. It can be used to evaluate the likelihood of an arbitrary structure. The covariance model (CM) on the other hand, was introduced by Eddy and Durbin in 1994 (35) to capture both the consensus and variability of a particular RNA family. Here, an RNA family refers to a collection of RNAs that are highly similar in structure and function. They also tend to share certain sequence similarity, usually localized, while base changes are common. Covariance models, in some sense, are analogous to profile Hidden Markov models (HMMs), which have been used to model the protein families. Similar to a profile HMM, the skeleton of the CM states specifies the consensus structure, while one state models the variability at a given position. A CM is parameterized by two types of probabilities: transition probabilities define the likelihood of transition between states (insertions and deletions are allowed), while emission probabilities define the likelihood of the observed bases at the corresponding states. The complexity of an ncRNA secondary structure requires covariance models to have extra states to model its nested tree-like structure. These include the "pair" states for modeling the base pairs in the structure, and "bifurcation" states, for modeling junctions of two branches in the structure. In particular, the pair states capture the covariation at the base paired positions, i.e. when a mutation occurs within a base pair during evolution, it tends to preserve the base pair, e.g. change from GC pair to AU pair, to maintain the same the secondary structure. Such covariation is considered as strong evidence of purifying selection on structure, as it suggests that while mutations are allowed, it is important that they conform to the functional structure of the ncRNA molecule.

Like Hidden Markov models, covariance models have a collection of algorithms for routine probabilistic analysis. Given a sequence, one can compute its most likely alignment to the given CM by the Viterbi algorithm. This alignment determines the closest mapping of the sequence to the consensus structure specified by the CM. Alternatively, we can also compute the probability of the sequence by summing all possible alignments by the inside-outside algorithm, and the posterior probability of a given position (or a pair of positions) mapping to a given state. The CM is usually constructed automatically based on an an-

notated structural alignment, using the maximum likelihood approach. This approach has been extended by the informative prior modeled as Dirichlet mixtures under a Bayesian framework (103), making it more adaptive to sequences that are dissimilar to the training data. These methods have been implemented in the software package COVE (35) and its successor Infernal (34).

COVE and Infernal have driven Rfam (58), one of the largest collections of ncRNA families. For each family, it contains a hand-curated structural alignment of trusted family members (referred to as the seed alignment), and the corresponding CM. The CMs are used to scan genome databases on a regular basis to find potential new members of the families. The Rfam database has been extensively used for ncRNA annotation in new genomic sequences.

1.3.2 ncRNA gene prediction

While most ncRNAs known currently are discovered by experimental methods, there are a few technical challenges that limit applications of these methods. NcRNAs are typically too small for genetic screens, and many are not polyadenylated, which makes them poor targets for polyA-selected cDNA-based expression studies. To further complicate the situation, many noncoding transcripts identified by expression platform such as microarrays may simply be transcription noise. Some transcripts have such low expression levels that they are difficult to be further validated by RT-PCR or northern blot.

Computational prediction of ncRNA gene based genomic sequences can potentially address the experimental limitations outlined above. Maizel et al. suggested using secondary structure for RNA gene detection (82), based on the observation that known ncRNAs have more stable secondary structures than expected by chance. Rivas and Eddy pointed out that this statistical effect is largely due to composition bias, and concluded that secondary structure alone is generally not statistically significant for ncRNA detection (119). This observation was then later revised by Peter Clote and colleagues, who pointed out that structural RNAs usually do have lower folding energies than random RNAs of the same dinucleotide frequencies (20), although the signals are still not statistically significant for

ncRNA detection.

Due to limited predictive power from a single sequence, continuing attempts for ncRNA prediction have been focused on comparative genomic analysis.

Methods that align then fold

Several methods have been proposed to predict secondary structure from alignments of RNA homologs. The input alignments are constructed based on sequences only, using tools such as CLUSTALW (139) and MULTIZ (the alignment tool used by the UCSC Genome Browser) (12). The most important information used for consensus structure prediction includes structure stability and covariation. As we mentioned previously, covariation is perhaps the strongest evidence of functional ncRNAs aside from experiments. Different methods use different measures of structure stability and covariation. In particular, RNAalifold (67), a tool implemented in the Vienna package, computes the minimum free energy (MFE) of all substructures, averaged over all sequences. It also measures covariation using a variant of mutual information (MI) (61), modified to favor canonical base pairs and penalize non-canonical base pairs. These two terms are then combined linearly, and used in a dynamic programming algorithm to search for the optimal structure. Pfold (80; 81) on the hand, uses SCFGs to capture structure stability, and uses an evolutionary model to capture the covariation. Compared to mutual information, which ignores the phylogeny of sequences, the evolutionary model examines whether a base pair is preserved at each branch of the tree and measures the corresponding likelihood. The two methods perform quite similarly in practice.

Methods that fold then align

An alternative method is to first fold the sequences, then align them. Shapiro (131) proposed a method to compare two RNA structures using a tree alignment model. An implementation of this method is included in the Vienna package, and it has been extended for multiple sequence comparison in a tool called RNAforester, which is also incorporated in the Vienna package. The key drawback of this method is that the secondary prediction

of a single sequence is usually unreliable, especially if the sequences are long. The situation gets worse if only parts of the input sequences contain real ncRNAs. Consequently, such methods are not frequently used in practice.

Methods that align and fold simultaneously

Alignment-based structure prediction methods are subject to alignment errors, therefore are not appropriate for sequences with limited sequence conservation. The seminal approach of Sankoff (124) performs simultaneous alignment and structure inference. This method in principle exhausts all possible structural conformations and all possible alignments using a dynamic programming algorithm. However, this approach is notoriously expensive, with time complexity $O(n^6)$ and space complexity $O(n^4)$ for only a pairwise alignment. Various restrictive approximations have been developed, including FOLDALIGN (64), Dynalign (94), Stemloc (70) and Consan (27), all attempting to increase performance while sacrificing some accuracy, but even these procedures remain computationally expensive.

There are also a collection of graph-theory based methods, such as Carnac (144), ComRNA (71) and RNAmine (62). Many methods in this set are typically efficient for small size datasets, but in the worst case have exponential time complexity.

1.3.3 Scoring ncRNA predictions

An important problem in genome scale ncRNA discovery is to evaluate the qualities of the predictions. Although most of the ncRNA prediction methods provide some metrics as optimization targets, these metrics usually measure sequence or structural homology, which are inherently dependent on features such as alignment length, sequence similarity etc. Therefore, such metrics are not appropriate for cross comparison of predictions with different alignment features. Several tools have been designed for scoring ncRNAs. The basic principle is to evaluate the amount of structural conservation among species, calibrated based on different alignment features. In particular, QRNA (120) has been used to scan pairwise alignment for ncRNAs with different evolutionary signatures from protein coding regions and background. Evofold (109), another phylo-SCFG based method somewhat

similar to Pfold, computes the likelihood ratio of the structural model vs. the background model. AlifoldZ (148) and RNAz (149) compute a folding energy z-score for each sequence, normalized based on appropriate background distribution. RNAz use this z-score and other features for the support vector machine (SVM) classification. These methods will be further discussed in Chapter 4.

1.3.4 ncRNA homolog search

Whenever a new ncRNA candidate has been found, a natural next step is to locate all existing copies, including homologs in other species. Structural conservation is a good indicator that the structure is functional. In addition, the folding constraints from different homologs tend to help clarify the ambiguities among a set of possible structures, leading to higher accuracy in folding prediction. In the context of the covariance model, more homologs usually provide better model of the family. As we will discuss in Chapter 5, more homologs also tend to give more clues for the functional roles of the ncRNA candidate. Traditionally, ncRNA homolog search has been performed by two popular approach: BLAST search based on sequence only, and model-based search (e.g. CM search) using both the structure and sequence. However, these two methods are at the two extremes of the spectrum: BLAST is fast, but fails to capture structure, so it tends to miss the homologs with strong structure but weak sequence conservation; CM search, on the other hand, is accurate but slow. Routine Rfam search using CMs could take over thousands of CPU years. To speed up CM search, Weinberg and Ruzzo (155) suggest building a rigorous HMM filter for a given CM, such that all candidates that will be accepted by the CM will pass the filter. This methods improves the efficiency significantly (typically 100 fold speedup) without loss of accuracy. Weinberg and Ruzzo proposed additional speedup techniques and further extension by relaxing the rigorous condition while still approximating the CM as closely as possible (153; 157) Besides CM, there is another statistical profile method called ERPIN (46), with different modeling constraints. There are other homolog search methods based on pattern search (59; 29). These methods however, are usually binary and with no statistical support, making it difficult to adjust sensitivity and specificity.

1.4 Large scale ncRNA database

Several large scale ncRNA databases have been constructed to curate the research development on ncRNA studies, and to provide easy and comprehensive access to the community. We have previously described the Rfam database. The latest version is Rfam 8.1, with 607 families. The main advantage of the Rfam database is that all its families have resolved consensus secondary structure, and its advanced computational ncRNA scanning function facilitates ncRNA annotation of new genome sequences. Besides Rfam, there are several general ncRNA databases, which include NONCODE (`http://noncode.bioinfo.org.cn` (85)), the *functional RNA database* (fRNAdb, `http://www.ncrna.org/frnadb/`) (77), and the *mammalian noncoding RNA database* (RNAdb, `http://research.imb.uq.edu.au/rnadb/`) (108). There are also specialized database. Some are designated for house keeping RNAs, such as the *European ribosomal RNA database* (`http://bioinformatics.psb.ugent.be/webtools/rRNA/`), the *Genomic tRNA Database* (GtRNAdb, `http://lowelab.ucsc.edu/GtRNAdb/`), the *RNase P Database* (`http://www.mbio.ncsu.edu/RnaseP/`), and the *Signal recognition particle database* (SRPdb, `http://rnp.uthct.edu/rnp/SRPDB/SRPDB.html`). Recently, several databases have been constructed for microRNAs and snoRNAs in response to the intense research efforts and dramatic progress in these fields. One of the most popular ones is the *microRNA database* (miRBase, `http://microrna.sanger.ac.uk/`) (56; 54; 55), which integrates information for both microRNA sequences and targets. The databases for snoRNAs include snoRNAbase (`http://www-snorna.biotoul.fr/`) for human, the plant snoRNA database (`http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home`), and the yeast snoRNA database (`http://people.biochem.umass.edu/fournierlab/snornadb/main.php`). A significant portion of the information provided by these databases has been integrated into the UCSC genome browser or Genbank, to provide a unified view of functional elements.

1.5 Contributions of this thesis

In this thesis, we will address several key issues in ncRNA discovery.

First, we target the problem of the RNA motif discovery. The RNA motif here refers to

a structural alignment of a set of ncRNA homologs. As we have introduced in the previous section, many existing methods are alignment-based methods. Although these methods use covariation for consensus structure prediction, the sequence alignments are unaware of the structure and consequently, tend to doubly penalize the covariation and misalign the corresponding positions (147). Most methods that perform structural alignment are either limited to pairwise alignment only, or only perform global alignment. We tackled a more challenging and more realistic problem, which is to construct a multiple, local structural alignment given a set of sequences, each ranging from hundreds to thousands of bases long. Our method, referred to as CMfinder, has been demonstrated to be efficient, robust and reliable, and capable of detecting real ncRNA signals hidden within large genomic datasets. This method is practical for most applications, and can be used with little precondition.

A genome scale ncRNA discovery study is bound to produce an enormous number of predictions. For the follow-up study, it is critical to have an evaluation system that put the most promising candidates on the top of the prediction list. We have studied a couple of existing methods, and found that they are not robust when applied to motifs with weak sequence conservation, due to their hidden assumption about the alignment quality. We proposed two scoring methods to address this issue; one is a heuristic ranking function and another is a probabilistic model-based method. Both methods work satisfactorily in appropriate settings. The model-based method has more theoretical support and better statistical behavior, but requires an evolutionary tree of the aligned sequences.

We have applied the RNA motif finding tool and the scoring tool in two large genome scale studies. In the first study, we attempted to discover *cis*–regulatory elements in bacteria. We have designed a computational pipeline for high-throughput ncRNA screening. A set of computational tools have been carefully selected and integrated to optimize the overall effectiveness. Prototype experiments in Firmicutes demonstrated that this pipeline has been highly efficient and accurate to recover most of the known *cis*–regulatory ncRNA elements and ncRNA genes in this group. Application of this pipeline to all bacterial clades produced a large number of high quality predictions, including 22 candidates that were selected based on a set of very stringent criteria. Among selected candidates, six were hypothesized to be riboswitches, and to this date, five have been experimentally validated.

There have been continuing efforts in mining this set of predictions, possibly still enriched with hidden treasures.

In the second study, we have applied CMfinder to ncRNA discovery in vertebrates. The pilot study in the ENCODE regions produced thousands of predictions. Among eleven top ranking candidates that were tested by RT-PCR, ten were confirmed to be present as RNA transcripts in human tissue. In addition, we found that about one quarter of our predicted motifs show revisions in more than 50% of their aligned positions, and most of our predictions do not overlap with previous studies. Our results suggest caution in any RNA structural analysis relying on multiple sequence alignments, and strongly argue for considering RNA structure directly in any searches for these elements.

In conclusion, our efforts have led to great advance in ncRNA prediction, by providing a set of convenient tools that eliminate a significant amount of manual work and speed up dramatically the discovery process. The key merit of our system is its flexibility and robustness at handling datasets with different sequence conservation, and its efficiency at handling large datasets. Considering that most known ncRNAs discovered currently via computational methods are within the regions with significant sequence conservation, our system offers exciting opportunities to venture into previously unexplored territories, possibly expanding our knowledge of ncRNAs in general. The effectiveness and great potential of this system have been demonstrated by our extensive search of ncRNAs in bacteria and the pilot study within the vertebrates genomes.

1.6 Thesis outline

In Chapter 2, I will first present one RNA motif finding tool - CMfinder. In Chapter 3, I will discuss several speedup techniques for CMfinder to deal with large datasets. Significance tests for evaluating CMfinder motifs are discussed in in Chapter 4. Finally, applications of CMfinder to genome wide ncRNA discovery for bacteria and vertebrates are presented in Chapters 5 and 6 respectively.

Chapter 2

CMFINDER: AN RNA MOTIF PREDICTION ALGORITHM

2.1 Introduction

We are interested in the problem of identifying conserved secondary structure motifs among related sequences, and characterizing them by models that can be used for homology search. For example, identification of such motifs in untranslated regions of orthologous bacterial genes has been critical to the discovery of new riboswitches (8), but available techniques require significant manual work.

Comparative sequence analysis is generally recognized as the most reliable method of RNA structure prediction, but these methods can fail when sequence conservation is too low (due to poor alignments) or too high (due to lack of sequence covariation). Structure prediction for single sequence is inaccurate, and simultaneous multiple sequence alignment and folding is computationally expensive. Finally, these tools generally do not interface smoothly with RNA homology search tools.

We present a new algorithm for solving this motif discovery problem. Oversimplifying considerably, it is an expectation maximization (EM) algorithm like MEME (6), but instead of weight matrix models, it captures RNA secondary structures with covariance models (35). Because of the greatly increased complexity of the problem, we applied several techniques to solve scalability and convergence issues. These include use of careful heuristics for choosing a set of candidate structure elements to initialize the EM iteration. To improve the accuracy of consensus secondary structure in the M-step, we have combined mutual information with data-dependent, position-specific priors for base pairing based on a thermodynamic model. The key merits of our solution include:

- Applicable to unaligned input with unrelated sequences, long flanking regions and/or low sequence similarity;

- Reasonably fast and scalable with respect to the number and length of input sequences;

- Producing a motif structural alignment and statistical model that can be directly used for homology search.

The third point is particularly important, because we view this tool as only one component of a discovery pipeline wherein a motif model built from a small dataset will be used to find more instances, allowing the model to be extended and refined iteratively.

Extensive testing demonstrates that these goals are largely met. For most tests, the hand-curated "seed" alignments from Rfam are our gold standard. CMfinder achieved better results than other methods in 17 of 19 tests, and predicted base pairs with average 77% sensitivity, 81% specificity and 79% accuracy, compared to at most 60% accuracy for the other methods. Most disagreements with Rfam are local perturbations such as small shifts or extra base pairs. The results presented in this chapter have been reported in (167).

2.2 Methods

The basic idea of CMfinder is to use a Covariance Model (CM) to model an RNA motif, a finite mixture model to describe motif distribution in sequences, and an EM framework to search the motif space. It is motivated by two previous techniques: the DNA motif finding program MEME (6), and the Covariance Model based RNA analysis tool COVE (35) and its successor Infernal (33). The combination of these two methods is non-trivial due to the increased complexity: in MEME, the motif model is an ungapped weight matrix with a relatively short, fixed length window, while RNA motifs are generally much longer with significant secondary structures and frequent insertions/deletions. COVE/Infernal assumes that each sequence is an instance of the model, and performs global alignment rather than the more difficult local alignment we need. In the EM framework, the expanded search space and higher model complexity suggest that it is infeasible to explore all subsequences as MEME does, and having a good starting point is critical. To address these issues, CMfinder chooses motif candidates with potentially stable secondary structures, selects a conserved set across all sequences, and aligns them heuristically. This step is loosely similar

to Carnac (144) and ComRNA (71). The subsequent EM iteration refines the model and alignment. The following section elaborates this idea.

2.2.1 Construction of heuristic initial alignment

The goal of this step is to identify the approximate location and structure of the motif efficiently. The key design issue is the tradeoff between accuracy and efficiency. As the motif will be refined later, we can tolerate alignment errors provided correctly aligned instances are well-represented, but require robustness to varieties of dataset sizes, sequence similarities and structures.

Candidate Selection

We first eliminate a large portion of the search space by focusing on strong candidates, i.e., segments with potentially stable secondary structure. We used RNAfold in the Vienna package (68) to compute the minimal free energy for all subsequences in a given sequence. The motifs whose length and number of stem-loops are within the range (default 30-100 bases, 1-2 stem loops), and are locally optimal (base paired at the ends, with no lower-energy states by extending or shrinking 2 bases at the ends) are selected. They are then sorted by the energy, scaled by sequence length. Candidates are selected from the top of the list iteratively. We allow overlapped candidates as long as one of them is significantly longer than the other one. Carnac (144) and ComRNA (71) use similar ideas to choose simple stems as candidates, but we allow a greater flexibility in candidate selection, which provides better robustness and ultimately leads to better performance.

Candidate Comparison and Alignment

To find the consensus alignment of the candidates, our next step is to compare their predicted secondary structures. We use the tree-edit algorithm (131; 68), modified to compare candidates at the single base or base pair level so that the comparison is sensitive to both sequence and structure. This improves its ability to distinguish between RNAs with relatively

simple structures. We initially used a primitive edit distance matrix defined below.

$$R_{a,b} = \begin{cases} 0 & \text{if a=b} \\ 2 & \text{if } a \neq b \text{ and both a and b are single bases} \\ 1 & \text{if } a \neq b \text{ and both a and b are pairs} \\ \infty & \text{else} \end{cases} \tag{2.1}$$

This edit distance matrix forbids single bases to be aligned with base pairs, therefore, penalizes the alignment heavily for incorrectly predicted base pairs. This matrix has been replaced with a more elaborate edit distance model (78) in chapter 3. The edit distance of two RNAs is scaled by the square root of the product of their lengths to weight comparisons of RNAs with different sizes properly.

Given the distances between all pairs of candidates, we would like to construct the initial alignment by choosing one candidate from each sequence so as to minimize the sum of pairwise scores. However, this is an NP-hard problem. We use the following heuristics to find an approximate solution: First, for each sequence i and each of its candidates c_{ij}, we locate its closest match $m_{ijk} = c_{kl}$ in sequence k, where $l = argmin_{l'} dist(c_{ij}, c_{kl'})$. Then for c_{ij}, we compute the sum of distances to its closest matches in all sequences $d_{ij} = \sum_{k \neq i} dist(c_{ij}, m_{ijk})$. The candidate with the minimal distance d_{ij} over all choices of i and j is chosen as the "consensus candidate". Then we iteratively and greedily choose one candidate from all the remaining sequences such that its sum of distances to the previously chosen candidates is minimized. If this distance is above some threshold, i.e., the remaining candidates are all significantly different from the chosen ones, and the number of the chosen ones is over $n \cdot f$, where n is the total number of sequences and f is the fraction of the sequences expected to contain the motif instances, this process terminates. As some sequences may not contain the motif at all, this technique prevents contamination by unlikely candidates. On the other hand, we require at least $n \cdot f$ candidates to be included to prevent model overfitting. Finally, we construct the initial alignment based on pairwise tree-alignment between each chosen candidate and the consensus candidate.

After the first initial alignment is determined, we select the next "consensus candidate" from the unchosen candidates, and build the alignment in the same manner. Other members

except the consensus candidates are allowed to appear in multiple initial alignments in case they are assigned to the wrong alignments initially. However, we forbid significant overlap between two initial alignments (over 40% candidates in common) as they tend to converge to the same solution. Each of these alignments will be used to initialize the EM algorithm described in Section 2.2.2.

The complexity of candidate comparison algorithm is approximately quadratic in the length of the candidates, thus far more efficient than the Sankoff-style algorithms (124; 51; 94), because we fix the local structures. On the other hand, errors of secondary structure prediction and the simplified edit distance model may still compromise its performance, although local structure predictions on short subsequences are generally quite reliable. Despite the efficiency of the tree-edit algorithm, pairwise comparisons of numerous candidates can be expensive. To improve efficiency further, we only align two candidates if they are compatible with locally conserved regions of the corresponding pair of sequences found by BLAST search. This heuristic also improves accuracy by preventing obvious misalignments. CARNAC and comRNA rely on similar anchor-based techniques to reduce their otherwise prohibitive computational cost; it is less critical in our case, but still valuable.

2.2.2 Refining Alignments via CM-based EM

To improve the quality of the initial alignments, we adopt an iterative EM-like algorithm based on covariance models. As described in chapter 1, the covariance model (35) is a probabilistic model for RNA families that cleanly describes both the secondary structure and the primary sequence consensus. We apply a finite mixture model to describe a sequence as a mixture of regions that follow the background distribution and regions that follow the motif covariance model, and then use the EM algorithm to estimate the model and the hidden variables representing motif instances simultaneously. For clarity, we first assume that motif instances only occur among the candidates as defined in section 2.2.1, a restriction we relax (see "Adjusting candidates"). The following notations are used:

M: the motif CM.

B: the background distribution.

$\Gamma = (M, B, \gamma)$: the finite mixture model,

where γ is the mixture probability that a sequence contains a motif.

N: the total number of sequences.

$S = (s_i)_{1 \leq i \leq N}$: the input sequences.

m: the maximum number of candidates allowed in each sequence.

$C_i = (c_{ij})_{1 \leq j \leq m}$: the candidate set of sequence s_i.

$\Pi_i = (\pi_{ij})$: the alignments of candidates in C_i with M.

$X_i = (x_{ij})$: hidden variables representing the occurrence of the motif in C_i

($x_{ij} = 1$ if c_{ij} is a motif instance, and $x_{ij} = 0$ otherwise).

$D = (L_1, L_2, ..., L_l)$: the sequence alignment used in the M-step. L_i: a column.

$\sigma = (\alpha, \beta)$: consensus secondary structure for D.

α: a set of (indices of) single-stranded columns.

β: a set of (pairs of indices of) base paired columns.

The aim is to find $\Gamma = (M, B, \gamma)$ that maximizes the log likelihood

$$\log P(S|\Gamma) = \log \prod_i \sum_{X_i} \sum_{\Pi_i} P(s_i, X_i, \Pi_i|\Gamma)$$

The E-step estimates hidden variables X_i and Π_i, and the M-step updates M and γ. Note that our framework differs significantly from the inside-outside algorithm which estimates the CM parameters from unaligned sequences. Although both are EM algorithm using unaligned sequences as input, the inside-outside algorithm assumes that the structure of the model is known and fixed, and performs global alignment. Essentially, it calculates the probability that a state/transition is used by summing over all possible parses involving the given state/transition, and uses such probability estimates to update CM parameters. In our framework, we need to determine where the motifs are located, and therefore, introduce an addition hidden variable X_i. In addition, due to relative poor quality of the initial model, the CM structure needs to be refined during iteration. Thus, the following algorithm description will focus on these two aspects. The EM algorithm is initialized by the alignments produced in the heuristic step, which is followed by the M step, then E step. The major CM functions

(e.g., alignment, scanning, etc) are adopted from (35), and will not be explained here.

The E-Step

For each candidate c_{ij} of sequence s_i, we need to estimate two hidden variables: alignment π_{ij}, and motif assignment x_{ij}. The motif assignments are estimated using a finite mixture model. Assuming zero or one motif occurrence per sequence, the probabilities that a motif candidate is an instance of a motif can be computed as:

$$\begin{aligned} P(x_{ij}=1|\Gamma) &= \frac{f_{ij}}{f_{i0}+\sum_{k=1}^{m} f_{i,k}} \\ f_{i0} &= (1-\gamma)P(s_i|X_i=(0,0,\ldots,0),\Gamma) \\ f_{ij} &= \lambda P(s_i|x_{ij}=1,\Gamma) \end{aligned} \tag{2.2}$$

where $\lambda = \frac{\gamma}{m}$. Let $\widetilde{c_{ij}}$ be the region in sequence i excluding c_{ij}, then

$$P(s_i|x_{ij}=1,\Gamma) = P(c_{ij}|M)P(\widetilde{c_{ij}}|B)$$

$$P(s_i|X_i=(0,0,\ldots,0),\Gamma) = P(c_{ij}|B)P(\widetilde{c_{ij}}|B)$$

Equation 2.2 can be rewritten as

$$\begin{aligned} P(x_{ij}=1|\Gamma) &= \frac{\lambda P(c_{ij}|M)P(\widetilde{c_{ij}}|B)}{(1-\gamma)P(s_i|X_i=(0,0,\ldots,0),\Gamma)+\sum_{k=1}^{m}\lambda P(c_{ik}|M)P(\widetilde{c_{i,k}}|B)} \\ &= \frac{\lambda P(c_{ij}|M)/P(c_{ij}|B)}{1-\gamma+\sum_{k=1}^{m}\lambda P(c_{i,k}|M)/P(c_{i,k}|B)} \end{aligned} \tag{2.3}$$

Ideally, we should compute $P(c_{ij}|M)$ by the inside algorithm. The inside algorithm computes the probability of a sequence given a covariance model by summing over probabilities of all possible alignment paths. It recursively fills a three-dimensional dynamic programming matrix with values $\alpha_v(i,j)$, where $\alpha_v(i,j)$ is the summed probabilities of all parse subtrees rooted at state v for subsequence $x_i,\cdots,x_j$. However, a specific alignment is needed to infer the covariance model in the M-step. Thus, we use the Viterbi algorithm (also known as the CYK algorithm in the context of covariance models) to compute $P(\pi_{ij}|M)$, the probability for the optimal alignment path given the covariance model, as an approximation to $P(c_{ij}|M)$. Similar to the inside algorithm, the Viterbi algorithm is a dynamic programming algorithm that fills a three-dimensional table, but instead of taking the sum of probabilities

for all parse subtrees, it computes the maximum probability of all subtrees. The resulting probability $P(x_{ij} = 1)$ can be interpreted as the probability for c_{ij} to be a motif instance given alignment π_{ij}. Although this approximation tends to underestimate $P(x_{ij} = 1)$, most real motif instances are distinguished from the background so dramatically that this approximation is sufficient.

The M-step

Here, we update M and γ.

To update M, we first need to adjust the structure of M, then infer the transition and emission probabilities. Given the structure, the second issue can be solved using a Bayesian posterior estimate with Dirichlet prior (35), hence, we focus on the first issue. This is easy in MEME because the structure is predetermined. In CMfinder, it is equivalent to finding a consensus secondary structure. We formulate this in the following Bayesian framework.

Our goal is to find $\hat{\sigma} = argmax_{\sigma} P(D, \sigma)$. Assuming independence of non-base paired columns, then

$$P(D|\sigma) = \prod_{k \in \alpha} P(L_k) \prod_{(i,j) \in \beta} P(L_i, L_j) \tag{2.4}$$

$$= \prod_{1 \leq k \leq l} P(L_k) \prod_{(i,j) \in \beta} \frac{P(L_i, L_j)}{P(L_i)P(L_j)} \tag{2.5}$$

$$\text{Let } I_{ij} = \log \frac{P(L_i, L_j)}{P(L_i)P(L_j)}$$

Using maximum likelihood parameter estimates and a multinomial model for each column/column pair, I_{ij} is the mutual information between columns i and j. Without prior information, the optimal structure maximizes $\sum_{(i,j) \in \beta} I_{ij}$, which is the approach adopted by COVE. This method works well in large, phylogenetically diverse datasets, but not when there is insufficient covariance, as would be expected with a few closely related sequences.

We solve the problem by introducing an informative prior on structures. Let q_i be the prior for column i to be single stranded, and p_{ij} the prior for columns i, j to be base paired, then $P(\sigma) = \prod_{k \in \alpha} q_k \prod_{(i,j) \in \beta} p_{ij}$, and $P(D, \sigma)$ can be rewritten as

$$P(D, \sigma) = P(D|\sigma)P(\sigma)$$

$$= \prod_{1 \le k \le l} P(L_k) q_k \prod_{(i,j) \in \beta} \frac{P(L_i, L_j)}{P(L_i) P(L_j)} \frac{p_{ij}}{q_i q_j} \tag{2.6}$$

$$\text{Let } K_{ij} = \log \left(\frac{P(L_i, L_j)}{P(L_i) P(L_j)} \frac{p_{ij}}{q_i q_j} \right) = I_{ij} + \log \frac{p_{ij}}{q_i q_j},$$

then the maximum likelihood structure σ maximizes $\sum_{(i,j) \in \beta} K_{ij}$. We infer a prior on structures based on a thermodynamic model. For each sequence, we calculate the partition function P_{ij} (68; 97), which estimates the probability of forming base pair i, j, averaged over all possible structures. We estimate the column pairing probabilities p_{ij} by averaging the partition functions of the aligned sequences. The probability that a column is unpaired is estimated as $q_i = 1 - \sum_j p_{ij}$. Note that candidates are weighted based on their probabilities to be motif instances when computing I_{ij}, p_{ij} and q_i. In the first iteration, all candidates in the intial alignment are weighted equally. Finally, we use a dynamic programming algorithm to choose a set of compatible base pairs maximizing the sum of K_{ij}. Since p_{ij} and q_i are predicted from the given sequences, they are not "priors" in a strict sense. However, the mutual information and the partition function look at the same data from different perspectives: the mutual information measures the conservation of base pairs in the particular sequences from a evolutionary point of view, while the partition function is based on a thermodynamic model that is generically applicable to all RNAs. Combining them gives us the power of both approaches: we rely on the energy model when there is little mutual information and use mutual information if the structure is ambiguous based on the energy model. In comparison with our method, RNAalifold (67) uses a linear combination of the energy contribution and mutual information. Our probabilistic integration provides some justification for combining these two seemingly disparate elements. Finally, to avoid prediction of isolated unreliable base pairs, we introduce a threshold parameter for helices; if the K_{ij} terms, summed over base pairs in a given helix, falls below this threshold, the helix is not reported as part of the optimal structure.

To update γ, its maximum likelihood estimate is simply $\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij}$. We introduce a pseudo-count to improve the robustness of the estimate:

$$\gamma = \frac{\sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij} + \gamma 0}{N + 1}$$

where $\gamma 0 = 0.3$. As the EM algorithm starts with M-step, the values of $P(x_{ij} = 1)$ are unknown in the first iteration, so they are actually updated at the end of the E step.

Convergence Criterion

The EM iterations terminates when the difference of the new score and old score is smaller than 2% of the old score, or when the score drops below -10, usually indicating a bad starting point which is not worthwhile for further improvement.

Adjusting candidates

We introduce a second EM phase identical to the first, except that the covariance model is used to scan each sequence, and the top hits in each are treated as candidates. This helps to discover motif instances missed by initial candidate selection and to refine the alignment. In fact, we found the second EM phase is usually sufficient by itself, so we use it to replace the first EM phase by default.

Adding additional family members

The EM framework enables automatic update of the covariance models based on new sequences that may contain the motif. We simply repeat the EM algorithm using the existing motif as seed. This feature is highly effective at discovering large and diverse RNA families from a small set of related sequences; see chapter 4.

Combining Motifs

In theory, this method can find arbitrarily large motifs, but in practice, it works best with relatively short ones (< 100 bases). To overcome this, we apply a greedy algorithm to merge multiple motifs hierarchically. To merge two motifs, we simply concatenate them, and the regions between the two motifs are aligned by padding with gaps. If two motifs overlap, the region of overlap is assigned to the motif with greater number of affected base pairs, while the involved base pairs in the other motif are removed. The merged motif is then iteratively refined using the EM module. To determine which two motifs should be combined, and the

order of combination, we applied a greedy hierarchical merging heuristic: for every pair of motifs, we compute a rough estimate of the alignment score after the combination, which is the sum of their alignment scores (the CM scores based on the Viterbi algorithm) penalized by the number of gaps or overlaps between them. The highest scoring pair of motifs will be combined and removed from future consideration. If the combined motif improves the scores relative to each of the two motifs, then it is added to the set of motifs for further combination, otherwise, it is removed.

Run time

The computation time of the EM algorithm is dominated by the CM scan with the Viterbi algorithm, with time complexity $O(NL^3|M|)$ for each iteration (where L is the maximum sequence length and $|M|$ is the number of states in the covariance model). In practice, the alignment is confined to a fixed window with length W, and time complexity in this case is $O(NLW^2|M|)$. The EM algorithm generally converges in fewer than 15 iterations. Overall, a typical CMfinder run (on $<$ 60 sequences of $<$ 1Kb average length) takes 1–60 minutes, depending on the number and the complexity of motifs, and is practical for most applications.

2.3 Results

As described in Chapter 1, Rfam is a large collection of multiple sequence alignments and CMs for ncRNA families (58). It contains a CM built from a hand curated "seed" alignment for each ncRNA family. Additional homologs are then predicted by searching genome databases with the model. We used Rfam seed alignments to evaluate CMfinder's performance on three increasingly difficult tasks:

1. Discovery: given unaligned seed members of an Rfam family (together with flanking regions), can CMfinder construct a good alignment and CM to characterize the family?
2. Robustness: How does the quality of the alignment degrade as more flanking sequence is added and as seed sequences are replaced by unrelated sequences?

3. Scale-up: Does it scale to plausible genome-wide discovery scenarios, where the initial set of sequences are taken, say, from a small group of orthologous genes? Are the sensitivity and specificity of the CMfinder model adequate when used to scan several gigabases of genome sequence?

In all three tasks, our results are strong. We will focus on the first two for the moment, and discuss the third in greater details in Chapter 5.

2.3.1 Discovery

We selected 19 families from Rfam (release 6.1, Aug 2004) as our test data, with varying length (26-216 bases), sequence identity (43% - 81%), and number of family members (9 - 75). This dataset captures the diversity of known RNA families, while excluding highly conserved ones, and emphasizing *cis*-regulatory elements, especially riboswitches.

For each family, we took the seed alignment as the target motif, and included 200 bases of genomic sequence flanking the motif, randomly distributed between the 5' and 3' sides, to simulate the realistic situation where motif locations are unknown.

We compared CMfinder with RNAalifold (67), Pfold (80), Foldalign (64; 51), ComRNA (71) and Carnac (144). The accuracy of all predictions are computed at the base pair level relative to Rfam annotation. Let TP be the number of correctly predicted base pairs, FP the number of falsely predicted base pairs, and FN the number of true base pairs that are not predicted. The sensitivity is defined as $TP/(TP+FN)$, specificity as $TP/(TP+FP)$ and overall prediction accuracy as their geometric mean. For decent sensitivity and specificity, the latter metric approximates Matthews Correlation (52). Note that this is a very stringent metric. Small shifts and extra base pairs are counted as false negatives/positives while they are considered neutral in some other work. More importantly, this penalizes incorrect motif boundaries in this local alignment setting.

Fairly benchmarking different programs created with disparate goals is tricky. We are sometimes using these tools for purposes that they were not designed or optimized to do. Nevertheless, we feel that our choices are plausible ones for our goals (automated, genome-scale RNA motif discovery) and constitute a reasonable comparison of these tools for that

purpose. The results may also illustrate the consequences of such tool abuse. CMfinder performs well in this context, but the other tools may excel in other contexts (or this one, if more cleverly used). In general, all programs received the same input, and were run with default parameters without any per-data-set tuning. As one exception, among the programs tested, only ComRNA can predict pseudoknots, but we chose non-default parameter settings preventing this, which may understate its performance. (Lacking a "gold standard" for pseudoknots in the test data, all obvious alternatives seemed worse.) As another important exception, both RNAalifold and Pfold require aligned inputs, so we used CLUSTALW alignments; other programs were given unaligned inputs. For ComRNA, we set a run time constraint comparable to the run time of CMfinder, and chose the motif with the best accuracy for each dataset. For Foldalign, we tried both multiple alignment (51) and pairwise alignment (64), and report the results for the latter as it is faster and generally gives better results. We summarize its accuracy as the average of all pairwise comparisons. Although pairwise alignments are not directly comparable with multiple alignments, the tool ultimately attempts to achieve the same goal as our tool, and it is interesting to test whether more sequences improve consensus structure prediction accuracy. For CMfinder, we produced up to 10 motifs for each dataset, combined them, and retained the motif with the best average alignment score as the final output.

The descriptions of Rfam families and prediction accuracies of all tested methods are summarized in Table 2.1. More detailed comparisons including predicted structures and alignments are included in the supplement website (166). It is worth noting that the test data used here is not independent of the data we used during development and tuning of CMfinder. Thus, the quantitative results reported in Table 2.1 may reflect some bias in favor of CMfinder. However, we believe that the heterogeneity of the test data and CMfinder's relatively small number of *ad hoc* parameters provide reasonable protection against significant "overfitting". Results reported in Chapter 5, especially Table 5.2, support that quantitative performance on Rfam families not included in our original test/training sets are broadly comparable to those reported here.

CMfinder achieved the best accuracy on all families except s2m and RFN. The disagreement with Rfam for s2m is due to a small shift of a helix. For RFN, the motif we produced is partial, and most prediction errors are constrained to local regions and in regions with

Table 2.1: Summary of Rfam test families and results. We included up to 200 bases of genomic sequence flanking the motif, randomly distributed between the 5' and 3' sides. #seqs: the number sequences in each family's seed alignment. (For ease of post processing, we only chose one sequence per EMBL ID.) %id: average sequence identity among family members. length: average length of family members (nucleotides). #hp: number of hairpin-loops in the consensus structure. Last 6 columns: accuracies; **bold** highlights the best result in each row. CW/Pfold: Pfold using ClustalW alignment. CW/RNAalifold: similar. (X: Carnac terminated abnormally, presumably due to memory problems. -: Foldalign (pairwise) not tested due to the heavy computation cost. RNAalifold, Carnac and ComRNA do not predict any structure in many cases, resulting in accuracies of 0.)

ID	Family	Rfam ID	#seqs	%id	length	#hp	CMfinder	CW/Pfold	CW/RNAalifold	Carnac	Foldalign	ComRNA
1	Cobalamin	RF00174	71	49	216	4	**0.59**	0.05	0	X	-	0
2	ctRNA_pGA1	RF00236	17	74	83	2	**0.91**	0.70	0.72	0	0.86	0
3	Entero_CRE	RF00048	56	81	61	1	**0.89**	0.74	0.22	0	-	0
4	Entero_OriR	RF00041	35	77	73	2	**0.94**	0.75	0.76	0.80	0.52	0.52
5	glmS	RF00234	14	58	188	4	**0.83**	0.12	0.18	0	-	0.13
6	Histone3	RF00032	63	77	26	1	**1**	0	0	0	-	0
7	Intron_gpII	RF00029	75	55	92	2	**0.80**	0.30	0	0	-	0
8	IRE	RF00037	30	68	30	1	**0.77**	0.22	0	0	0.38	0
9	let-7	RF00027	9	69	84	1	**0.87**	0.08	0.42	0	0.71	0.78
10	lin-4	RF00052	9	69	72	1	**0.78**	0.51	0.75	0.41	0.65	0.24
11	Lysine	RF00168	48	48	183	4	**0.77**	0.24	0	X	-	0
12	mir-10	RF00104	11	66	75	1	**0.66**	0.59	0.60	0	0.48	0.33
13	Purine	RF00167	29	55	103	2	**0.91**	0.07	0	0	-	0.27
14	RFN	RF00050	47	66	139	4	0.39	**0.68**	0.26	0	-	0
15	Rhino_CRE	RF00220	12	71	86	1	**0.88**	0.52	0.52	0.69	0.41	0.61
16	s2m	RF00164	23	80	43	1	0.67	**0.80**	0.45	0.64	0.63	0.29
17	S_box	RF00162	64	66	112	3	**0.72**	0.11	0	0	-	0
18	SECIS	RF00031	43	43	68	1	**0.73**	0	0	0	-	0
19	Tymo_tRNA-like	RF00233	22	72	86	4	**0.81**	0.33	0.36	0.30	0.80	0.48
				Average Accuracy:			**0.79**	0.36	0.28	0.17	0.60	0.19
				Average Specificity:			0.81	0.42	0.57	**0.83**	0.60	0.65
				Average Sensitivity:			**0.77**	0.36	0.23	0.13	0.61	0.17

great sequence conservation in which the structure is ambiguous. CMfinder significantly outperformed the other methods on families with low sequence conservation or short motifs such as the SECIS family. This is a difficult test case due to low sequence similarity, and two conserved non-canonical G-A pairs in the stem-loop. The other 4 methods tested predict no base pairs, while CMfinder correctly aligns and annotates the region enclosed by the two G-A pairs. A comparison of the CMfinder and Rfam motifs for SECIS is given in Figure 2.1. RNAalifold, Pfold, Carnac and ComRNA have relatively weak performance for such families, presumably because sequence conservation is insufficient to delineate possible alignments. Inclusion of arbitrary flanking regions makes sequence-based alignment even harder. Although our initial pairwise alignment algorithm is much simpler than pairwise Foldalign, we gained more information by comparing all sequences.

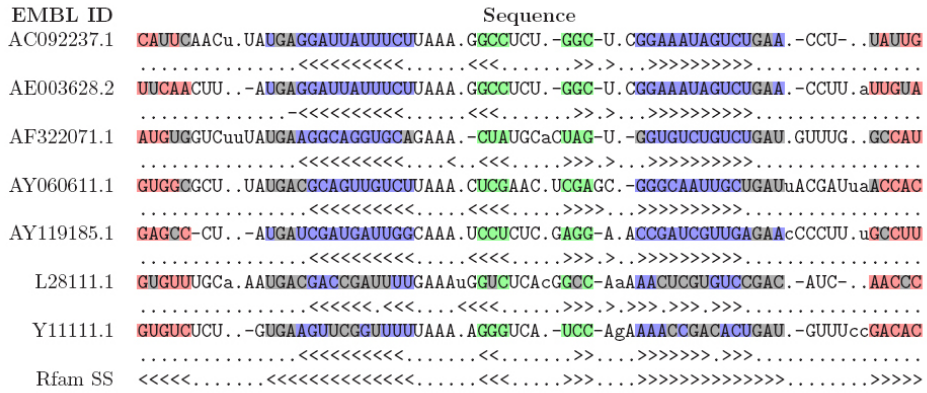


Figure 2.1: **SECIS element** Comparison of Rfam motif (RM) and corresponding CMfinder motif (CMM) alignment. Only 7 sequences are shown due to space limits. The colored block alignment corresponds to RM: blocks marked by the same color indicate a helix. Non-canonical base pairs in expected helices are colored light gray. Below each sequence is CMM secondary structure annotation; the last line shows Rfam consensus secondary structure. Although missing the short helices outside the non-canonical G-A pairs, the CMM alignment is generally in excellent agreement with Rfam.

We also tested CMfinder and other methods in the context of global alignments to see if CMfinder is competitive in this scenario. We compared CMfinder with 4 methods: RNAalifold, Pfold,Coveb (a program in the COVE package to construct a CM from a multiple alignment), and Covet (a companion of coveb, but using an EM algorithm to refine

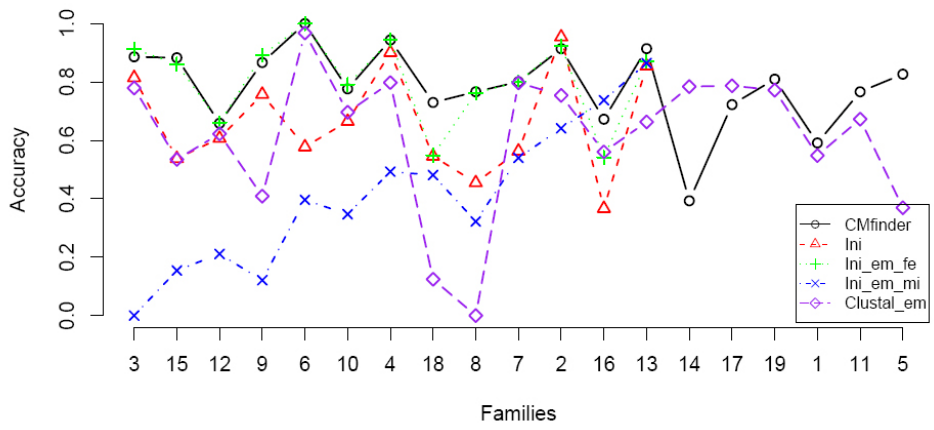


Figure 2.2: Comparison of CMfinder with its variants. The initial alignment corresponding to the best output motif is selected for each family. For the rightmost 6, final CMfinder motifs are combinations of multiple motifs, precluding comparison to ini, ini_em_mi, and ini_em_fe.

the CM). For fairness, each method is tested on the same CLUSTAL alignment for every dataset. The major difference between CMfinder and covet in this test is the use of partition function-based priors in secondary structure prediction. Both CMfinder and covet use an EM algorithm to refine the alignment and prediction, while the other three methods do not. The comparison of these methods is shown in table 2.2

To quantify the effectiveness of each component of our algorithm, we compared the performance of CMfinder with its four variants: 1. heuristic initial alignment before EM iteration (*Ini*); 2. initial alignment with EM iteration based on mutual information only (*Ini_em_mi*); 3. initial alignment with EM iteration based on folding energy only (*Ini_em_fe*); 4. CLUSTALW alignment with EM iteration (*Clustal_em*). To speed the EM iteration, we trimmed regions with more than 10% gaps from both ends of the CLUSTALW alignment.

The motif prediction accuracy of these 5 methods are shown in Figure 2.2. First, we observe that the EM iteration improves the prediction accuracy considerably, from an average of 66% in the initial alignments to 83%. Second, the energy-based partition function (*Ini_em_fe*) generally outperforms mutual information (*Ini_em_mi*) in the EM algorithm, while using mutual information in addition to folding energy yields improvements on SECIS and s2m. Third, CMfinder has better performance than Clustal_em except for RFN and S_box. For RFN, the CMfinder motif is only partial. Meanwhile, CMfinder is far more effective at locating poorly conserved and/or short motifs, such as IRE, and SECIS.

Table 2.2: Test on Rfam global seed alignments without the flanking sequences. The three numbers in each cell for a given family and a given method specify the accuracy, sensitivity and specificity respectively. The best accuracy, sensitivity and specificity for each family are highlighted in bold font.

Family Name	RFAM_ID	CMfinder			Pfold			RNAalifold			Covet			Coveb		
Cobalamin	RF00174	**0.61**	0.59	0.64	0.41	0.29	0.60	0.32	0.12	**0.87**	0.51	**0.60**	0.44	0.26	0.23	0.30
Entero_CRE	RF00048	0.77	**1.00**	0.59	**0.95**	**1.00**	0.90	0.71	0.50	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Entero_OriR	RF00041	**0.95**	**0.97**	0.93	0.80	0.63	**1.00**	0.70	0.63	0.76	0.06	0.05	0.09	0.00	0.00	0.00
Histone3	RF00032	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.70	0.67	0.74	0.57	0.33	0.96
IRE	RF00037	**0.92**	**0.95**	**0.90**	0.67	0.55	0.83	0.64	0.59	0.69	0.00	0.00	0.00	0.12	0.04	0.34
Intron_gpII	RF00029	0.71	**0.79**	0.64	**0.78**	0.62	**0.99**	0.76	0.58	0.99	0.67	0.69	0.66	0.55	0.52	0.58
Lysine	RF00168	**0.77**	**0.76**	**0.78**	0.59	0.49	0.70	0.22	0.10	0.48	0.51	0.49	0.53	0.41	0.38	0.44
Purine	RF00167	**0.90**	**0.93**	**0.87**	0.76	0.71	0.81	0.00	0.00	0.00	0.73	0.87	0.61	0.67	0.67	0.68
RFN	RF00050	0.73	**0.85**	0.62	0.72	0.81	0.64	**0.77**	0.81	**0.74**	0.62	0.80	0.47	0.58	0.51	0.66
Rhino_CRE	RF00220	**0.96**	**1.00**	**0.93**	0.66	0.48	**0.93**	0.77	0.75	0.79	0.33	0.28	0.40	0.00	0.00	0.00
SECIS	RF00031	**0.68**	**0.56**	**0.82**	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.03	0.00	0.00	0.00
S_box	RF00162	**0.81**	**0.89**	0.75	0.77	0.68	0.88	0.72	0.57	**0.92**	0.55	0.59	0.52	0.41	0.32	0.53
Tymo_tRNA-like	RF00233	**0.77**	**0.80**	0.75	0.62	0.52	0.74	0.73	0.54	**0.99**	0.58	0.54	0.63	0.45	0.34	0.58
ctRNA_pGA1	RF00236	0.89	**0.90**	0.88	0.86	0.80	0.92	**0.92**	**0.90**	**0.94**	0.16	0.12	0.21	0.00	0.00	0.00
glmS	RF00234	**0.74**	**0.71**	0.77	0.62	0.47	**0.81**	0.49	0.31	0.78	0.30	0.29	0.32	0.27	0.18	0.42
let-7	RF00027	0.79	**0.75**	0.83	**0.84**	0.71	**1.00**	0.80	0.68	0.93	0.00	0.00	0.00	0.00	0.00	0.00
lin-4	RF00052	**0.76**	**0.78**	0.73	0.72	0.57	0.91	0.73	0.56	**0.94**	0.18	0.14	0.24	0.18	0.08	0.38
mir-10	RF00104	**0.85**	**0.94**	0.77	0.75	0.61	0.93	**0.85**	0.84	0.86	0.00	0.00	0.00	0.01	0.01	0.02
s2m	RF00164	0.68	0.65	0.71	**1.00**	**1.00**	**1.00**	0.64	0.59	0.69	0.71	0.65	0.78	0.68	0.47	**1.00**
Average		**0.80**	**0.83**	0.78	0.71	0.63	0.82	0.62	0.53	**0.85**	0.35	0.36	0.35	0.27	0.21	0.38

This suggests that our heuristics are generally more robust, but CLUSTALW can be a complementary alternative for constructing initial alignments. Finally, there is significant improvement of Clustal_em over Pfold and RNAalifold. All three are based on CLUSTALW alignment, yet Clustal_em achieves 61% average prediction accuracy, compared to 36% for Pfold and 27% for RNAalifold. On families where both RNAalifold and Pfold fail, such as S_box, Cobalamin and Purine, Clustal_em has 55% to 80% prediction accuracy. To summarize, both the initial alignment procedure and the EM module are effective components of CMfinder, which make it reliable on a variety of datasets.

2.3.2 Robustness

We tested CMfinder on more challenging datasets with larger flanking regions and only a subset of the sequences containing real RNA motifs. To form each dataset, we randomly selected n family members, including a given length of flanking sequence, then permuted the motif regions in k of them (again, randomly selected). The latter sequences serve as *control sequences*; the rest, with real RNA motifs, are referred to as *test sequences*. We performed this test on Histone3, IRE and SECIS families. Figure 2.3(a) shows the CM scores of the motif instances produced by CMfinder when tested on datasets with flanking regions varying from 50 to 400 bases on both the 5' and 3' sides, with 1/4 control sequences, while Figure 2.3(b) varies the fraction of control sequences from 1/8 to 1/2, all with 100-base flanking regions. We categorized three types of predicted motif instances: true and false motif instances in the test sequences (**Tt** and **Ft** respectively), and those in the control sequences (**Fc**). Tts and Fts are determined by whether their overlaps with corresponding Rfam motif instances are greater than 10 bases. Most overlaps were shorter than 5 or longer than 25 bases.

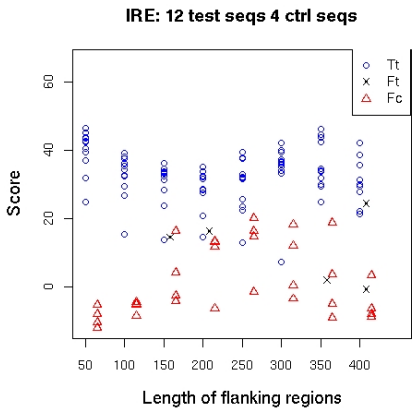
Figures 2.3(a) and 2.3(b) generally show good score separations between true motif instances (Tt) and false ones (Ft and Fc). As the flanking region increases, candidate selection becomes more difficult due to the larger number of stable local structures (we used the same number of candidates in all the test cases), and good local alignments occur more easily by chance. Although the score differences between the true and false ones

tend to decrease as the flanking region increases, the distinction is generally clear enough to differentiate the two types. In the IRE test, there are total of 20 false motif instances in all datasets above score threshold 10, but closer examination reveals that 11 of them in fact correspond to real IREs (which often occur in tandem) present in Rfam, but not seed members. The score difference between true and false motifs is even more apparent in Figure 2.3(b). The only Ft turns out to be a real IRE element. In both figures, there is only one Tt with score below 10. For IRE in Figure 2.3(a), there are five false motif instances in the test sequences (Ft), three with scores above 10, occurring at flanking length 150, 200, 350, and 400. On closer examination, three good scoring instances in fact correspond to real IRE (which often occur in tandem) in Rfam, but not seed members.

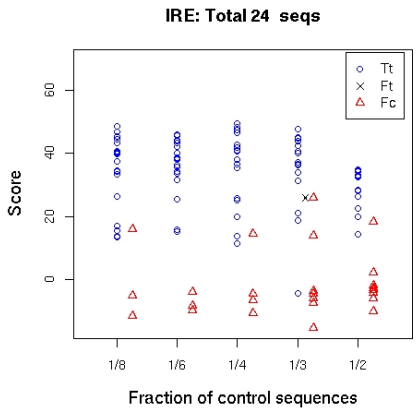
We have similar results for Histone3 and SECIS family (see Figure 2.3(c), 2.3(d), 2.3(e), 2.3(f)). Overall, even in the presence of significant amounts of extraneous sequence, CMfinder successfully predicted Histone3, IRE and SECIS motifs, which are among the more difficult of our test families.

2.4 Discussion

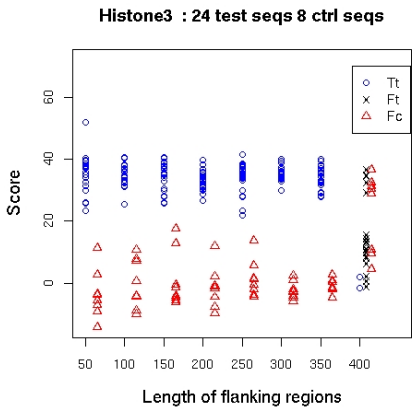
In summary, we have automatically learned highly accurate CMs from small automatically constructed datasets. In contrast, the Rfam models are learned from hand curated seed alignments, usually containing many more sequences. Automated model construction shouldn't supplant the high-quality curated Rfam seed alignments, but we are optimistic that it will allow broad-scale screening for new *cis*-regulatory ncRNAs with minimal manual intervention. Further extension of CMfinder and applications are discussed in the following chapters.



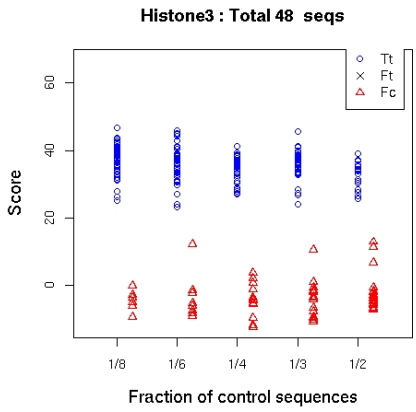
(a)



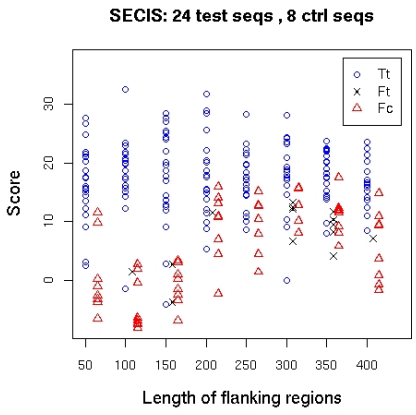
(b)



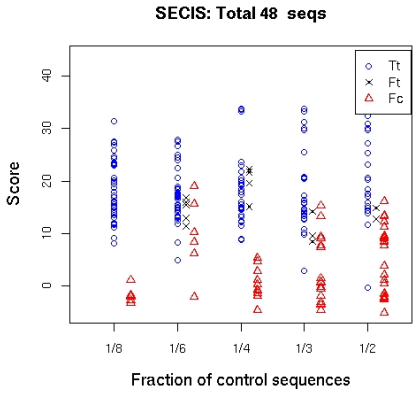
(c)



(d)



(e)



(f)

Figure 2.3: Robustness Test on IRE, Histone3 and SECIS family. Each column represents a test dataset, each point represents a motif instance predicted in a sequence. Tt: True motif in test seq.; Ft: False motif in test seq.; Fc: False motif in control seq. (see text).

Chapter 3

SCALING CMFINDER TO LARGE DATASETS

3.1 Introduction

Although CMfinder is robust to noise, and is capable of identifying various types of noncoding RNAs, it is not scalable to long sequences. In an attempt to make CMfinder a better tool for analyzing the eukaryotic genomes with large intergenic, intronic or UTR regions, we propose several techniques to tackle the performance issues relating to long sequences. The key ideas include partitioning the sequences into regions enriched of sequence homology, speeding up the covariance model scan by using a heuristic Hidden Markov Model filter, and re-formulating the probabilistic model under more realistic assumptions. The method described in this chapter has been implemented in CMfinder version 0.3.

Our experimental results demonstrate that the scaled-up CMfinder correctly predicts motifs in datasets with up to 10 sequences of 50KB bases each in a few hours. We have carefully benchmarked its performance in varying application settings, and showed that it is still robust to noise and varying sequence conservation levels in its input dataset.

3.2 Methods

3.2.1 Reduce search space using sequence similarity

Because it is expensive to consider both sequence and structure simultaneously for RNA motif prediction, our initial attempt to reduce the search space is to limit the RNA motif search within regions based on sequence conservation only. However, as we have emphasized previously, the presence of covariation, optional and variable length stems, and large indels can make this task difficult. On the other hand, it is common for ncRNAs to contain patches of local sequence conservation, and as functional elements, they tend to be more conserved than their local background, which we hope to exploit to locate regions enriched with sequence homology as potential hosts of ncRNA motifs.

We could potentially exploit a traditional multiple sequence alignment tool for this purpose. However, our results suggest that many of these methods are not robust enough extract RNA motifs with low sequence similarity (see Results table 3.2). In addition, many such tools only perform global alignment, or assume that homologous regions are in a consistent order in different sequences.

We attempt to solve this problem without computing the exact multiple sequence alignment. The key idea was to use all against all pairwise sequence comparison to locate the matched segments, and cluster these segments to maximize the pairwise similarity between the members. Consider the case in which the motif instances contained in three sequences all have relatively low pairwise sequence similarity. In the context of pairwise sequence comparison, the similarities between these motif instances are likely to be obscured by those of random matches. However, since the random matches are uniformly distributed across the whole sequence, it is unlikely that matches in sequence A with sequence B overlap with matches in sequence A with sequence C by chance. Therefore, by searching for clusters in which all members share reasonable pairwise sequence conservation, there is a better chance to detect these motifs with relatively weak sequence conservation.

To do this, we first perform all against all pairwise Smith-Waterman local alignment. We use a slide window of 200 bps, with offset of 100 bps, and align each window to the other sequence. For all matches at a given end position, we only select the maximum one, given that its match score is above a very conservative threshold (equivalent of fasta score 70). If two matches overlap significantly (i.e. more than 80%), the one with lower score will be removed. We report the top 200 such matches. Then, we try to select a set of segments, one from each sequence, such that the sum of all pairwise similarity scores will be maximized. Since not all pairwise local matches are kept in the comparison step, it is common that for two overlapping segments $A1$ and $A2$ in sequence A, the match score between $A1$ and a segment $B1$ in sequence B $s(A1, B1)$, and match score between $A2$ and a segment $C2$ in sequence C $s(A2, C2)$ are included while the scores $s(A2, B1)$, and $s(A1, C2)$ are not included. In this case, we estimate the similarity score as the following:

$$s(A2, B1) = s(A1, B1) * len(A2)/span(A1, A2)$$

where "len" refers the length of the given segment, and "span" is the length of the shortest interval that covers both the segments. We use this estimation scheme so that for non-overlapping nearby segments $A1$ and $A2$, $s(A2, B1)$ will still be nonzero, providing support for $A2$ and $B1$ to be contained in the same cluster. Next, we choose a seed segment such that sum of the similarity scores to its best matches to all other sequences is maximized. From all segments of the remaining sequences, we select the one with the maximum sum of match scores to selected segments. This process terminates if this sum divided by the number of selected segments is below a certain threshold, or if there is already one segment selected from every sequence. While this process is effective to find the most conserved core of an RNA motif, it may miss the parts that are less conserved. Therefore, we try to extend the boundary of the cluster to increase the chance to include the whole RNA motif within one cluster. We do this by including the segments nearby the segments within the cluster. Again, we choose iteratively the segment that is within a specific range of the already selected segment of the same sequence, which has the maximum sum of match scores with the selected segments of other sequences. The process terminates if there are no more segments to add. We define a cluster member as the shortest interval that covers all selected segments within a sequence. If a cluster member has length smaller than 300bp, we extend it evenly at both ends to 300bp if possible.

We use this process to iteratively produce multiple clusters. All matches that overlap with previously selected clusters are down weighted based on the length of the overlap. For each cluster, we only search candidate motif instances within cluster segments, and subsequent steps are mostly the same as in the previously described CMfinder. The time complexity of the this step is dominated by pairwise Smith-Waterman algorithm, which is $O(m^2n^2)$, where m and n are the lengths of the two input sequences respectively. The cluster range is typically between 300-600bp.

Besides the pairwise sequence comparison step that is required by most multiple sequence alignment algorithms, the clustering step is very fast, as it avoids computing the exact alignment. This speed advantage enables it to check many pairwise sequence comparisons simultaneously to determine the approximate location of conserved regions, making it robust to detect weak yet consistent sequence conservation. This method is based on local sequence

conservation, so it can deal with genome rearrangement, a problem that plagues many multiple sequence alignment tools. In case of sequence duplication, this clustering procedure includes one copy in the cluster.

3.2.2 Applying heuristic HMM filter

The EM algorithm for CMfinder works as follows: Given an initial seed alignment, CMfinder predicts a consensus structure, and constructs a CM. The CM is then used to scan the sequence dataset for motif instances. The scanned motif instances are aligned based on the CM model. The process is repeated until convergence. The bottleneck of this EM algorithm is using the CM to scan the sequences. We can use the clusters found in the previous step to narrow down the search range. On the other hand, it is possible that a cluster may miss some motif instances that are more distant from others. It would be nice to have another chance for recover these missing instances during CM search. To speed up CM search, we apply the filter based on the heuristic Hidden Markov Model (HMM) proposed by Weinberg and Ruzzo (157). This method creates a profile HMM from a given CM based on the maximum-likelihood principle, and then use the HMM to search the sequences for candidate CM hits. The HMM search is far more efficient than the CM search, with time complexity of $O(nH)$, where H is the number of states in the HMM, and n the sequence length. Using this technique we can afford to search sequences at the megabase level during the iteration. For easy integration with the existing CMfinder package, instead of using the RaveNnA package, we use the HMM filter implemented in Infernal (version 0.72). We still keep the CM scan option. This is useful for short RNA motifs such as iron response element (IRE), for which the corresponding HMM filter lacks the necessary specificity. We incorporate other new features from the Infernal package, such as Query-Dependent Banding (QDB), which pre-calculates regions of the dynamic programming lattice that have negligible probabilities, resulting in approximately 4-fold speed up for average RNA query. We also use HMM banded CYK algorithm to speed up the CM alignment algorithm. This new Infernal feature has not yet been described in the literature.

3.2.3 New model for motif instances weighting

Two component model (TCM)

In our previous work, we described an input sequence as a mixture of regions that follow the background distribution and regions that follow the CM, assuming zero or one motif occurrence per sequence (a.k.a ZOOP model). See section 2.2.2 for modeling details. We made several approximations when we computed the probability of a candidate to be a true motif instance. First, instead of computing all possible motif instances, we only searched for candidates whose CM scores are above 0, and let their prior be $\frac{\gamma}{n}$, where n is the length of the sequence, and γ the prior that one sequence contains a motif instance. Secondly, due to pragmatic reason, we estimate the log likelihood of a candidate to be an instance of the CM vs. the background model as the CM score based on its CYK alignment. Theoretically, the likelihood of candidate given the CM should be computed by the inside-outside algorithm, which summarizes probabilities of all possible alignment paths. CM scores for motif instances tend to follow bimodal distribution, so the separation of true motif instances and the background is usually quite clear. The prior, and the fact that we use CYK score instead of the inside-outside score does not make much difference. As the sequence length increases, a candidate needs to have much higher CM scores to compensate for the lower prior. This can be problematic for short RNA motifs, whose highest possible CM scores are still quite low. We address these issues by using the two component model (TCM) (7) instead of the ZOOP model. The TCM model is described as the following:

$$P(x_{ij} = 1) = \frac{P(c_{ij}|M)\lambda}{P(c_{ij}|M)\lambda + P(c_{ij}|B)(1-\lambda)}$$

where λ is the prior for a candidate to be a motif instance. The CM score based on the inside-outside algorithm corresponds to the log likelihood ratio $s_{ij} = \log_2 \frac{P(c_{ij}|M)}{P(c_{ij}|B)}$, and the above formula is rewritten as

$$P(x_{ij} = 1) = \frac{\lambda}{\lambda + 2^{-s_{ij}}(1-\lambda)}$$

We approximate the inside-outside score by CYK score if the latter is greater than $\log_2(100/\lambda)$, as the resulting difference is negligible, otherwise the inside-outside score is used.

We let λ be a user defined value. When searching for large, highly conserved RNA motifs, we recommend using a small value for λ to minimize false positives. If a user is interested in searching small, nonspecific RNA motifs, we recommend using a larger value for λ.

The TCM model can also handle multiple motif instances within one sequence, which is problematic with the ZOOP model.

The Extreme Value Distribution (EVD) model

There is another problem for both TCM and ZOOP models. When the CM is trained over an alignment with very few sequences, the model becomes very specific, and tends to give very pessimistic score for even a good hit. This problem can be potentially addressed by the extreme value distribution (EVD). It has been shown that un-gapped local alignment scores for random sequences follow EVD, which is also empirically true for gapped local alignments based on simulation studies (2). It has been suggested that the gapped structural alignment scores roughly obey EVD as well (78). Under the EVD assumption, given an arbitrary query sequence, the probability that its alignment score is greater than or equal to a given score x is given by the formula

$$P(S \geq x) = e^{-\lambda(x-\mu)}$$

where λ is the scaling factor and μ the mean for EVD. Given a random database of length n, the probability that its best alignment score is greater than or equal to score x is

$$P(x, n) = 1 - e^{-nP(S \geq x)}$$

In the framework of CMfinder, we assign the probability of candidate with CM alignment score of x to be a true motif instance as $1 - P(x, n) = e^{-nP(S \geq x)}$, i.e., the probability that the best hit of a random database of n has score less than x. For a high specific CM, the mean μ tend to be low, therefore increase the weights of the weak hits.

Estimation of the EVD parameters is a challenging task. The common practice is to sample random sequences, collect their CM scores, and then fit the curve. The problem is, we need fairly large sample size to obtain a reliable estimate, which is infeasible due to

CM scoring cost. To evaluate the robustness of the parameter estimates, We have fitted the EVDs of RFAM CM models using 10,20,50, 100 and 1000 samples, and observed that while the μ estimates are relatively independent of the sample size, the estimates of λ vary significantly. For most high scoring, or low scoring CM hits, the value of λ has minor effects on their probabilities. However, this is not true for boundary cases. The situation is aggravated by the fact that in the EM algorithm, the CM changes at each iteration, which requires the EVD parameters to be adjusted correspondingly.

Due to the computation cost associated with the EVD training, we found this approach less attractive. We still leave it as an option, using only 20 random samples for training, and only doing this if the CM scores changed significantly during the EM iteration. We can imagine under certain circumstances, when one needs to expand a narrow group of RNA homologs, this option can be useful. Unless specified otherwise, TCM is used by default.

3.2.4 Other miscellaneous changes

We have made several additional minor changes in CMfinder.

When we create heuristic initial alignments, we perform pairwise tree-edit comparison of motif instance candidates based on their putative secondary structures. This comparison was based on a naive score matrix that did not discriminate different types of base pairs. We now use RIBOSUM matrix from the Rsearch package (78), trained over large ribosomal RNA database. We also make a couple of changes in our strategy to create an initial alignment. To avoid producing highly similar motifs, we previously examined whether the overlaps between a newly created alignment and previously selected alignments is above certain threshold, and discarded it if it does. The performance of this procedure was very sensitive to the choice of the threshold value. In the new version, to select a new consensus candidate, its match scores with its best matches in other sequences are weighted based on the overlap of the matched segments with the selected initial alignments. Therefore, if matched segments of a candidate are already contained in the selected alignments, this candidate is unlikely to be selected.

To address the issue that the candidate secondary structure based on single sequence

prediction is unreliable, causing candidates corresponding to true motif instances to have low similarity score, we increase the similarity of pair of candidates that share significant sequence similarity. To do this, we select all pairwise local sequence matches with pvalue less than 0.1 (The pvalues are computed based on the match scores and the extreme value distribution, using the same EVD parameters as in FASTA), then compute the overlap between a pair of candidates with these sequence matches. The tree-edit similarity score is scaled by one plus the fraction of the overlap. This is based on the intuition that while CMfinder is tolerant to errors in the initial alignment, it requires the initial alignments to cover the approximate region of the true RNA motif.

In the EM algorithm, we re-computed the partition function for all motif instances at each iteration. In the current implementation, if a motif instance is contained in another motif instance in previous iterations, and boundaries match closely, we try to reuse the available partition function as much as possible. To avoid wasting computation time on bad motifs, the EM iteration terminates before convergence if the current alignment contains little structure (with fewer than three consensus base pairs), or the sum of motif probabilities is less than 2.

We also introduce minor enhancement to our algorithm for combining motifs. In the previous version, when two motifs are merged, the regions between them are simply aligned by padding with gaps. To capture the sequence similarity more effectively, they are now aligned by CLUSTALW. In the process of merging motifs progressively, the newly merged motif is compared to the two original motifs. If there is not much additional secondary structure within the new motif, it will be eliminated.

We also simplify the usage by adding a command that incorporates all major steps of CMfinder, and introducing a convenient default mode that infers appropriate input parameters based on features of the input sequences, which makes it more accessible for average users. It infers the following options:

- **The number of clusters**. If the average input sequences are shorter than 500 bases, the clustering step is skipped. Otherwise, the number of clusters is about 3-5 depending on the sequence length.

- **The expected fraction of motif instances.** This parameter is used to prevent the contamination by irrelevant sequences. When the number of selected candidates is more than n times this fraction, all other candidates are ignored if they are not similar to the consensus candidate at all. By default, we requires at least five instances (or the total number of input sequences if it is less than five) to be included in initial alignment, and this fraction set to be *5/seq_num*. If this fraction is below 10%, it is set to 10%.

- **The number of predicted motifs and length of motifs.** By default, CMfinder produce up to five single stemloop motifs and five double stemloop motifs (before merging). The lengths of single stemloop motifs range from 30 to 100 bases, while the double stemloop motifs range from 40 to 100 bases.

- **The use of the HMM filter and HMM banded constraints.** The HMM options are only relevant for the EM phase. As CM scans for short and simple motifs are fast and the risk of HMM scans to miss some good hits is considerably high for such motifs, we do not use the HMM speedup options when the initial motif is shorter than 40 bases. Otherwise, we use both the HMM filter and HMM banded constraints for scan and alignment respectively.

The motifs are combined by default, and the HMM options in the combination step are also determined based on alignment length.

3.3 Results

3.3.1 Create benchmark

To test the robustness of CMfinder 0.3 on datasets with different features, we created a benchmark from 18 RFAM (58) families, all of which are used in our previous study (see Section 2.3.1). We dropped one family as its alignment has been revised in Rfam since we published the previous version, and it is a relatively easy case. For each family, we took several subsets with different numbers of sequences and sequence similarity, and

included varying lengths of the flanking regions within corresponding genomic context. In particular, for each family, we sampled 5 sequences from RFAM seed alignments with high sequence similarity, (i.e. all pairwise sequence similarities are within the range of 65% to 95%), and their 2000 bp, 10000 bp, and 50000 bp genomic flanking regions, randomly distributed between $5'$ and $3'$ side. (The flanking regions are retrieved from the EMBL database, and it is constrained by the sequence length of given accession ID.) We refer to this set as the *high* similarity set. We then sampled another 5 sequences, 3 of which have pairwise sequence similarity of 65% - 95%, while the pairwise similarity of these 3 and other 2 remote homologs are within the range of 30% and 65%. This set is referred to as *mix* similarity set, and we included in their flanking regions as above. We used the same sampling process to create *high* and *mix* sets with 10 sequences, and the *mix* set in this case includes 3 outliers. To simplify comparison in case of tandem RNA motif occurrences, we scanned each dataset for all possible motif instances using RFAM CMs, and remove the ones that are not included the sampled seed alignment. Finally, we use RepeatMasker (`http://repeatmasker.org`) to remove repeat regions. For each family, we could potentially create up to 12 datasets with various features. However, in many cases, we could not sample enough sequences that satisfy our criteria, and we lower the the number of sequences by one if necessary. Datasets whose total sequence length is less than half of the required flanking length are removed. The number of datasets that fit in each category is described in Table 3.1. Predicting RNA motifs in *high* and *mix* datasets represents two

Table 3.1: Summary of benchmark datasets. "#seq":the number of input sequences. "sim": the similarity category. Under "2000", "10000" and "50000", the numbers specify the number of datasets that fall into the corresponding categories, with flanking region of length 2000, 10000 and 50000 bp respectively.

#seq	sim	2000	10000	50000
5	high	12	7	5
5	mix	15	11	8
10	high	7	3	1
10	mix	5	4	3

different types of challenges. In the *high* datasets, the flanking regions of the RNAs tend to be conserved too. While it is easy to construct the sequence alignment in this case, it is difficult to discriminate structural conservation from purely sequence conservation. Without sufficient constraints from compensatory mutations, the consensus structures tend to be unreliable and ambiguous, and identification of the exact motif boundaries is difficult. For the *mix* datasets on the other hand, the challenge lies in locating the remote homologs. However, once the remote homologs are included in the motif alignment, they are very effective to constrain the consensus structure, and to provide sequence diversity in the covariance model, enabling it to find even more diverged homologs in the database search. Short of experimental validation, the presence of diverged homologs is among the strongest evidence of true RNA motifs in computational analysis.

3.3.2 Performance of clustering heuristics

We used the clustering heuristics to locate the homologous regions that are likely to host RNA motifs. Here, we tested how well this technique clusters real homologous RNA motif instances. Since standard multiple sequence alignment is a natural alternative to locate homologous regions, we used CLUSTALW (139) as a comparison baseline. We also compared our technique with DIALIGN (100), an algorithm that aligns the sequences based on local sequence conservation, and MAFFT (76; 75), another multiple alignment algorithm that is generally regarded as fast and accurate. For CLUSTALW and DIALIGN, we used the default setting. For MAFFT, we use option "–clustalout –auto –ep 0.0". To evaluate their performance, we used a sliding window of 600bp to scan the whole alignment, and counted the maximum number of motif instances that were contained in one window. For our technique, we counted the number of true RNA motifs that were included in one cluster (if a partial motif was included, we took the length fraction of the included region over the whole motif). Since the sequence segments in each cluster were almost always less than 600 bases, this is a fair comparison. We produced 3, 4, 5 clusters respectively for datasets with 2K, 10K and 50K flanking regions. Our method is penalized by using a small number of clusters when the flanking regions are also conserved. The alignment algorithms are not

penalized in the same way in this comparison.

In the table 3.2, we list in each dataset category, the average fraction of the true motifs that were included in one sliding window (for the alignment methods) or in one cluster (our technique). Not surprisingly, the performance of CLUSTALW degrades dramatically as the length of flanking regions increases, or as the motif sequence similarity decreases. DIALIGN, on the other hand, is quite robust as the length of flanking region increases. DIALIGN ran very slow on the datasets with 50K flanking regions, and in many cases, it failed due to out of memory error, so we did not report its performance in such circumstances. The performance of MAFFT is worse than DIALIGN, but slightly better than CLUSTALW. The advantage of MAFFT over CLUSTALW, however, disappears when tested on the 50K datasets.

Our clustering heuristics consistently included all real motif instances in one cluster for 70 out of 80 datasets in the benchmark. Most of the remaining 10 datasets correspond to short RNA motifs such as IRE, Histone3 and Entero_CRE, in which the clustering algorithm missed a few instances or a few bases at the boundary. There is only one case in which our clustering technique failed completely, which corresponds to the glmS family with *mix* similarity and 50K flanking regions. It fails in this case because the RNA motif is sandwiched by the coding sequences of upstream and downstream genes, which are far more conserved than the motif. Compared to the tested alignment algorithms, our clustering technique was more sensitive to include the remote RNA homologs within the same cluster.

3.3.3 Performance of CMfinder

In Tables 3.3 and 3.4, we show the performance of CMfinder under the default configuration. Then of all motifs predicted, we selected up to 4 motifs per cluster based on a heuristic scoring function (see Section 4.2 for details). All predicted motifs are compared to Rfam seed alignments at the base pair level. The definition of sensitivity and specificity is the same as in Chapter 2. We consider sensitivity to be more critical than specificity for de novo motif discovery, because it is easier to disregard the degenerate parts of a motif than to recover the missing parts in manual inspection that is subsequent to automatic prediction.

Table 3.2: Comparing the cluster algorithm with other alignment algorithms. The first three columns specify the characteristics of the dataset. Under "CLUSTALW", "MAFFT" AND "DIALIGN", each number specifies the maximum fraction of true RNA instances covered by any 600bp window, averaged over all datasets within the given category. Each cell under "cluster" is the maximum fraction of true RNA instances covered by all clusters for the dataset, again, averaged over all datasets within the given category. The method with the best performance within each category is highlighted in bold font.

			CLUSTALW	MAFFT	DIALIGN	cluster
5	high	2000	0.72	0.85	0.94	**1.00**
		10000	0.47	0.60	0.94	**0.98**
		50000	0.35	0.30	NA	**1.0**
	mix	2000	0.67	0.72	0.89	**0.97**
		10000	0.45	0.54	0.82	**1.00**
		50000	0.39	0.37	NA	**0.88**
10	high	2000	0.80	0.82	**0.94**	**0.94**
		10000	0.51	0.55	**0.97**	0.96
		50000	0.30	0.20	NA	**1**
	mix	2000	0.65	0.70	0.92	**0.98**
		10000	0.33	0.47	0.77	**1.00**
		50000	0.27	0.23	NA	**0.97**

For each dataset, we picked one motif from the selected motifs whose sum of sensitivity and half of specificity is optimized, and the corresponding sensitivity and specificity are represented in Tables 3.3 and 3.4, Under each category, the sensitivity is mostly within the range of 0.7 - 0.8, while specificity is mostly within the range of 0.5 - 0.65. The relatively low specificity reflects our bias in choosing the motifs and the fact that CMfinder tries to merge motifs as much as possible. In addition, some of our selected motifs are more conserved than the original RFAM seed alignments because relatively few sequences are selected. As a consequence, the predicted motifs tend to expand beyond the real RNA boundaries.

The secondary structure prediction accuracy is determined by several factors: the occurrences of the motif within the input sequences , the exact motif boundaries and the secondary structure. There are some noticeable trends. First, the performance of CMfinder is stable while the flanking regions increase except for the glmS family when flanking region

is 50K. As we discussed earlier, the clustering fails to locate the homologous region in this case. Second, the prediction accuracy improves as more motif instance become available. For the Intron_gpII family, the prediction accuracy of the *mix* set is worse than *high-sim* set, because in the *mix*, there is a large indel in the Rfam alignment, which is not correctly aligned by CMfinder. On the other hand, For the RFN family, the prediction for the *mix* set has higher accuracy than the *high* set, because the secondary structure is less ambiguous with constraints from diverged sequences. Our method has problem with the *mix* set of the IRE family with the 2K flanking region, as it fails to identify two remote homologs. Instead, a spurious instance was included due to its homology to the IRE stem regions. Another false positive by our definition that is right next to a real IRE appears real. It was not filtered as we intended to, probably due to its close proximity to the real one. Although both sensitivity and specificity are low for this IRE dataset, the motif structure is actually very close to the real one.

We generally expect the performance of CMfinder to get worse as the length of the flanking sequences increase. However, this is not always the case. It can be attributed to the heuristic step of the algorithm, which may choose different sets of candidates when the boundaries of input sequences change, and cause considerable variation in the output motifs.

3.3.4 Computation time analysis

The CMfinder running time is approximately linear with the sequence length, and is sub-quadratic with the number of sequences (see Table 3.5). Within our benchmark, the running time ranges from a few CPU minutes to a few CPU hours, which is practical for most applications.

3.3.5 Performance of CMfinder in noisy datasets

In this experiment, we tested the performance of CMfinder in the presence of random sequences that do not contain the RNA motif instances (see Table 3.6). We took all the datasets in the *mix* category with 2K flanking regions. For each dataset, we sampled the

same number of sequences from the original dataset, and randomly permuted them while maintaining the first-order Markov properties. For example, for a dataset with five real motif instances, we added another five random sequences. Out of 20 such datasets, there are only four cases where we did not predict the right motif instances, i.e. predictions do not overlap the real motif instances in corresponding sequences. Three of them correspond to short RNA motifs with a simple hairpin: IRE and SECIS. The errors are partly due to the initial clustering algorithm which missed a few instances with very weak sequence identity (as low as 30%), and these missing cases have never been recovered. In addition, as these motifs are short and non-specific, they are likely to be contaminated by random sequences. The relative low sensitivity and specificity in secondary structure prediction are due to these false positive and false negative motif instances. Our predicted motifs nevertheless share the same structure and same consensus sequence as the real motifs. The Purine motif includes a false positive motif instance from random sequences. The alignment score is very high, although still much lower than the real motifs. Overall, we were able to predict the bulk of the motif structure in almost all test cases, and secondary structure prediction has average 71% sensitivity and 64% specificity.

3.3.6 Evaluation of CMfinder components

We have introduced several techniques to address various scalability issues. Among these, the clustering step is the most critical, without which the method would fail completely on the large datasets that we have tested, so we always use this option. We compared the default mode with two other variants: **var1** does not use the HMM filter, and **var2** uses the HMM filter for scanning, but does not use HMM banding constraints for alignment, and results on the benchmark datasets are summarized in Table 3.7. Note that for CM scanning, CMfinder limits search within the range defined by the cluster, and expands the search to the whole sequence only if no good hits are found. Since the clustering algorithm is usually very successful at narrowing down the search range, the full sequence CM scans are pretty rare. Due to this reason, the running time difference between the default mode and var1 is less significant than we would expect, but the ratio is still about 2-6 fold. If

testing on the noisy datasets wherein many sequences do not contain the motif, we would expect greater difference as full sequence scans become much more frequent. Comparison of the default mode and var2 implies that the HMM banded constraint technique brings considerable amount of speedup additional to that of the HMM filter. The comparison of the sensitivities and specificities on the secondary structure prediction suggests that both speedup techniques cause little sacrifice of accuracy, although var1 with CM scan performances slightly better.

We also tested CMfinder with the original ZOOP model and the EVD model for motif instances weighting on the benchmark (data not shown). However, the difference of their performance is almost negligible. This is partly due to the fact that we set the prior of the ZOOP model to be dependent on the length of the cluster segment computed in the initial clustering step, not the length of the entire sequence. In addition, the motif signals in the benchmark datasets are general very strong so that motif instances can be distinguished rather easily. We have observed occasionally that for the ZOOP model, some predicted motifs miss a couple of true instances, but these instances are later recovered during the motifs merging process. For example, one motif (A) covers a conserved region of S_box and include all true instances, while another motif (B) covers a less conserved region of this RNA and misses a true instance. The merging process recovers the missing instance in B due to the strong signal in the nearby region in A. Due to many such compounding factors, we found that while TCM model tends to be more robust on noisy datasets and on long sequences, the effect does not manifest on this benchmark study.

3.4 Discussion

In this chapter, we have presented some enhancements of CMfinder that make it more scalable and easier to use. Extensive tests on benchmark datasets, many of which are intractable by traditional methods, demonstrate that the upgraded CMfinder performs satisfactorily and consistently, except for a very few cases. We have yet to test whether this method is capable to detect known ncRNAs in a realistic setting while other traditional methods can not, because most of these ncRNAs are either predicted based on routine sequence-based computational analysis, or discovered experimentally but not conserved be-

yond closely related species. However, we believe this phenomenon is simply due to lack of appropriate methodology. With the availability of the powerful new tool, it is now practical to conduct large scale comparative study on nonconserved genomic regions for presence of ncRNAs. We will continue to investigate in this direction.

CMfinder still suffers from a few limitations. As an EM algorithm, it can be trapped at local optima. For example, when a CM is trained over a set of highly conserved sequences, it may not discover remote homologs. Some local misalignments may never to be corrected. Experts with trained eyes can refine the predicted alignments manually based on a few general principles, but this is a time consuming and challenging task. It is possible to simulate this trial and error process by an algorithm, probably formulated in the Markov chain Monte Carlo (MCMC) framework. We are also interested in improving the design of the covariance models. Current covariance models can not describe variable size stems effectively. It may be as simple as adding states for insertion of base pairs (only insertion of single bases is currently allowed), but we do not know yet the ramification of such a change. Taking the phylogeny into account in motif search is also important. As input sequences are not evolutionarily equidistant, a group of closely related sequences sometimes biases the motif model, and marginalize the more distant ones. One potential solution is to weight the sequences so that the distant ones receive higher weight. Such methods have been used in CLUSTALW and Infernal. One major obstacle is how to obtain an accurate phylogeny without an alignment. How to incorporate such information in both the heuristic and EM steps is another challenge.

We have investigated how to make this method scalable to sequence length in this study, but scaling with regarded to the number of sequences is still unsolved. For example, CMfinder has problems dealing with the datasets with hundreds of sequences and with only very few sequences containing the motif. The challenge here has more to do with higher signal to noise ratio rather than the computational overhead. This is a very difficult and important problem that we hope to make progress on in the future.

Table 3.3: CMfinder performance on benchmark: Sensitivity of secondary structure prediction at the base pair level.

family	5-high			5-mix			10-high			10-mix		
	2K	10k	50k	2K	10k	50k	2K	10k	50k	2K	10k	50k
ctRNA_pGA1	0.93	0.93		0.96	0.96					0.93	0.95	
Rhino_CRE	0.53						0.54					
Entero_CRE	0.91	0.88					0.92	1				
let-7	0.7	0.73	0.7	0.81	0.81	0.81						
mir-10	1	1	1	0.94	0.94							
lin-4				0.8	0.7	0.7						
Histone3				1			1					
s2m	0.51						0.76					
IRE	0.85			0.38			0.84			0.6		
SECIS	0.41			0.47								
Intron_gpII	0.89	0.89	0.89	0.72	0.87	0.74				0.86	0.86	0.85
Tymo_tRNA-like				0.81								
RFN	0.24	0.34	0.33	0.56	0.46		0.69	0.38		0.6	0.61	0.78
S_box	0.66	0.64	0.66	0.61	0.87	0.45	0.92	0.92	0.92	0.53	0.7	0.67
Purine				0.93	0.93	0.89						
Lysine				0.89	0.91	0.74						
glmS				0.77	0.77	0						
Cobalamin				0.45	0.45	0.58						
mean	0.69	0.77	0.72	0.74	0.79	0.61	0.81	0.77	0.92	0.70	0.78	0.77

Table 3.4: CMfinder performance on benchmark: Specificity of secondary structure prediction at the base pair level.

family	5-high			5-mix			10-high			10-mix		
	2K	10k	50k	2K	10k	50k	2K	10k	50k	2K	10k	50k
ctRNA_pGA1	0.57	0.57		0.97	0.97					0.42	0.41	
Rhino_CRE	0.28						0.66					
Entero_CRE	0.87	0.68					0.8	0.57				
let-7	0.91	0.94	0.91	0.83	0.83	0.83						
mir-10	0.73	0.67	0.73	0.66	0.66							
lin-4				0.55	0.58	0.46						
Histone3				0.31			1					
s2m	0.46						0.52					
IRE	0.6			0.26			0.61			0.7		
SECIS	0.33			0.69								
Intron_gpII	0.62	0.62	0.62	0.67	0.79	0.62				0.68	0.7	0.58
Tymo_tRNA-like				0.76								
RFN	0.1	0.13	0.19	0.38	0.39		0.34	0.26		0.38	0.4	0.37
S_box	0.51	0.61	0.81	0.44	0.68	0.41	0.64	0.84	0.84	0.58	0.71	0.71
Purine				0.48	0.37	0.35						
Lysine				0.73	0.77	0.87						
glmS				0.75	0.77	0						
Cobalamin				0.49	0.48	0.47						
mean	0.54	0.60	0.65	0.60	0.66	0.50	0.65	0.56	0.84	0.55	0.56	0.55

Table 3.5: CMfinder running time (mins), default mode. The numbers without parentheses refer to the means within each category, while those within the parentheses refer to the standard deviations.

#seq	sim	2000	10000	50000
5	high	3.44 (2.33)	10.96 (6.22)	60.00 (20.45)
5	mix	3.34 (1.76)	12.38 (5.38)	62.64 (21.24)
10	high	13.29 (8.33)	36.92 (12.00)	295.85
10	mix	38.33 (14.01)	45.49 (11.13)	231.27 (78.38)

Table 3.6: CMfinder performance on noisy datasets. "mem sens": the fraction of true RNA instances included in the motif. "mem spec": the fraction of the motif instances that are true. "bp sens" and "bp spec": the sensitivity and specificity of secondary structure prediction at base pair level.

	family	mem sens	mem spec	bp sens	bp spec
5	Cobalamin	1	1	0.43	0.76
	ctRNA_pGA1	1	1	0.97	0.94
	glmS	1	1	0.65	0.78
	Histone3	1	1	1.00	0.62
	Intron_gpII	1	1	0.69	0.78
	IRE	0.4	0.67	0.38	0.24
	let-7	1	1	0.71	0.89
	lin-4	1	1	0.83	0.54
	Lysine	1	1	0.82	0.95
	mir-10	1	1	0.68	0.84
	Purine	1	0.8	0.93	0.37
	RFN	1	1	0.72	0.35
	S_box	1	1	0.57	0.56
	SECIS	0.6	0.75	0.46	0.53
	Tymo_tRNA-like	1	1	0.80	0.77
10	ctRNA_pGA1	1	1	0.93	0.55
	Intron_gpII	1	1	0.86	0.68
	IRE	0.67	1	0.60	0.78
	RFN	1	1	0.60	0.38
	S_box	1	1	0.53	0.58

Table 3.7: Effects of HMM filters and banded constraints. In each category, three methods are listed in the order of **def** default mode, **var1** without HMM filter, **var2** with HMM filter but no HMM band constraints. "sens" and "spec" are the sensitivity and specificity of secondary structure prediction at the base pair level, and "time" is the running time. All numbers are averaged over all datasets within the same category.

#seq	sim	flank	sens			spec			time(mins)		
			def	var1	var2	def	var1	var2	def	var1	var2
5	high	2000	0.69	0.72	0.71	0.54	0.60	0.56	3.45	12.99	7.62
		10000	0.77	0.78	0.78	0.60	0.57	0.59	10.96	25.48	13.71
		50000	0.72	0.74	0.59	0.65	0.59	0.42	55.99	107.00	56.13
	mix	2000	0.74	0.75	0.72	0.60	0.62	0.63	3.34	10.52	4.91
		10000	0.79	0.73	0.75	0.66	0.64	0.64	12.38	34.85	10.73
		25000	0.61	0.61	0.53	0.50	0.51	0.44	62.65	139.46	65.06
10	high	2000	0.81	0.87	0.77	0.65	0.64	0.66	13.30	78.52	27.18
		10000	0.77	0.84	0.85	0.56	0.59	0.59	36.92	209.85	60.13
		50000	0.92	0.91	0.89	0.84	0.84	0.85	295.85	753.05	279.93
	mix	2000	0.70	0.80	0.67	0.55	0.47	0.56	38.34	91.78	67.95
		10000	0.78	0.80	0.84	0.55	0.54	0.57	45.49	190.87	69.47
		50000	0.77	0.70	0.73	0.55	0.48	0.55	231.27	940.81	319.74

Chapter 4

EVALUATING SIGNIFICANCE OF RNA MOTIFS

4.1 Introduction

Application of RNA motif finding tools such as CMfinder to genome scale ncRNA discovery generally produces a large number of predictions. The problem of how to evaluate these predictions and select high quality candidates for further analysis is critical. In the context of comparative sequence analysis, compensatory mutations between species - mutations that maintain base pairs in the putative conserved secondary structure - are among the best evidence for functional ncRNAs. Several algorithms detect ncRNAs in genome alignments based on both secondary structure stability and conservation. Evofold (109) combines two separate, phylogeny-aware probabilistic models: one to describe functional RNAs and the other to describe background genomic regions. Regions that can fold into stable secondary structure, and contain compensatory mutations within the presumed base paired regions are preferred by the RNA model. QRNA uses a similar approach, scanning pairwise alignments for conserved secondary structures based on stochastic context-free grammars and pairwise mutation models (120). AlifoldZ (148) computes averaged folding energy of aligned sequences constrained to a proposed common structure, and infers a Z-score based on a permutation test to estimate statistical significance. RNAz (149), a successor to AlifoldZ, uses machine learning techniques to classify RNAs, and computes significance based on classification scores.

While the above methods all search for conserved secondary structure within the alignments, they quantify the level of structure conservation differently, and so have very different performance characteristics. A recent study applied Evofold, RNAz and AlifoldZ to search for ncRNAs in the ENCODE selected regions of the human genome (150), and found that the overlap between RNAz and EvoFold predictions is quite low. RNAz is sensitive on alignments with moderate to high G+C content and relatively low sequence similarity,

while Evofold is more sensitive in AU rich regions with relatively high sequence similarity. Both Evofold and RNAz have high estimated false positive rates ($> 50\%$) (150) in the ENCODE genome scan.

4.2 A heuristic ranking scheme

To search for an appropriate scoring function for CMfinder motifs, we have tried a few existing methods, all of which are somewhat unsatisfactory for our purpose. For example, RNAz cannot detect ncRNAs with less significant folding energy. It fails to detect the Lysine and SAM-I riboswitches even given perfect alignments. Evofold, on the other hand, produces too many false positives on diverged sequences, giving high scores to spurious alignments of short stem-loops or gappy alignments.

Our initial attempt was to use a heuristic ranking scheme. We first calculated several motif features including secondary structure stability, phylogenetic distribution and sequence conservation pattern. Intuitively, motifs with relatively stable consensus secondary structures and that conserved in many species are good candidates for being functional RNAs. In addition, we noticed that even motifs with low overall sequence conservation contain mosaic patterns of high local sequence conservation. These conserved regions are likely sites of inter- or intramolecular interactions, and hence under strong selection. Therefore, we would like to capture such local sequence conservation to distinguish real RNA motifs from structural motifs that arise by chance.

We used the following motif features in the ranking function, which is used to score CMfinder motifs produced in Bacteria genome scans (See Chapter 5):

- *sp*: the number of different species in which the motif occurs, measured by distinct genome IDs.

- *mc*: average number of motif instances per species. Most riboswitches occur upstream of multiple genes in one species. However, if there are too many motif instances in one species, it is likely to be a repeat element.

- *bp*: the (weighted) number of base pairs in the consensus structure. To discriminate

weak base pairs from stronger ones, we weight the base pairs according to the partition function (68; 97), which estimates the probability of forming a base pair averaged over all possible structures.

- *lc*: local sequence conservation. To measure local sequence conservation, we first identify the conserved columns in the given alignment, defined as the columns with more than 70% sequence identity. Then we locate all blocks with at least 4 consecutive conserved columns, and computed the total size of all such blocks in a given alignment as *lc*.

- *sid*: average pairwise sequence identify. Motifs with high sequence similarity are generally "suspicious", as they are plausibly caused by lack of divergence rather than conservation due to functional importance.

The features bp, lc, sid are computed as the weighted average of all motif instances in a given alignment. The motif instances are weighted based on two criteria. First, CMfinder weights the instances based on their alignment scores, so that poor ones receive low weights. Secondly, we used the Gerstein, Sonnhammer, Chothia algorithm (47) implemented by Infernal (34) to downweight instances with high sequence similarity. The final weight is the product of the two.

The features are integrated in the following ranking function:

$$r = sp \cdot \sqrt{lc \cdot bp/sid} \cdot (1 + \log(mc))$$

We applied the square root and log transformations to prevent a single feature from dominating the overall ranking score. Using this ranking scheme, we have effectively distinguished real ncRNAs in Bacteria genome scans (See Chapter 5). We have also tried to integrate our motif features for scoring by machine learning algorithms including support vector machine (SVM) and logistic regression, but these methods did not perform well. It is nontrivial to apply such machine learning algorithms because of the heterogeneity of the features and positive samples, as well as the difficulty to collect negative samples. In particular, known ncRNAs are highly diversified. It is generally difficult to capture them by a few features

based on our limited knowledge. In addition, some class of ncRNAs have far more examples than others (microRNAs, snoRNAs etc), causing the classifier to bias significantly towards these classes, and to ignore the minorities. Further more, it is difficult to construct negative samples for the classifier that are representative of the background for all applications. It is dangerous to learn a classifier based on a set of negative samples in one scenario while testing it in a different scenario. In this context, we can not foresee all possible applications of this scoring function: motifs could be predicted from bacteria or vertebrate sequences, from conserved regions or nonconserved regions. While it is possible to construct a nice classifier that works well for some applications (e.g. RNAz), it is difficult to make it robust for general purpose.

4.3 A probabilistic ranking scheme

For robustness and interpretability, we prefer a more principled approach that measures the statistical confidence of predicted motifs. Intuitively, we would like to evaluate the likelihood that the motif instances from different species have evolved from their common ancestor with a conserved RNA structure. A probable solution for this problem lies in phylo stochastic context free grammars (phylo-SCFGs), which combine the phylogenetic model's ability to describe evolutionary history, and SCFG's ability to model RNA secondary structure. Phylo-SCFGs have previously been used in Evofold (109), Pfold (80; 81), and a few other studies (110; 79; 111). Evofold has been used extensively in genome scale ncRNA prediction. Its phylo-SCFG has two phylogenetic models, a single-nucleotide model for single-stranded loop and non-structured regions and a base pair model for base paired regions. The overall alignment can be partitioned into structured and non-structured regions, each described by corresponding grammars. Although Evofold has been demonstrated to perform well on conserved alignments (109), we found it inappropriate for scoring CMfinder motifs, many of which have low sequence similarity. In particular, CMfinder sometimes finds structural motifs with a stable hairpin within unrelated sequences. Evofold tends to give very significant scores for such motifs. This may be in part due to the fact that its models are only trained on well conserved alignments, but we believe that a more fundamental problem is its "one size fits all" model for the genomic sequences, which in practice

are a heterogeneous mixture of well- and poorly conserved segments. To address this weakness, we add a third model to describe poorly conserved regions that are under neutral selection or are due to alignment errors. The basic idea underlying our approach is that we test a given alignment under three competing hypotheses: it is a nonconserved region, a sequence conserved region or a structurally conserved region. This approach predicts a structural alignment as a functional RNA if the alignment favors the structural RNA model significantly against both alternatives.

Like all other phylo-SCFG methods, our method has two components: an SCFG and an evolutionary model.

4.3.1 SCFG

As we described in Chapter 1, in the context of RNA secondary structure analysis, SCFGs are probabilistic models that define the landscape of all possible pseudoknot-free structures for sequences. Several SCFGs for RNA structures have been discussed in (28; 80). We have designed a novel SCFG that first partitions the RNA structure into single stranded and base paired regions, then partitions the single stranded regions further into conserved and nonconserved regions. Below is our SCFG grammar :

$$
\begin{array}{lcllll}
S & \rightarrow & L & |T & |F \\
F & \rightarrow & H & |TL & |HL \\
H & \rightarrow & LT & & \\
T & \rightarrow & P & |PT & |PH \\
P & \rightarrow & pP\hat{p} & |pF\hat{p} & |pL\hat{p} \\
L & \rightarrow & cC & |nN & \\
C & \rightarrow & cC & |cN & |\epsilon \\
N & \rightarrow & nC & |nN & |\epsilon
\end{array}
$$

Here, S is the start symbol. The nonterminal symbols used here correspond to particular structural components: L for single-stranded regions, T for structured regions in which both ends are paired (not necessarily with each other), F for structural regions with dangling $5'$ end (described by H) or $3'$ end, and P for stemloops in which both ends are paired with

each other. Single stranded regions L are then further decomposed to conserved C and nonconserved N segments. In the context of the phylo-SCFG, instead of emitting single terminal symbols, C, N generate alignment columns c, n in conserved and nonconserved mode respectively, and P generates pairs of columns $p\hat{p}$. The likelihood of these columns can be computed based on corresponding evolutionary models, and are included in the overall likelihood of the given alignment.

In this SCFG, we do not further discriminate non-structured regions from single stranded regions with structured regions. As stated by the first production rule, a given alignment (corresponding to symbol S) can be all single-stranded or in other words, non-structured (corresponding to L), or contain structured regions, which may or may not be flanked by single-stranded regions (corresponding to F and T respectively).

This grammar is more complicated than other proposed RNA SCFG grammars, because we want to model the transitions between conserved and nonconserved modes within single-stranded regions to capture the typical mosaic conservation pattern in RNA structures. The single-stranded regions are modeled by the last three production rules, which taken together, are equivalent to a phlyo-HMM used in phastCons (133). Given the secondary structure and conservation annotation of an alignment, this grammar is unambiguous.

4.3.2 Evolutionary model

Our evolutionary model follows the general probabilistic framework described in (38), which typically includes an evolutionary tree, a substitution rate matrix, and a vector for equilibrium frequencies. Given the rate matrix R, the substitution matrix for a branch of length t is computed as

$$Q(t) = e^{tR} \tag{4.1}$$

The likelihood of a given column (or a pair of columns for base pair model) is computed via Felsenstein's dynamic programming algorithm using the following recursion:

$$P(D^x|x) = (Q(l_{xy})P(D^y|y)) \cdot (Q(l_{xz})P(D^z|z)) \tag{4.2}$$

where tree nodes y, z are children of x, and l_{xy}, l_{xz} are corresponding tree branch lengths. D^x refers to bases in the column that correspond to the leaves in the subtree rooted at

x. $P(D^x|x)$ is the vector of length equal to the alphabet size (i.e. 4 for single nucleotide model, and 16 for base pair model). The value at the ith position in this vector indicates the probability of D^x, conditioned that the base at node x is the symbol i. The "$\cdot$" operation here returns a vector in which each element is the product of the corresponding elements in the original two vectors. The overall probability of the column is given by

$$P(D|T) = \pi P(D^r|r)$$

where D stands for a given column in the input alignment, T the evolutionary tree, π the equilibrium frequency, and r the root of the tree.

In our phylo-SCFG, each column in the alignment can be in categorized by three different modes: base paired, conserved and nonconserved single stranded. We build an evolutionary model for each mode. The details are described below:

- Evolutionary tree

  We assume that an external evolutionary tree is provided. A few other tools, including Pfold (80) and XRate (79), estimate the evolutionary tree based on the given alignment. We do not use the same approach because the sequence alignments may not truly reflect their evolutionary distance. For example, if irrelevant sequences from two closely related species are forced into an alignment, the evolutionary tree learned from this alignment would be greatly distorted, and based on this tree, we can not predict the correct conservation modes. For application in vertebrate species, we use the species tree learned by phastCons (133) based on the 17 way MULTIZ alignments, which is also used in Evofold. The same evolutionary tree is used for all three evolutionary modes.

- Base pair model for base paired regions. For simplicity, we ignore gaps for the moment. The base pair mutation model is defined by $M^p = (\pi^p, R^p)$, where π^p is a length 16 equilibrium frequency vector, and R^p a 16×16 instantaneous substitution rate matrix. Due to relative large dimensionality of both π^p and Q^p, we design that parameterization scheme with a reasonable number of free parameters. First, we assume that $\pi^p_{ab} = \pi^p_{ba}$. Then we define $R_{ab,cd}$ as the following:

$$R_{ab,cd} = \pi^p_{cd} \cdot \begin{cases} r_{ab,cd} & \text{if ab,cd are both canonical} \\ \gamma_1 & \text{else if either ab or cd is canonical, } a = c \text{ or } b = d \\ \gamma_2 & \text{else if either ab or cd is canonical, } a \neq c \text{ and } b \neq d \\ \beta_1 & \text{else if both ab and cd are non-canonical, } a = c \text{ or } b = d \\ \beta_2 & \text{else if both ab and cd are non-canonical, } a \neq c \text{ and } b \neq d \end{cases} \tag{4.3}$$

The canonical base pairs include AU, GC and UG pairs. We assume two types of symmetry for $r_{ab,cd}$. First, we let $r_{ab,cd} = r_{cd,ab}$, which guarantees the reversibility of the model. Second, we assume that $r_{ab,cd} = r_{ba,dc}$. The second assumption and the symmetric assumption on equilibrium frequencies are based on our observation that reversing the sequence typically has a very minor effect on structural stability. In total, we have 22 parameters, 9 for background frequencies, 9 for canonical substitutions, and 4 for non-canonical substitutions. This parameterization is similar to the base pair model in Evofold, with the exception of our symmetry assumption, which saves 12 free parameters. Our method also differs from Evofold by using 2 more parameters to model non-canonical substitutions, in order to distinguish between substitutions involving one and two mutations, the latter of which we believe are much less likely during evolution. In total, we have 10 fewer parameters than Evofold. We believe that our model is more parsimonious with little sacrifice of descriptive power, leading to more robust estimation. Due to the concern that the vertebrate Rfam motifs we collected here have significant sequence conservation overall, the mutation rate for the base paired regions may be underestimated, which may limit its use to discover more diverged motifs. Therefore, we scale the mutation rate matrix by a constant factor of 2. Our experiments suggests that the scaled model performs equally well for Rfam datasets (based on pvalues), and better in genome scale discovery (based on false discovery rate).

- Conserved single-nucleotide evolutionary model

  For the conserved single-nucleotide model, we choose the commonly used general time-

reversible model:

$$R_{a,b} = \pi_b r_{a,b}$$

and assume that $r_{a,b} = r_{b,a}$. There are 3 free parameters for equilibrium frequencies, and 6 for the instantaneous rate matrix.

- Nonconserved single stranded evolutionary model

  We consider two possibilities that contribute to a poorly conserved region in the alignment: the region is under neutral selection, or a misalignment has occurred. For the first case, we simply scale the evolutionary tree by a constant c, and use the same substitution rate matrix and the same base frequencies as the conserved single-nucleotide model. This is equivalent to scaling the substitution rate matrix by c according to equation 4.1. If a misalignment occurs, we simply compute the column likelihood as the product of frequencies of corresponding bases, ignoring the evolutionary tree. We do not use different states to further discriminate these two cases. To compute column loglikelihood, we simply choose the larger one of these two cases.

- Gap model

  Now, we are ready to consider gaps. There are commonly used methods for handling gaps. The first method treats a gap as an unknown base according to the overall base distribution. This method is used in Pfold and Evofold. Although convenient for likelihood computation, this model is biologically unrealistic. In practice, we found that gappy columns are more likely to be predicted as base paired using this method. An alternative is to treat gap as an extra character. This method is more descriptive than the former, as it can model insertion/deletion (indel) of single nucleotides effectively. It has its own technical issues, however. First, it can not model indels involving multiple nucleotides, and secondly, the concept of equilibrium frequency for indels does not exist. The third method models indels explicitly via a birth-death process as in the TKF91 model (140), and it has been extended to RNA structure (69). This method can handle long indels, and can be used directly for pairwise alignment. However, it is

still unclear how to extend this model to multiple sequence alignment with practical computational cost. Rivas (117) developed a method to extend standard substitution models to include gaps. This method takes a typical instantaneous rate matrix without gaps, then adds gap as an extra character in such a way that the model maintains a relative stationary background distribution among non-gap characters, while the gap frequency evolves towards 1 at time infinity. This process is referred to as quasi-stationary. For a pair of species at a given distance, its equilibrium frequency and substitution matrix satisfies the reversible condition. While this is a more realistic method, we do not know how to maintain its reversible quasi-stationary property for multiple species. On the other hand, we are ready to sacrifice theoretical properties for practical convenience, as long as covarying insertions and deletions of bases pairs can be modeled effectively. We add gap to the base pair instantaneous rate matrix described in equation 4.3 as the following:

$$\begin{aligned}
&R_{ab,a-} = R_{ab,-b} = \alpha_1 \\
&R_{ab,--} = \alpha_2 \\
&R_{a-,b-} = R_{a,b} \\
&R_{a-,--} = \alpha_3 \\
&R_{ab,c-} = 0 \qquad\qquad a \neq c \\
&R_{ab,-c} = 0 \qquad\qquad b \neq c \\
&R_{a-,cd} = R_{-a,cd} = 0 \\
&R_{--,cd} = 0
\end{aligned} \tag{4.4}$$

At the root, we define the background frequency of gaps as 0. Here we describe an evolutionary process that only allows deletions but no insertions. For a given alignment column, gaps can be interpreted as deletions, or alternatively, nongap bases can be treated as insertions. For simplicity and efficiency, before applying phylo-SCFG, we first filter alignment columns that contain more than 50% gaps. In this context, treating all gaps as deletions is reasonable. When $a \neq c$, we set $R_{ab,c-} = 0$ as we believe it involves two events, a substitution and a deletion, so it is not considered in the instantaneous rate matrix. On the other hand, we set $R_{ab,--}$ to a nonzero

value, because due to selection pressure to maintain the structure, the chance of a base pair deletion tends to be greater than losing one nucleotide on one strand, and losing another on the other strand. Pragmatically, it is hard to model such pair deletions as a compound of two independent single deletions such that the resulting event probabilities match the observed frequencies. Similarly, for the single-nucleotide model, we add the gap to the instantaneous rate matrix as the following:

$$\begin{aligned} R_{a,-} &= \alpha_4 \\ R_{-,a} &= 0 \end{aligned} \tag{4.5}$$

Like Rivas's model, our model maintains relatively stable equilibrium frequency of nongap characters, while the gap frequency increases with time.

4.3.3 Training

We train our phylo-SCFG parameters by a maximum likelihood approach. Given a set of structurally annotated alignments, we try to estimate parameter values such that the likelihood of the alignment given the model is maximized. If ignoring the conservation mode, for a given alignment with consensus structure annotation, its likelihood is given by the following formula

$$L(A, S|G, M) = L(S|G)L(A|S, M) \tag{4.6}$$

where A is the alignment, S is consensus structure, G the grammar, and M the evolutionary model. Given the structure annotation, the maximum likelihood estimation of the grammar is independent of the alignment and the evolutionary model, and the estimation of the evolutionary model only relies on the alignment and the structure. Therefore, grammar and evolutionary model can be estimated separately. Alternatively, we can first learn the grammar from structurally annotated alignments, then estimate the evolutionary model that maximizes

$$L(A|G, M) = \sum_S L(S|G)L(A|S, M) \tag{4.7}$$

which considers the uncertainty of secondary structure annotation by integrating all possible structures. This approach has been adopted by Evofold and xrate. We choose to use the fixed structures for training efficiency.

- Training data

  We use annotated Rfam alignments to train the SCFG and evolutionary models. Since we use the 17 vertebrate species phylogeny, we limit training to vertebrate Rfam members only. To obtain a set of structural alignments of homologous ncRNAs, we collected the data from the study by Wang *et al.* (147), which for each seed Rfam member in human, includes matches of the corresponding Rfam Covariance model in all multiple alignment blocks within 10K range of the human member. We take these Rfam CM matches, and align them to the corresponding CM model using the "cmalign" method in the Infernal package (34). If multiple CM matches are found in one species, we choose the one with the highest sequence identity to the human seed member. We produced in total 264 structurally annotated alignments.

- SCFG

  Since the structural alignments do not contain the conservation modes for single stranded regions, we ignored the corresponding production rules for the moment. Similarly, since all training alignments are structured, we can not learn the probability that an alignment is non-structured, so we ignore it as well. Given the consensus structure, the emission probabilities for the columns are fixed, so the reduced grammar becomes the following:

$$
\begin{array}{lcllll}
S & \rightarrow & T & | & F & \\
F & \rightarrow & H & | & TL & | \quad HL \\
H & \rightarrow & LT & & & \\
T & \rightarrow & P & | & PT & | \quad PH \\
P & \rightarrow & < P > & | & < F > & | \quad < L > \\
L & \rightarrow & .L & | & \epsilon &
\end{array}
$$

  where terminal symbols are .,< and >, representing single-nucleotide, and left/right base of a base pair respectively. Since this grammar is deterministic, the production probabilities can be computed based on the number of times each production rule has been used.

We set the remaining parameters manually, and tuned based on training data. We varied the probabilities for $S \rightarrow L$ in the range 0.9 - 0.999, and found that the results are very robust except for very short motifs. We also do not have appropriate training data to learn the transitions parameters between the two conservation modes, as the Rfam alignments for vertebrates are in general strongly conserved, so a very small fraction of the alignment columns fit the nonconserved mode. This situation, however, is likely to change as we extend the search for ncRNAs into less conserved regions. We set these parameters rather arbitrarily, while taking account of coverage constraints and smoothness constraints as discussed in (133), so that the fraction of conserved and nonconserved columns are reasonable and isolated conserved or nonconserved blocks with only a couple of columns are avoided.

- evolutionary model

  To learn the base pair model, we extracted base paired columns from the Rfam alignments as training data. The maximum likelihood estimation of base pair parameters are found using the BFGS quasi-Newton algorithm (169).

  We use the same approach to learn the conserved single-nucleotide model, using all columns of the Rfam alignments, except for the ones annotated as gapped. Although the Rfam alignments also contain columns with weak sequence conservation, they are generally highly conserved. For the nonconserved model, we set the scaling factor $c = 5$, i.e., the mutation rate is 5 times faster for the nonconserved mode than the conserved mode. Although this ratio is arbitrarily selected, we set the value large enough to make the distinction between two modes significant, and to minimize potential ambiguity.

4.3.4 Scoring

We have tried several strategies for scoring motifs. First, we use $-\log(P(S \rightarrow L))$, i.e. the negative log of the probability that there is no structured region in the alignment. This scoring scheme considers all possible structures, which together contribute to $1-P(S \rightarrow L)$, the probability that the alignment contains a structured region. We refer to this score as the *totRNA* score. To evaluate the quality of a given structural annotation, we score each

annotated base pair by $\log(1 - P(pair(i, j))$, where $P(pair(i, j))$ is the posterior probability that columns i, j are paired by summing over all parses that emit a pair at i, j. The score for an overall structure is the sum of scores for all its base pairs. Note that this score does not correspond to any probability, because the posterior probabilities of base pairs are not independent. Compared to the alternative using the posterior probability of the given whole structure, this score is robust to partial prediction errors: if a predicted base pair has very low posterior probability, its score is near zero and ignored. The overall score is dominated by a set of high quality base pairs. We refer to this score as the *pair* score. Occasionally, some pairs of columns share great covariation by chance, without support from corresponding folding energy. Such pairs may still receive significant posterior probabilities. To take this issue into account, we multiply the emission probabilities of base paired columns by the values of corresponding partition function (97). We define the partition function for a pair of columns as the geometric mean of the partition functions of all sequences at given positions. To limit the effect of outliers with near zero values, we set all values below 0.05 to 0.05. With addition of the partition function, this analog of the *totRNA* score is referred to as the *par-totRNA* score, and the analog of the *pair* score is referred to as the *par-pair* score.

4.4 Results

4.4.1 Heuristic ranking function

The heuristic ranking function has been applied to ncRNAs prediction in bacteria, and the results are discussed in Chapter 5. A variation of this function was used in ncRNA prediction in the vertebrate ENCODE regions (142), and the results are discussed in Chapter 6.

4.4.2 Probabilistic ranking function

Training Results

Based on the parameters learned by parsing structurally annotated Rfam alignments and other manually determined parameters, our SCFG grammar is determined as follows:

S	$\rightarrow$	L	0.900	$\mid$	T	0.016	$\mid$	F	0.084
F	$\rightarrow$	H	0.147	$\mid$	TL	0.179	$\mid$	HL	0.674
H	$\rightarrow$	LT	1.000						
T	$\rightarrow$	P	0.826	$\mid$	PT	0.011	$\mid$	PH	0.163
P	$\rightarrow$	$pP\hat{p}$	0.794	$\mid$	$pF\hat{p}$	0.130	$\mid$	$pL\hat{p}$	0.076
L	$\rightarrow$	cC	0.950	$\mid$	nN	0.050			
C	$\rightarrow$	cC	0.840	$\mid$	cN	0.008	$\mid$	ϵ	0.152
N	$\rightarrow$	nC	0.008	$\mid$	nN	0.840	$\mid$	ϵ	0.152

The trained single nucleotide model (conserved) and base pair model are shown in Table 4.1 and 4.2. The single nucleotide model we learned is quite similar to the Pfold (80) unpaired model, although at different scale. Both show distinctive differences between transitions and transversion rates. We also notice a higher G + U frequency at 52.7%, which agrees with the observation by Green *et al.* (53) that there is strand bias with an excess of G+U on the coding strand of transcripts, which presumably reflects biases intrinsic to mutation and repair mechanisms. The ratio between phastCons's (133) rate matrix trained on the fourfold degenerate sites (4d) and the rate matrix of this model is approximately 3, compared to the ratio of 5 for our nonconserved mode. It indicates that our nonconserved mode supports higher mutation rate than the rate specified by phastCon for regions under neutral selection.

The base pair model also shows significant similarity to Pfold's rate matrix, and at similar scale. Note that the mutation rates between non-canonical and canonical base pairs differ significantly depending on whether one or two mutations are involved. For example, mutations from AA to AU are about 5 times more likely than the mutations from AA to GC. This difference justifies our choice to use two parameters to capture the mutation rates between non-canonical and canonical base pairs. Also note that the rate for a single deletion α_1 is approximately the same as those for double deletions α_2, while treating each gap independently would doubly penalize deletion of a base pair. Therefore, covariant deletions tend to favor the base pair model for base pairs. This is consistent with the empirical criteria for ncRNA candidate selection in bacteria (152) (see Chapter 5 for details).

Table 4.1: Single-nucleotide evolutionary Model (conserved)

	A	C	G	U
Freq	0.238	0.235	0.261	0.266
Rate Matrix				
A	-0.395	0.076	0.229	0.090
C	0.077	-0.397	0.069	0.250
G	0.209	0.062	-0.348	0.077
U	0.081	0.222	0.075	-0.378
Gap $\alpha_4 = 0.057$				

Testing results

We tested the performance of different variants of our scoring functions, and compared with Evofold and RNAz. We tried to replicate the Evofold scoring function (110) by using the same control file, and the same evolutionary tree, and taking the following steps as suggested by Jakob Pedersen (personal communications): first parsing the consensus structure annotation for a given alignment by the CYK algorithm, then trimming the structured regions, scoring each region by the log of the posterior likelihood ratio of the region under the structural model vs. the non-structural model, and finally normalizing the score by length of the region and scaling by 10. RNAz works the best for datasets with equal or fewer than 6 sequences, so if a given alignment contains more than 6 sequences, we used the rnazSelectSeqs.pl script provided in the RNAz package to select up to 6 sequences. We ran RNAz under the default setting.

1. Test on RFAM alignments

   We first took the 264 RFAM alignments used as training data, which include 22 tRNAs, 171 snRNAs (primarily C/D box or H/ACA box snoRNAs), 29 miRNAs, 25 *cis*–regulatory elements, and others. We shuffled them each 100 times while maintaining the approximate gap pattern, local conservation pattern and base pair frequencies

Table 4.2: Base pair Evolutionary Model

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
Freq	0.0081	0.0160	0.0091	0.1554	0.0160	0.0084	0.2315	0.0098	0.0091	0.2315	0.0093	0.0596	0.1554	0.0098	0.0596	0.0114
Rate Matrix																
AA	-2.6469	0.2967	0.1677	0.6306	0.2967	0.0191	0.1380	0.0221	0.1677	0.1380	0.0210	0.0355	0.6306	0.0221	0.0355	0.0258
AC	0.1503	-2.6528	0.1677	0.6306	0.0362	0.1564	0.1380	0.0221	0.0205	0.9395	0.0210	0.0355	0.0926	0.1810	0.0355	0.0258
AG	0.1503	0.2967	-2.8426	0.6306	0.0362	0.0191	0.9395	0.0221	0.0205	0.1380	0.1719	0.0355	0.0926	0.0221	0.2417	0.0258
AU	0.0329	0.0650	0.0367	-0.4806	0.0095	0.0050	0.0075	0.0397	0.0054	0.0351	0.0055	0.1618	0.0208	0.0058	0.0034	0.0464
CA	0.1503	0.0362	0.0205	0.0926	-2.6528	0.1564	0.9395	0.1810	0.1677	0.1380	0.0210	0.0355	0.6306	0.0221	0.0355	0.0258
CC	0.0184	0.2967	0.0205	0.0926	0.2967	-3.1968	0.9395	0.1810	0.0205	0.9395	0.0210	0.0355	0.0926	0.1810	0.0355	0.0258
CG	0.0048	0.0095	0.0367	0.0051	0.0650	0.0343	-0.4204	0.0397	0.0054	0.0064	0.0377	0.0034	0.0235	0.0058	0.1364	0.0068
CU	0.0184	0.0362	0.0205	0.6306	0.2967	0.1564	0.9395	-2.8812	0.0205	0.1380	0.0210	0.2417	0.0926	0.0221	0.0355	0.2116
GA	0.1503	0.0362	0.0205	0.0926	0.2967	0.0191	0.1380	0.0221	-2.8426	0.9395	0.1719	0.2417	0.6306	0.0221	0.0355	0.0258
GC	0.0048	0.0650	0.0054	0.0235	0.0095	0.0343	0.0064	0.0058	0.0367	-0.4171	0.0377	0.1364	0.0051	0.0397	0.0001	0.0068
GG	0.0184	0.0362	0.1677	0.0926	0.0362	0.0191	0.9395	0.0221	0.1677	0.9395	-3.0629	0.2417	0.0926	0.0221	0.2417	0.0258
GU	0.0048	0.0095	0.0054	0.4223	0.0095	0.0050	0.0002	0.0397	0.0367	0.5301	0.0377	-1.1687	0.0088	0.0058	0.0067	0.0464
UA	0.0329	0.0095	0.0054	0.0208	0.0650	0.0050	0.0351	0.0058	0.0367	0.0075	0.0055	0.0034	-0.4806	0.0397	0.1618	0.0464
UC	0.0184	0.2967	0.0205	0.0926	0.0362	0.1564	0.1380	0.0221	0.0205	0.9395	0.0210	0.0355	0.6306	-2.8812	0.2417	0.2116
UG	0.0048	0.0095	0.0367	0.0088	0.0095	0.0050	0.5301	0.0058	0.0054	0.0002	0.0377	0.0067	0.4223	0.0397	-1.1687	0.0464
UU	0.0184	0.0362	0.0205	0.6306	0.0362	0.0191	0.1380	0.1810	0.0205	0.1380	0.0210	0.2417	0.6306	0.1810	0.2417	-2.5544

α_1	α_2	α_3
0.0210	0.0196	0.1319

(P. Anandam, E. Torarinssion and W.L. Ruzzo, in preparation). Then we score each shuffled alignment. Since the structures are not given for these alignments, we inferred the structures by using the CYK algorithms. For each RFAM alignment, we computed its empirical p value as the fraction of the shuffled alignments with equal or higher scores.

We refer to our probabilistic scoring function as **pscore**. Figure 4.1 shows the distribution of pvalues of 4 variants of pscore. The y-axis shows the pvalues, and x-axis shows the corresponding pvalue ranks. It demonstrates that addition of the partition function significantly improves the p values by penalizing the structures with poor stability. We found that the "par-pair" has the best performance, so it becomes the method of choice for the following analysis. In the following analysis, pscore refers to the par-pair score in this analysis.

Next, we compare the CDFs of pscore Evofold, and RNAz (see Figure 4.2). RNAz and pscore share similar performance, both significantly outperform Evofold. We also plot the score ranks against the p values for pscore, Evofold and RNAz in Figure 4.3, coloring the points based on the corresponding RNA types. All three schemes shares the common trend that alignments with higher scores have more significant p values. Note that for all three methods, there are a considerable number of motifs that have score zero, and pvalue one, and they all collapse to one data point in this plot. The highest scoring motifs for all three methods are tRNAs, which are deeply conserved with a significant amount of covariation. The lowest scoring motifs for all three methods include C/D-box snoRNAs, which have highly conserved sequences but weak secondary structure. Another major sub-type of snRNAs, H/ACA-box motifs also have very strong sequence conservation, but with much more compact secondary structure. Some H/ACA-box motifs have considerable number of non-canonical base pairs, resulting in low scores for all three methods. For other H/ACA-box motifs, pscore and RNAz give pretty high scores due to their highly stable structures, while for Evofold, favorable structures are not sufficient to compensate for lack of covariation. MiRNAs have reasonable scores for all three methods, although, the pvalues

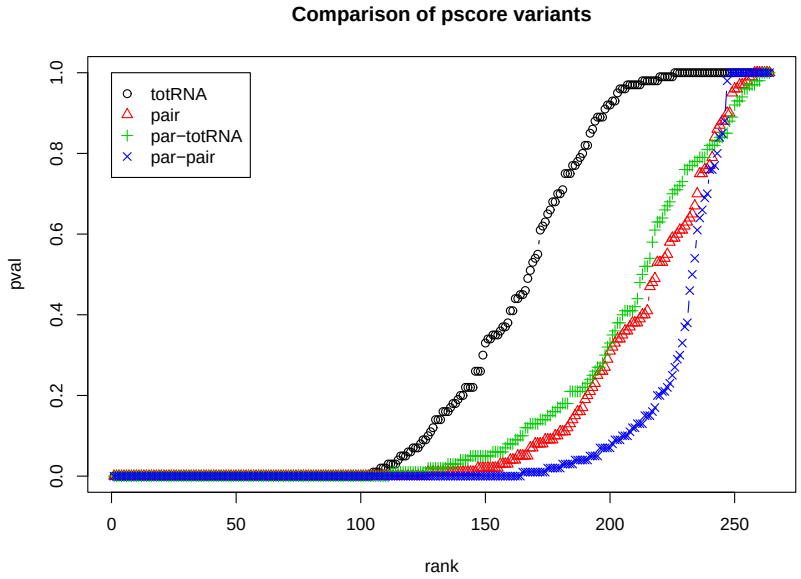


Figure 4.1: Pvalue distribution for pscore variants. x axis:pvalue rank, y axis:pvalue

for Evofold are less significant than the other motifs, suggesting that many shuffled motifs also have good Evofold scores. Closer examination suggests that many shuffled alignments with high Evofold scores either are too short, contain many gaps, or have weak structures. Preprocessing or postprocessing filters based on prediction length, gap frequency and structure compactness may eliminate a significant fraction of such false predictions. On the other hand, the performance of such filters is very sensitive to the threshold values. Short *cis*–regulatory elements such as IRE and HISTONE3 may not pass such filters. Therefore, although the performance of Evofold can be improved by additional processing steps, we still think it is necessary to calibrate the

effects of motif length, gaps, and structures inherently by the scoring method itself.

Next, we compared the scoring of base pairs versus Rfam annotated base pairs. For each alignment, we selected the base pairs with posterior probability greater than 0.5 as threshold, and counted the number of such base pairs that are annotated based on Rfam, and the number of not annotated. There are 91 alignments that do not have any high scoring base pairs. Most of these alignments are C/D-box snoRNAs. For the rest, the average false positive rate is 17.8%.

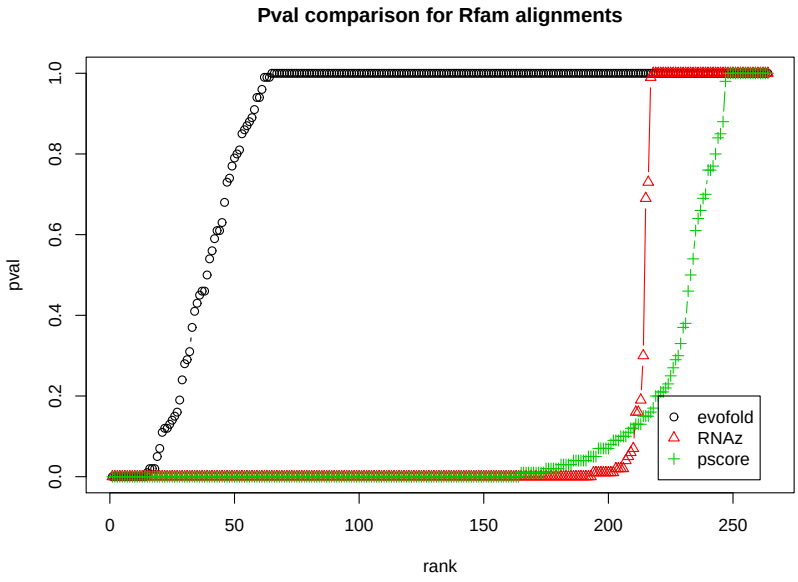


Figure 4.2: Pvalue distributions of Evofold, RNAz and pscore. x axis:pvalue rank, y axis:pvalue

Our analysis shows that pscore has comparable performance to RNAz, and both out-

perform Evofold in this context. However, we used the same test dataset for training, which may cause significant bias towards pscore. To correct this potential bias, we perform 5-fold cross validation, using 80% of the dataset for training, and the remaining 20% for testing. Figure 4.4 shows that the scores from cross validation of 264 Rfam motifs are almost identical to the scores using all training data. Close examination of model parameters indicates that their values have very small variation (standard deviations are within approximately 5% of the absolute values of means).

As shown in Figure 4.3, we can classify the Rfam motifs collected in this experiment into five categories. Except for "cis-regulatory", the motifs within each category share a common structural pattern. To test whether our model is robust regarding different types of RNAs, we perform cross validation by RNA types by using all the motifs of a particular type as testing data, and the rest as training data. As shown in Figure 4.5, all C/D box motifs have slightly higher scores, while miRNAs and H/ACA-box motifs have slightly lower scores by cross validation. We do not know exactly what causes such systematic bias yet, but the general trend indicates that the scores by this cross validation agree significantly with scores based on all training data.

2. Test on annotated snoRNA and miRNAs in vertebrates

   We collected a set of 991 human snoRNAs and miRNAs from the UCSC browser and retrieved the MULTIZ 17-way alignments of these ncRNAs. The alignment blocks were concatenated using the Galaxy tool (48). To cope with possible annotation errors, we extended the alignments by 20 bases on both strands. The consensus structures of many of these snoRNAs across species are not known, so we use the given alignments as input to CMfinder. CMfinder attempts to improve the alignment and predict consensus structure. After eliminating alignments with few than 3 sequences, 834 alignments remained, including 210 C/Dbox, 96 H/ACAbox, 511 miRNA and 17 scaRNA. We computed the pscore of all alignments. The pscore distribution stratified by the RNA types is shown in Figure 4.6.

   The distribution of pscore of miRNA, H/ACA, and scaRNAs are quite similar, all

significantly higher than C/D box. This is what we would expect based on previous discussion. We are a little surprised that 188 miRNAs score below 10. Closer examination reveals that these low scores are caused by primarily a few reasons. Some miRNAs are only conserved in primates and other close species with little covariation. Some other miRNAs are conserved in many species, but there is no covariation at all in the stems. A third class of such miRNAs also seem deeply conserved, but the structures are not conserved in diverged species due to conflicting base pairs in the stems, or missing big fragments of the structures. This can be caused by miRNAs gaining or losing function during evolution, missing sequences in the genome assemblies, or genome alignment errors. After removing the motif instances that do not conserve the structure well, the pscores usually improve significantly. To evaluate the extent of such effect, we selected 302 miRNAs with more than three motif instances, and pscores smaller than 40. For each of these motifs, we removed up to 4 lowest scoring motif instances, re-scored each filtered motif, and kept the highest scoring one if it was better than the original score. We found that 243 of 302 miRNAs have improved scores after this procedure, with average improvement of 16, and 78 of these 243 miRNAs then score above 40.

For the motifs that have low score due to lack of covariation, we found that if the unscaled base pair model (the default scaling factor is 2) is used, their scores tend to improve significantly. To evaluate this effect, we selected 202 motifs with pscores smaller than 10, and average pairwise sequence similarity greater than 90%, and rescored them using the unscaled base pair model. 176 of these motifs have improved scores, and their average improvement is 19, with 84 scoring above 40. Tests on the shuffled alignments suggest that unscaled and the original scores have very similar distribution.

3. Test on motifs predicted in the ENCODE regions.

   Next, we would like to investigate the performance of the scoring function on CMfinder motifs in a more realistic scenario. We have applied CMfinder to genome scale ncRNAs discovery in the vertebrate ENCODE regions (see Chapter 6 for details). Unlike the RFAM motifs, which are generally strongly conserved, many motifs predicted in

this experiment have poor sequence similarity. We then used pscore, Evofold and RNAz to evaluate the significance of predicted motifs. Note that Evofold and RNAz were applied to CMfinder motifs, rather than the original multiple alignment blocks, however, both ignore the secondary structure annotations of the motifs provided by CMfinder.

Because we did not know which motifs are real ncRNAs, we replied on false discovery rate (FDR) (10) to measure the capacity of these scoring methods to discriminate real motifs from ones occurring by chance. Our goal was to estimate for a given score threshold, the fraction of motifs with scores above that threshold that can be expected by chance. To do this, we shuffled the columns of the alignment blocks using a method included in the RNAz package (149), which besides keeping the same base composition and global sequence similarity, also maintains the approximate local conservation and gap pattern. It would be best to shuffle each input alignments multiple times, but in the interest of reducing computation time, we shuffled each only one time. We then applied CMfinder to the shuffled alignments, and re-scored them. Because each dataset may produce multiple motifs, we selected the highest score of all motifs using each of the three methods. We define the empirical FDR at a given threshold t as

$$FDR(t) = \min(\frac{S(t)}{R(t)}, 1)$$

where $S(t)$ is the number of motifs with scores greater than t in the original dataset and $S(t)$ is the corresponding number in the shuffled dataset. For each method, we ignored part of the FDR curve when the threshold $t = 0$. The comparison of FDR for all 3 methods are shown in figure 4.7. To plot the FDR, instead of using the score threshold values as X axis, whose range varies significant among the three methods, we use rank of motif scores in the original dataset, and compute the FDR values based on the corresponding motif scores as threshold values.

This plot clearly demonstrates that pscore has more favorable FDR than RNAz and Evofold, and higher pscores generally correlate to smaller FDRs. For the top 100 motifs, the FDR for pscore is 30% or less, while for both RNAz and Evofold, the

FDRs of the 100th motif are over 50%. While the estimated FDRs are still not as significant as we wish, considering the fact that our permuted alignments are very similar to the original alignments, and CMfinder is good at optimizing structural homology, it is not surprising that some motifs from permuted alignments also look promising.

Figure 4.8 shows the score distribution for all motifs from the shuffled datasets. Since both pscore and Evofold are correlated with log likelihood, we transform the probabilities produced by RNAz to log likelihood format. The background score distributions for all three methods are in fact quite similar, although the values are at different scale. These scores (now all at log scale) in general have approximate linear correlation with log of score ranks except at the tails, which suggests that after careful calibration, they can serve as rough estimates of p values.

To measure whether each method has systematic bias with respect to certain motif features we plotted the scores of all motifs from the shuffled datasets against motif length, GC content, the number of sequences and average sequence similarity for all three methods(see Figure 4.9, 4.10, 4.11). We first discretized all motif features, and used box-and-whisker plots to display the distribution of scores for a given factor value. To reduce clutter in the figures, we did not plot the outliers. Note that the scales of the scores are different for each method, so the y axis values are not directly comparable. For RNAz, we can see that the score does not correlate with motif length. There is some bias for extremely AU rich or GC rich regions, motifs with very low sequence similarity, and motifs with a small number of instances. For Evofold, there is strong bias for short motifs, and motifs with low sequence similarity (note that 95th percentile of Evofold scores for the unshuffled ENCODE motifs is 2). For pscore, the median scores remain rather constant across the spectrum, although there is still some bias for GC rich regions. These observations agree with suggestions by the authors of RNAz and Evofold that these tools should not be used for motifs that are too short, with significant composition bias or too few sequences.

Table 4.3: Pscore variants tested on the ENCODE motifs

method	partition function	nonconserved mode	scale factor for base pair model
pscore(par-pair)	Yes	Yes	2
var1	No	Yes	2
var2	Yes	No	2
var3	No	No	2
var4	Yes	Yes	1

4. Comparison of pscore variants on the ENCODE motifs.

   Next, we would like to evaluate different variants of pscore on the motifs predicted in the ENCODE regions. The explanation of the pscore variants are given in Table 4.3.

   Figure 4.12 shows the FDR curves of these variants. To our surprise, the pscore variants behaved very similarly in this experiment. We would expect var1, which does not use the partition function, to produce more false positives. After closer examination, however, we realized that due to the fact that all pscore variants use the CMfinder structural annotation, all predicted base pairs are supported by the energy model. In fact, for var1, the pair scores based on the CMfinder structures are often much lower than corresponding ones based on the optimal CYK annotation. Therefore, the use of CMfinder annotation prevents inflation of scores by spurious base pairs. When structural annotation is not available, use of the partition function is especially important. The elimination of the nonconserved mode seems to have limited effect on the FDR. We found that although this model tends to assign higher scores to poorly conserved motifs, our base pair evolutionary model assumes a reasonable amount of sequence conservation in the stem regions, so it still favors the motifs that are reliably aligned over those aligned by chance. In addition, due to the selection process of CMfinder motifs, motifs that are aligned by chance are chosen only if there are no better motifs in the same dataset, so they contribute to only a small fraction of

the collection. The var4 with unscaled base pair model has significantly worse FDR than other variants, due to its preference of strongly conserved motifs. The shuffled alignments of these motifs still tend to score high, making it difficult to discriminate real ncRNA signals.

4.5 Discussion

In this study, we presented a new RNA motif scoring scheme based on a phylo-SCFG model. Compared to Evofold, we have revised the parameterization of evolutionary models to better categorize structural alignments with non-canonical base pairs and gaps. We also incorporated folding energy into the model by use of the partition function. Finally, we modeled the nonstructured regions as a mixture of conserved and nonconserved segments, so that structure predictions are scrutinized with respect to more alternative hypotheses.

Our results demonstrate that this method performs robustly on motifs with heterogeneous characteristics, measured by the pvalues in Rfam datasets and false discovery rate in the ENCODE dataset. Due to the revised evolutionary models, the new scoring scheme is robust to the presence of gaps and local alignment errors. Incorporation of the partition function enable the base pairs that agree with the energy model to be favored. We also show that this scoring scheme has no obvious systematic biases to alignment features such is alignment size (the number and lengths of sequences), sequence similarity and GC content, which have no direct bearing to ncRNA signals. Overall, this scoring scheme makes fewer assumptions about the characteristics of the motifs, thus does not require additional processing of the input.

This method only annotates the significant base pairs. The goal is to identify strong evidence of RNA signals, so it is not necessarily the right tool to predict the whole structure, a function which we assume has been performed by other tools such as CMfinder. It also annotates the segments in the single stranded regions that are strongly conserved, which is a good indication of how reliable the alignments are.

Like Evofold, our scoring scheme is very sensitive to spurious motif instances in the alignment. Removing such instances from analysis is important to obtain the correct scores. CMfinder provides a CM score for each motif, which measures the homology to the consen-

sus, therefore it can be used as a reasonable criterion for filtering. However, sometimes it can be tricky to determine the correct threshold. It is especially difficult to evaluate the motif instances with partial homology with the consensus, which suggests that they probably evolved from a common ancestor, while conflicts in the structure or other sequence regions may suggest loss of function in corresponding lineages, or alternatively, the predicted structure is not functional. As we do not know which the case is here, we can not conclude whether these motif instances should be included. This is an issue that requires further investigation.

Tests on the ENCODE motifs suggest that our scoring scheme has the nice property that the FDR decreases as the motif score increases, and it is superior to both RNAz and Evofold. (Note, however, that we haven't performed the postprocessing steps for RNAz and Evofold as recommended by previous studies.)On the other hand, FDR is not entirely satisfactory yet. After manually reviewing the top scoring motifs in the shuffled datasets, we think these shuffled motifs fit our selection criteria as good ncRNA candidates, in that they are reliably aligned, deeply conserved, structurally stable and contain covarying base pairs. It is possible that the shuffling procedure we used did not sufficiently remove the secondary structure signals in these alignments, as reported by previous studies using the same shuffling procedure (151; 150; 121), or alternatively, we do not have sufficient evidence to discriminate real ncRNAs from those that arise by chance.

We have shown that the parameters of our evolutionary models and SCFG are robust to the training datasets by cross validation. Among the manually set parameters, we found that the SCFG mixture parameters have small effects on scores, while the mutation rates on the nonconserved mode and base pair model play important roles. Small ratio of mutation rates between base pair and conserved single nucleotide favors prediction of highly conserved ncRNAs, while small ratio of mutation rates between base pair and nonconserved single nucleotide biases towards prediction of poorly conserved ncRNAs. It is still unclear to us, what should be the correct preference list for ncRNAs in general. On the other hand, understanding the behavior of these rate parameters enables us to adjust the scoring scheme to select specific type of RNAs that are of interest.

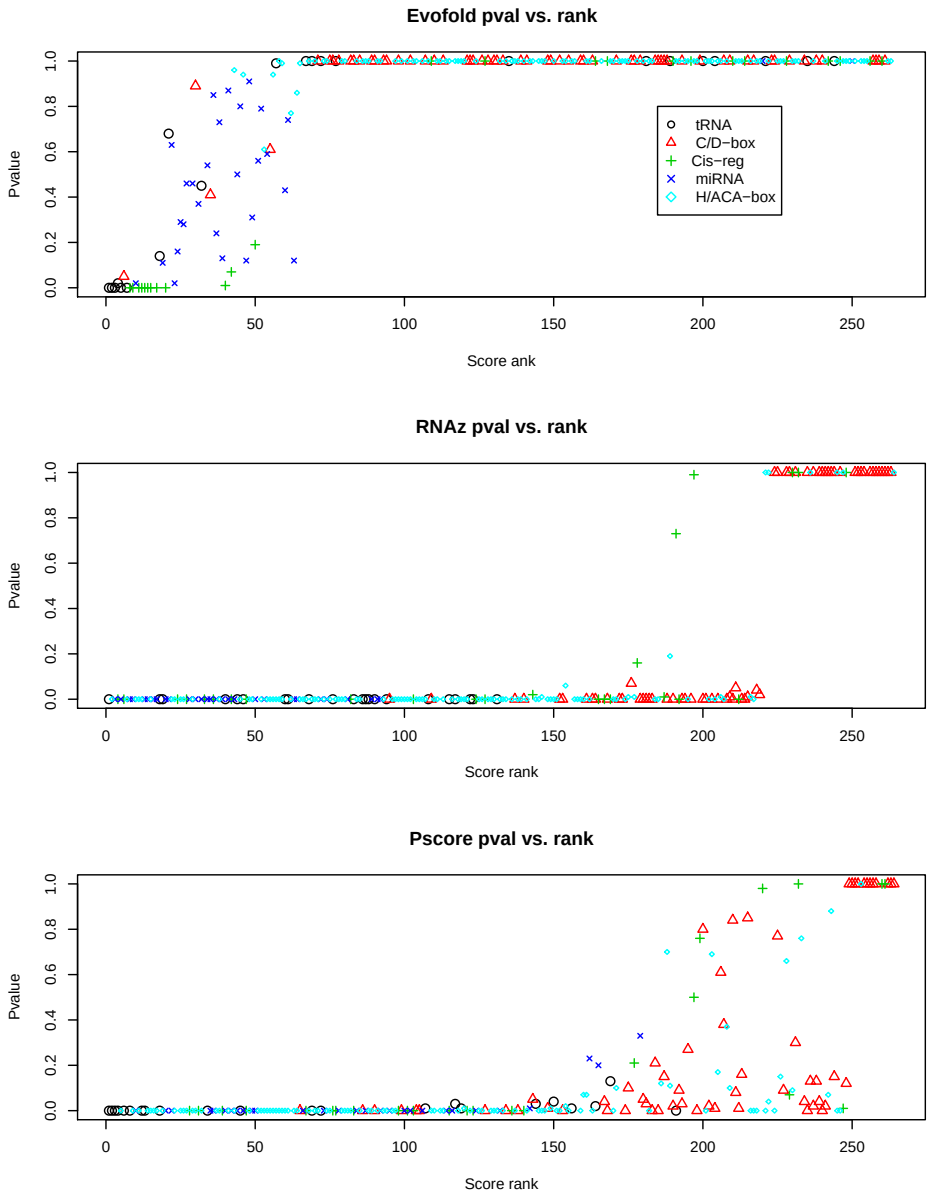


Figure 4.3: Pvalue vs. score rank for Evofold, RNAz and pscore.

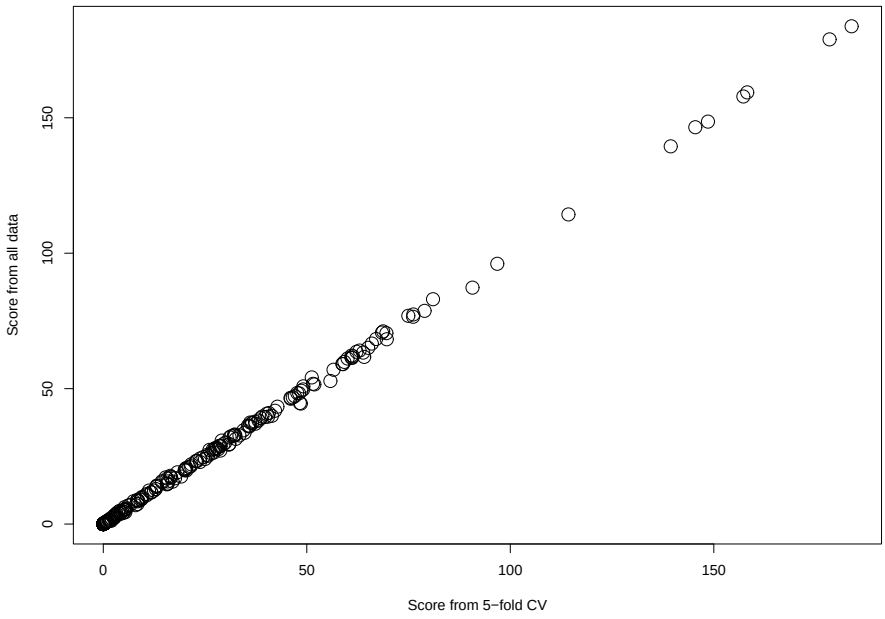


Figure 4.4: Pscore cross validation scores vs. scores from all training data

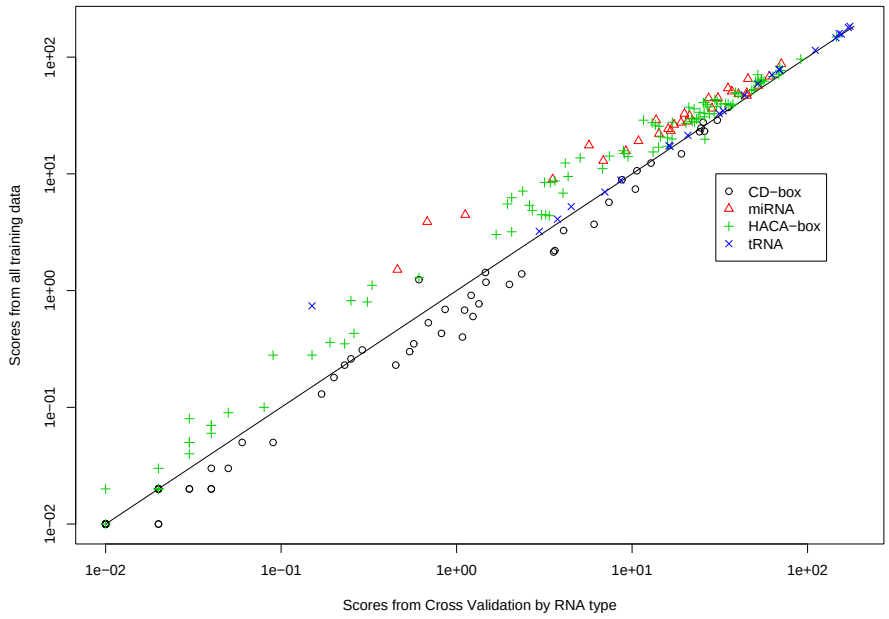


Figure 4.5: Pscore cross validation scores by RNA type vs. scores from all training data

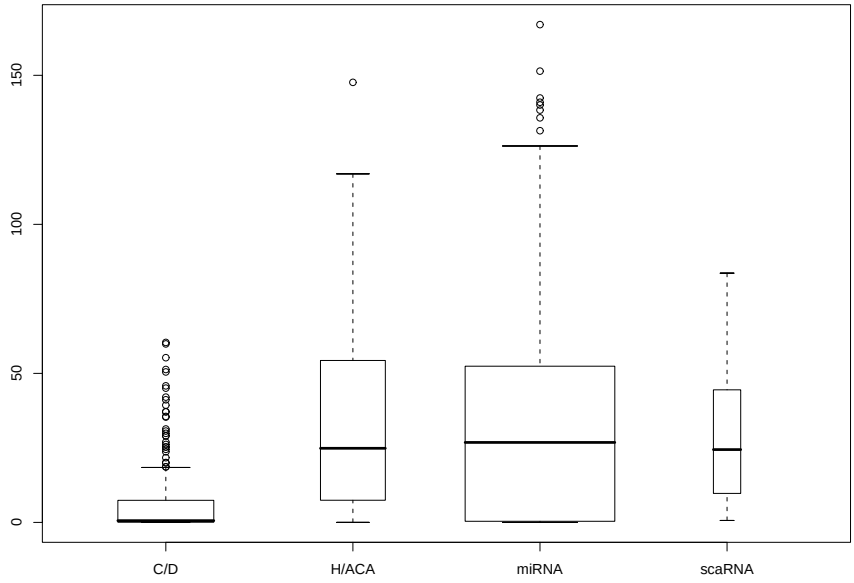


Figure 4.6: Pscore distribution of snoRNAs, miRNAs and scaRNAs in vertebrates, stratified by RNA type. The boxes are drawn with widths proportional to the square-roots of the number of observations in the groups. The box corresponds to 25th and 75th percentiles. The whisker extends to 1.5 times the interquantile range (IQR) below the 25th and above the 75th percentiles.

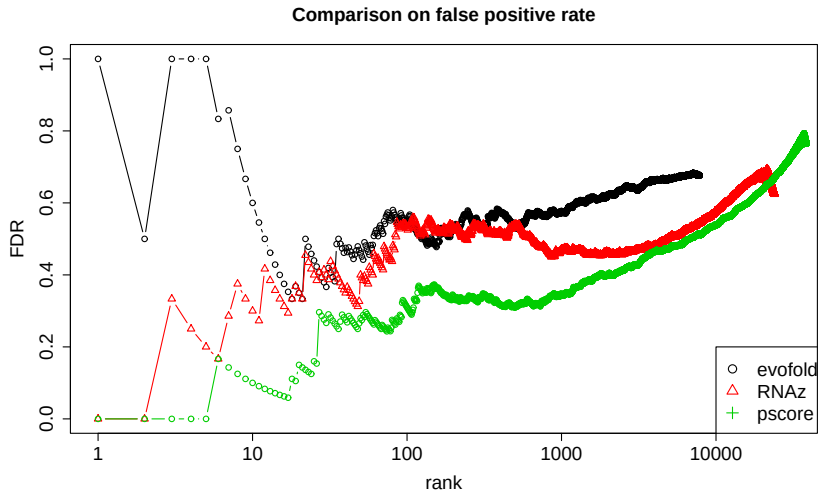
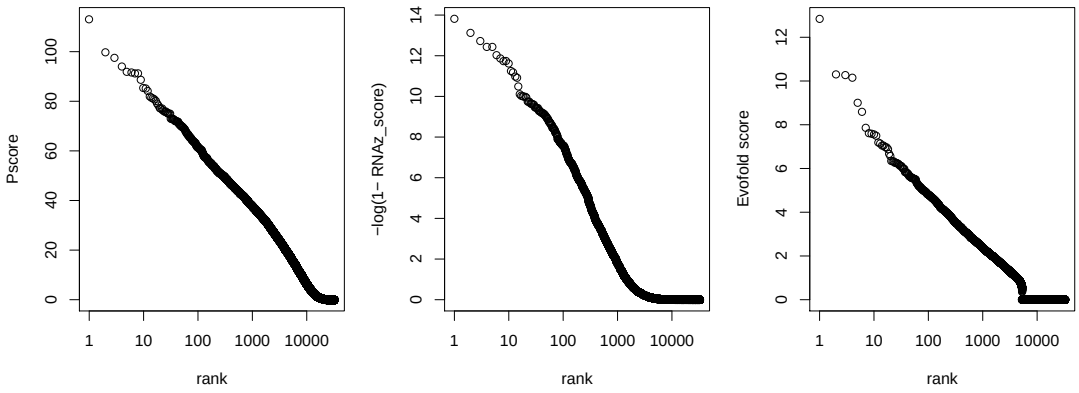


Figure 4.7: FDR of pscore, RNAz and Evofold on CMfinder motifs predicted within the ENCODE regions. The x axis shows the score ranks of the motifs in the original dataset, and y axix is the corresponding FDR. Scores of 0 are ignored. Note that the top ranking motifs are different for different scoring methods.



(a) pscore (b) RNAz (c) Evofold

Figure 4.8: Score Distribution of pscore, RNAz and Evofold on motifs from shuffled datasets. For each dataset, only the highest score of all motifs is included.

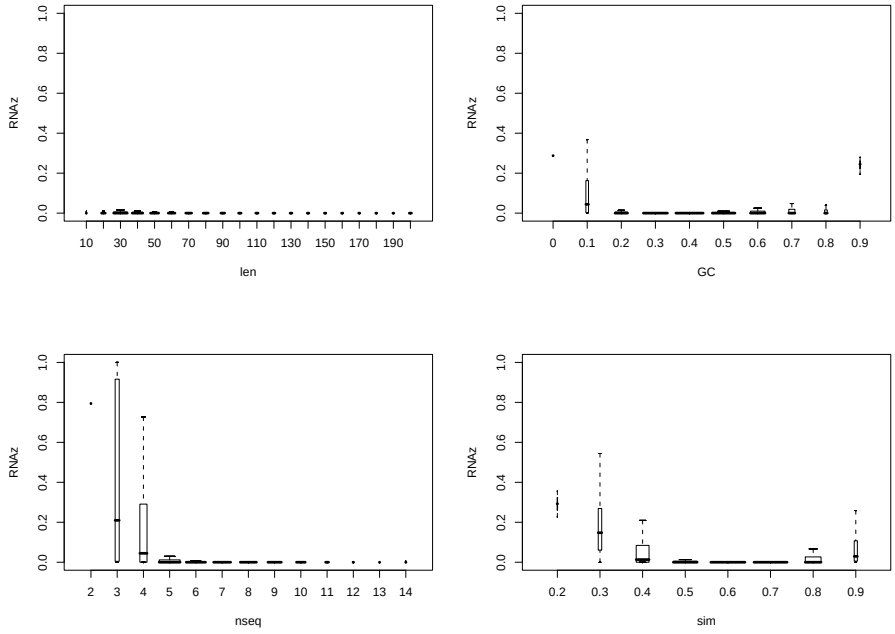


Figure 4.9: RNAz scores vs. motif features. The boxes are drawn with widths proportional to the square-roots of the number of observations in the groups. The box corresponds to 25th and 75th percentiles. The whisker extends to 1.5 times the interquantile range(IQR) below the 25th and above the 75th percentiles.

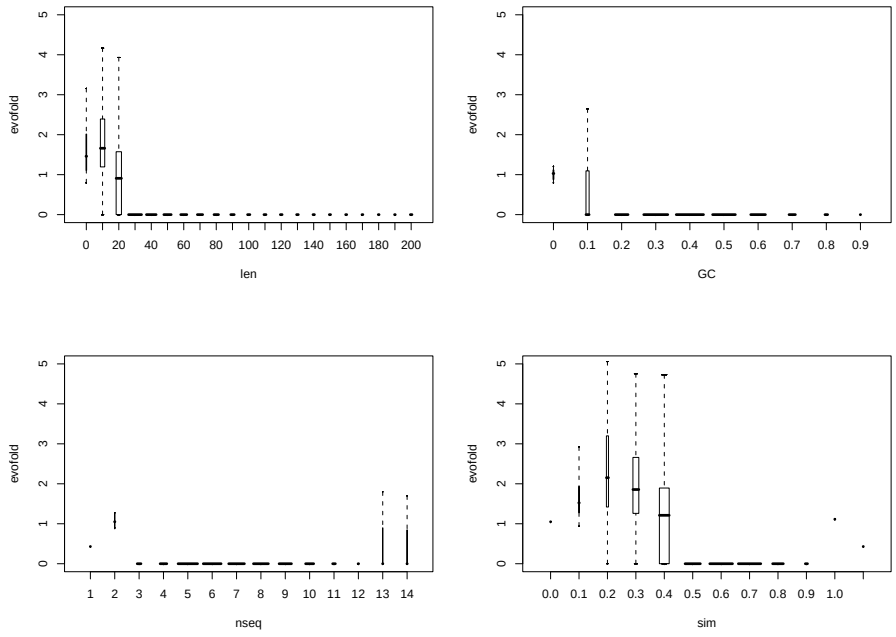


Figure 4.10: Evofold scores vs. motif features. Plot setting is the same as figure 4.9

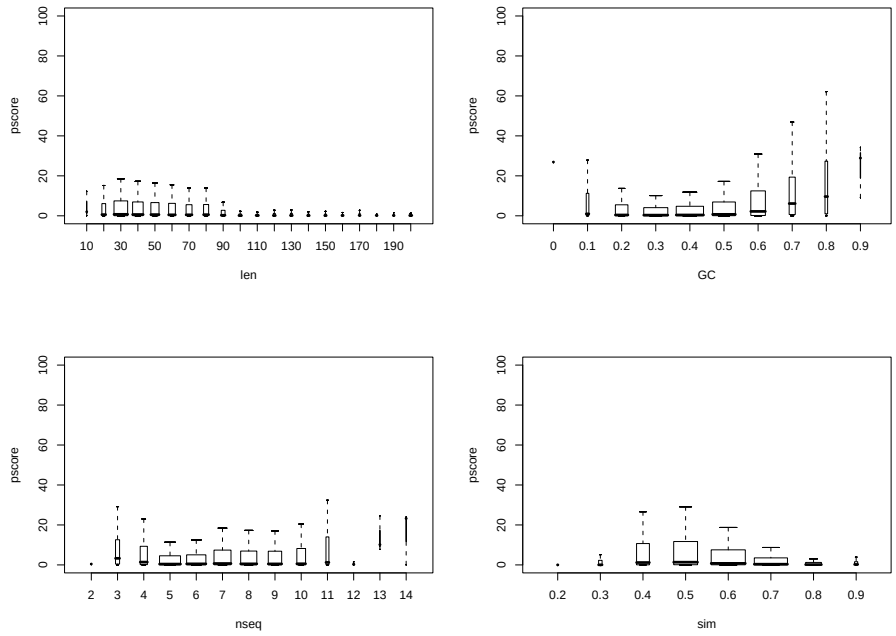


Figure 4.11: Pscores against motif features. Plot setting is the same as figure 4.9

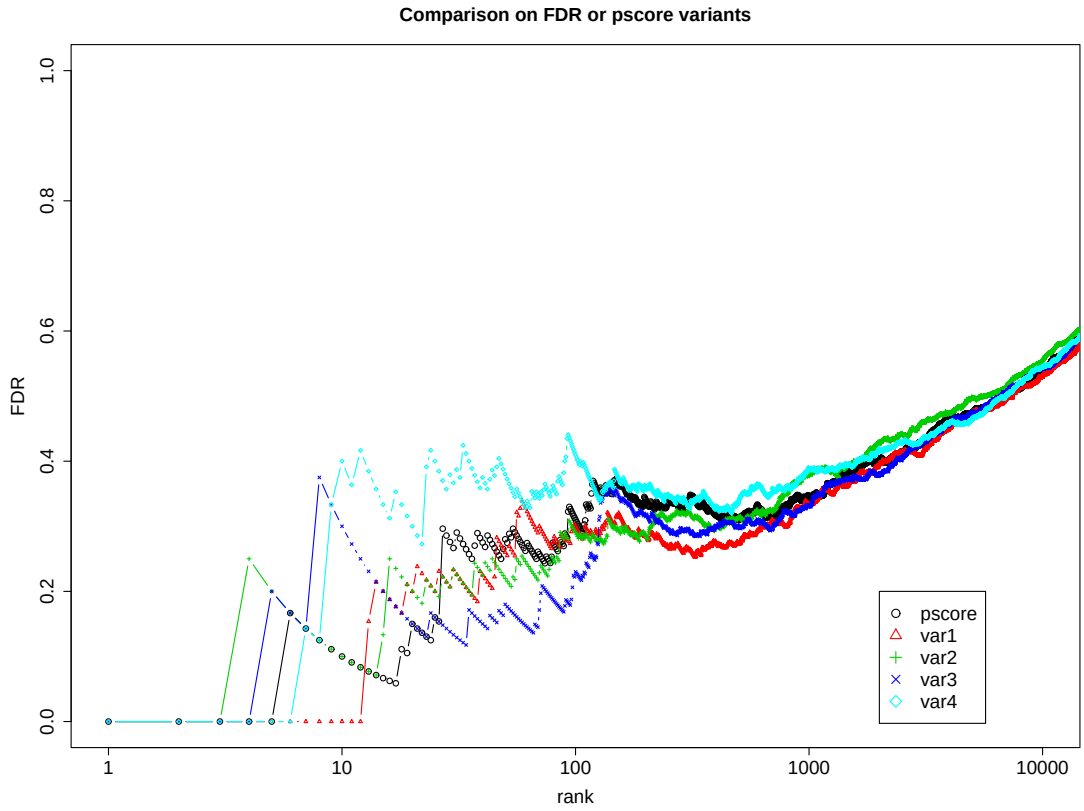


Figure 4.12: FDR for Pscore variants on ENCODE motifs

Chapter 5

CMFINDER-BASED COMPUTATIONAL PIPELINE FOR NCRNA DISCOVERY IN BACTERIA

5.1 Introduction

We would like to apply RNA prediction tools like CMfinder to discover novel noncoding RNA candidates. In the last few years, several groups have performed genome scale computational ncRNA predictions based on comparative genomic analysis. In particular, a pairwise, BLAST-based approach was used to discover novel riboswitch candidates in bacterial genomes, many of which now have been experimentally verified (8). Similar studies have been conducted in various bacterial groups, including Cyano-, Actino-, Alphaproteo- and Extremophilic Eu-bacteria (4; 129; 23; 115). More recent work has extended these searches to eukaryotes (98; 25; 151; 109; 143; 121; 138), which will be discussed in chapter 6.

With some exceptions such as (8) and (143), most of these approaches follow a similar paradigm, which is to search for conserved secondary structures on multiple sequence alignments that are constructed based on sequence similarity alone. Typically, these schemes use measures such as mutual information between pairs of alignment columns to signal base-paired regions. However, the signals such methods seek, namely compensatory base-pair mutations, are exactly the signals that may cause sequence-based alignment methods to misalign, or alternatively refuse to align, homologous ncRNA sequences. Even local misalignments may weaken this key structural signal, making the methods sensitive to alignment quality, which is especially problematic on diverged sequences.

In this chapter, we will present a structure-oriented computational pipeline for genome scale prediction of *cis*–regulatory ncRNAs. It exploits, but does not require, sequence conservation. The pipeline differs from previous methods in three respects: First, it searches in unaligned upstream sequences of homologous genes, instead of well aligned regions con-

structed by sequence-based methods. Secondly, we predict RNA motifs in unaligned sequences using CMfinder, due to its sensitivity on datasets with low sequence conservation, and robustness to inclusion of long flanking regions or unrelated sequences. Finally, we integrate RNA motif prediction with RNA homology search. For every predicted motif, we scan a genome database for more homologs, which are then used to refine the model. This iterative process improves the model and expands the motif families automatically.

Our approach has two key advantages. First, it is efficient and highly automated. Earlier steps are more computationally efficient than later steps, and we can apply filters between steps so that poor candidates are eliminated from subsequent analysis. Thus, even though we use some computationally expensive algorithms, the pipeline is scalable to large problems. Besides providing RNA motif prediction, the pipeline also integrates gene context and functional analysis, which facilitates manual biological evaluation. Secondly, this pipeline is highly accurate in finding prokaryotic ncRNAs, especially RNA *cis*–regulatory elements.

We have applied this computational pipeline for high throughput ncRNA discovery in bacteria. We first tested in Firmicutes, a relatively well-studied group of species that contain most of the known ncRNAs in bacteria, to calibrate the performance of this system. Then we extended the search to all major groups of the bacteria, with focus on discovering novel ncRNA candidates. To both tasks, we have achieved very satisfactory results. This is a joint project with Ronald Breaker's lab from Yale University. The methodology and the pilot study on Firmicutes have been reported in (165) (the online supplement is at `http://bio.cs.washington.edu/supplements/yzizhen/pipeline/`) and the full bacteria genome scan has been reported in (152).

5.2 ncRNAs discovery in Firmicutes: A prototype system

5.2.1 Methods

Our pipeline consists of the following steps. First, we used NCBI's Conserved Domain Database (CDD) (91) to identify homologous gene sets. For each gene, we collected its $5'$ upstream sequence. We call the set of $5'$ sequences associated with one CDD group a "dataset". *Cis*–regulatory elements are often conserved within such groups. Second, we

applied FootPrinter (14), a DNA phylogenetic footprinting tool, to select datasets that are likely to host ncRNAs. In our experience, functional RNAs such as riboswitches often show relatively low overall sequence conservation, but contain interspersed patches where conservation is high. FootPrinter is very effective at highlighting the latter regions. Third, we inferred RNA motifs in each unaligned sequence dataset using CMfinder, which is robust to varying sequence conservation and length of extraneous flanking regions. Fourth, we used RaveNnA (155; 153; 157) to find additional motif instances by scanning the prokaryotic genome database. Riboswitches, for example, often regulate multiple operons that contribute to a single pathway, but no single CDD domain will be common to all of these operons. Thus the search step was a powerful adjunct to the motif discovery process. These newly discovered motif members were incorporated into a refined motif model, again using CMfinder, and in some cases the search and motif refinement steps were repeated. Both CMfinder and RaveNnA rely on the Infernal covariance model software package (34) for RNA motif modeling and search. Finally, we performed gene context analysis and literature searches for the top ranking motifs. The flowchart of the pipeline is illustrated in Figure 5.1, and the details of major steps are described below.

Genome sequence data

We obtained genome sequences from 67 fully sequenced Firmicutes species from the NCBI microbial database (RefSeq (114) release 14, 11/20/2005). To reduce redundancy, we removed near duplicate genomes from analysis. Specifically, each complete sequence was converted to a "CDD vector," wherein the i^{th} component of the vector is the number of predicted occurrences in that genome of the i^{th} conserved domain from NCBI's Conserved Domain Database (91). We eliminated all records whose length-normalized dot product with the CDD vector of a record with a lower accession number was > 0.95. This left 44 complete genome (and 10 plasmid) sequences; accession numbers are given in the online supplement.

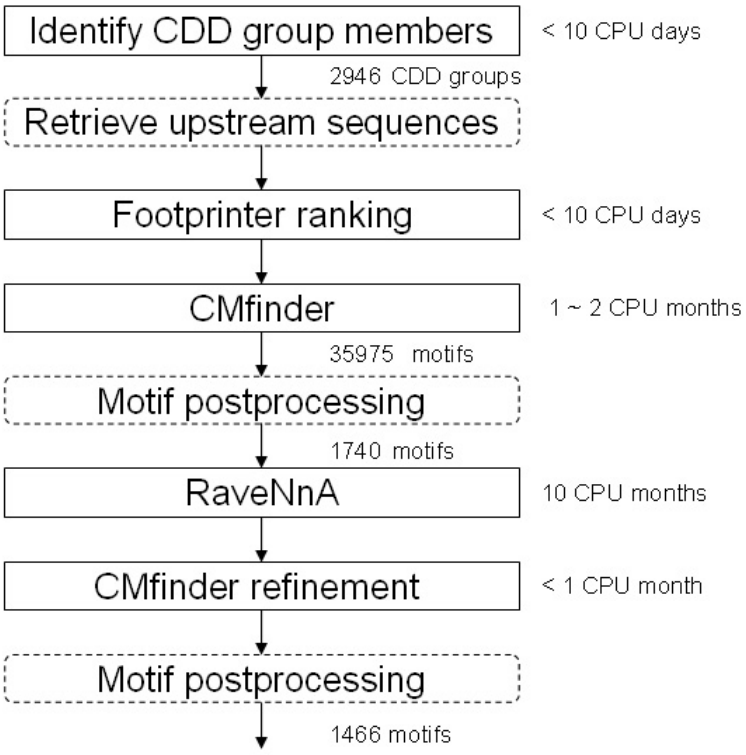


Figure 5.1: Pipeline flowchart. The boxes with solid lines indicate steps involving intensive computation (approximate running time is specified next to each). Other intermediate steps are specified in the boxes with dashed lines.

Identifying homologous gene sets

We first collected amino acid sequences from all annotated protein-coding genes in these species, and categorized them based on NCBI's Conserved Domain Database (version 2.05). The CDD domain models are curated from various resources including Pfam, SMART, and COG. By definition, all members of a CDD group contain a conserved domain in their protein sequences. A group typically includes both orthologs and paralogs. In the NCBI microbial database, 92% of all functionally annotated proteins (i.e., with non-hypothetical description field) are assigned to at least one CDD group, as are 32% of "hypothetical" proteins. We removed CDD groups that contained too few members (4 or fewer), since motif discovery is unreliable on such small groups. We also removed 145 groups with too many members (70 or more) since motif discovery is expensive on such large groups.

Collecting upstream sequences

For each gene in a CDD group, we collected up to 600 nucleotides upstream of its start codon, which typically includes both $5'$ UTR and promoter sequences. The prevalence of operons in prokaryotic genomes complicates the extraction of the regulatory regions, as the desired regulatory region may be upstream of the entire operon rather than immediately upstream of the selected gene. To handle this complication in a conservative manner, we extracted the noncoding sequences upstream of the gene and upstream of its plausible operon using MicroFootPrinter (105).

After collecting the upstream sequences, we removed redundant sequences (95% sequence identity across 80% of the sequence according to BLAST), and masked regions that match tRNA or rRNA models in the Rfam database.

Ranking using FootPrinter

FootPrinter (14) identifies conserved sequence motifs in a set of unaligned homologous sequences using phylogenetic analysis. We applied FootPrinter separately to the set of upstream sequences of each CDD group. We scored each FootPrinter motif by the number of motif instances minus the corresponding parsimony score, and scored each dataset as the sum of its top 30 motif scores. The resulting scores are used to rank all datasets. This ranking is performed by MicroFootPrinter (105), a front end to FootPrinter. This step is used for optional filtering. For this prototype system, we did not actually removed the low scoring datasets so that we can evaluate the performance of each step of the pipeline.

RNA motif discovery

We used CMfinder-0.2 (167) for RNA motif prediction in unaligned sequences. For each dataset, we produced up to 5 single stem-loop motifs, 5 double stem-loop motifs, and used CMfinder heuristics to combine the motifs into more complicated structures if possible. We ranked all CMfinder motifs using a heuristic scoring function described in Section 4.2. We refer to these scores as *composite scores*.

Genome scans for RNA motif homologs

One of the key strengths of our method is the integration of motif discovery with motif search. Motif discovery is focused on groups of orthologs defined by common CDD membership, since such groups seem likely to be enriched for common *cis*–regulatory elements. However, many *cis*–regulatory elements such as riboswitches will be found near a variety of operons involved in a coherent pathway, which may *not* share a common CDD group. Hence, genome-scale search for additional motif instances is an important component of our approach. Additional instances allow us to construct more accurate motif models, as well as giving insight into potential biological roles for the elements.

Given RNA motifs produced by CMfinder, we searched for additional instances using Infernal CMs (34) accelerated with the ML-heuristic filter (157) implemented in RaveNnA 0.2f. For reasons of speed, two levels of search were used. The initial search database was derived from all 75 finished Firmicutes genomes in RefSeq17 (4/30/2006) (114), a total of approximately 200 million nucleotides. Based on sequence annotations, we extracted only intergenic regions for searching, but extended each by 50 nucleotides in each direction to account for common errors in protein-coding gene annotations. The resulting database contained approximately 34 million nucleotides. This small database made it feasible to perform searches for all motifs (averaging 4.8 CPU hours per motif), and reduced false positives when compared to the full genome database. After motif refinement (incorporating hits from this "mini" scan), we performed "full" scans with selected motifs. Full scans examined the prokaryotic subset of the 8 gigabase RFAMSEQ data set (version 7.0, March 2005, built from "finished" portion of the EMBL nucleotide database), a total of approximately 900 megabases. In particular, comparisons to Rfam (e.g., Table 5.2) were based on full scans, since Rfam full alignments are also derived from scans of RFAMSEQ. For model refinement, we ran CMfinder on all hits with RaveNnA E-value less than 10. E-values were calculated as in (78). The necessary extreme value distribution calculations dominate the run times for mini-scans, but not full scans.

Clustering overlapped motifs

We identified the overlap between refined motifs according to their genomic coordinates. One motif is grouped with another if at least half of its members overlap, and the overlapped regions are longer than half of the motif length. The motifs are clustered progressively with high ranking motifs processed first. Finally, we ranked clusters based on their highest scoring motifs.

Identifying known Rfam motifs

To find which of our predicted motifs were already known, we compared them against the Rfam database. As some of our input sequences were not included in RFAMSEQ, we BLASTed our motif instances against Rfam full family members (produced by scanning Rfam covariance models on the RFAMSEQ genomic database; see (58)). For BLAST, we used a word size of 12, and selected the hits with length $\geq$ 30bp, E-value $\leq$ 10, and sequence identify $\geq$ 90%. These permissive BLAST thresholds resulted in a few isolated hits that we believe to be false positives. These false positives match fragments, each of about 30 bases, of the Rfam RNA-OUT, Intron-gpII, QaRNA, and RNaseP_bact_a families. In general, they are too short, weak and/or isolated to be compelling, in sharp contrast to the matches reported in Table 5.1.

5.2.2 Results

To evaluate the performance of this approach, we first tested in Firmicutes, a Gram-positive bacterial division that includes *Bacillus subtilis*, a relatively well-studied model organism with many known ncRNAs. The method exhibits low false positive rates on negative controls (permuted alignments), and low false negative rates on known Firmicutes ncRNAs. Rfam includes 13 ncRNA families categorized as *cis*–regulatory elements with representatives in *B. subtilis*. Of these, 11 are included among our top 50 predictions and a 12th appears somewhat lower in our ranking. Two other Rfam families are also represented among our top 50 predictions. In addition, both the secondary structure prediction and identified family members are in excellent agreement with Rfam annotation. For 14 Rfam

families mentioned above, we achieved 91% specificity and 84% sensitivity on average in identifying family members, and 77% specificity and 75% sensitivity in secondary structure prediction. Many promising novel ncRNA candidates were also discovered and are discussed below. We included 44 completely sequenced Firmicutes species (see online supplement) and 2946 CDD groups in this study. For each of the three main steps—FootPrinter, CMfinder, and RaveNnA-based refinement—we produced scores to determine which candidates were worthy of continued analysis. For evaluation purposes, we recorded the scores of candidates at each step, but eliminated none; in the future we may use them as filters.

CMfinder produced 35975 motifs in total. To identify distinct motifs corresponding to different RNA elements, we removed poor and redundant motifs and clustered the rest based on overlap. This treatment produced 1740 motifs grouped into 1050 clusters. After RaveNnA-based refinement, more motifs were identified as redundant and removed. 1466 motifs remained, grouped into 1060 clusters. (A few of the original clusters were subdivided based on divergent search results.) The full list of candidates, as well as the details for CMfinder motif postprocessing are available in the online supplement.

Negative controls: Permuted alignments

To evaluate how many of our top candidates could have arisen by chance, we performed a randomized control experiment. We first computed CLUSTALW alignments of the top 100 sequence datasets with the highest motif scores (before the RaveNnA scan). We then randomly permuted the alignments 50 times, maintaining the approximate gap pattern by swapping two columns only if their gap patterns have over 80% similarity. (The gap pattern of each column is described by a binary vector, which is 0 if the sequence has a gap in the corresponding position in the column, and 1 otherwise. The similarity is simply measured as the ratio of the number of matches between the two binary vectors over the length of the vectors.) This is a much simpler algorithm than the one included in the RNAz package. After degapping each permuted alignment (treating it as a set of unaligned sequences), we applied CMfinder, retaining the top ranking motif from each randomized dataset. We used this collection of 5000 motifs to estimate the background score distribution, and to infer

p-values for predicted motifs in the original datasets. Results are shown in Figure 5.2. By this measure, all 100 top scoring motifs have p-values less than 0.1, with median at 0.016. Additionally, 73 of the 100 candidates in the original dataset score higher than all motifs in the corresponding randomized datasets.

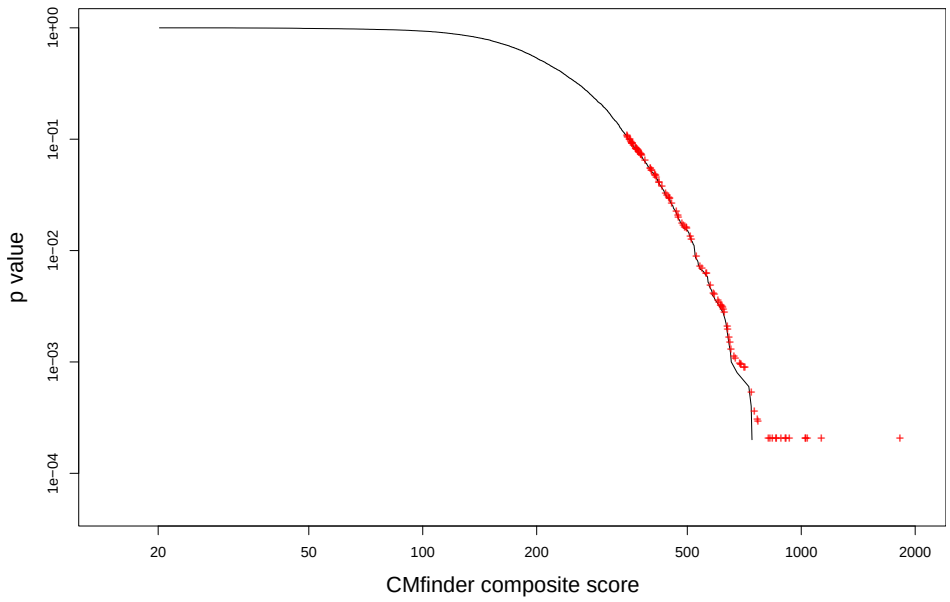


Figure 5.2: The empirical p-value distribution based on the permutation test. The black curve shows the complementary cumulative distribution function for the composite scores on randomized datasets, i.e., for each score x, the fraction of permuted alignments exceeding that score. The red cross data points show the p-values for the composite scores of the motifs on the original datasets. All p-values are greater than or equal to 2×10^{-4} as there are only 5000 samples in the background distribution.

Note that this estimation of p-values is imperfect. In particular, with the scoring scheme we used, datasets containing phylogenetically close sequences tend to score well in comparison to more diverged sets, because permuting the CLUSTALW alignments preserves their sequence conservation. (Independently permuting individual sequences instead of alignments would be less realistic, since in practice, *cis*–regulatory RNA motifs are often embedded in regions exhibiting some sequence conservation for other reasons.) Although imperfect, the

significance of real motifs tend to be underestimated by this method.

Positive Controls: Discovering known Rfam families

To roughly assess the sensitivity with which the method discovers true ncRNAs, we looked at its recovery of known Rfam families. We masked matches to Rfam's tRNA and rRNA models since otherwise these widespread, strong motifs might hide nearby, weaker, but still interesting ncRNA structures. Other Rfam families were not masked and serve as a positive control for our methods. Table 5.1 shows the distribution of known Rfam families in our candidate list after running FootPrinter, CMfinder and RaveNnA. We used the refined motifs as the final output.

According to Rfam (version 7.0), *B. subtilis* contains members of 21 families, categorized into 13 *cis*–regulatory families, 1 intron element, and 7 RNA gene families. We masked tRNAs and rRNAs (4 of the 7 gene families). Of the 17 remaining families, 13 appear within our top 50 candidates: 11 of the 13 *cis*–regulatory families present in *B. subtilis* together with two of the gene families (RNaseP_bact_b and SRP_bact). The 4 families not represented are the other 2 *cis*–regulatory elements (*ykkC-yxkD* and *ydaO-yuaA*), the remaining RNA gene tmRNA, and the intron element (Intron_gpI). The exclusion of Intron_gpI is not surprising, as we did not search introns. The *ydaO-yuaA* motif escaped detection because it is present in only 3 of the 68 sequences in its CDD group. The *ykkC-yxkD* and tmRNA motifs, although not among our top 50, would still have been ranked high enough to be discovered in a blind test. Note that, although our computational pipeline is oriented towards discovery of *cis*–regulatory elements, we sometimes find RNA genes like RNaseP, SRP and tmRNA because they happen to be conserved in synteny. We also found a partial tRNA motif, not masked since parts of the tRNA lie outside of the collected upstream sequences.

We can potentially filter the candidates at each step to scale this pipeline for larger genomes. In particular, we could have applied CMfinder to only the top half of the datasets according to FootPrinter, and performed genome scans on only the top 500 motifs, without missing any real Rfam families as listed in Table 5.1. On average, it takes FootPrinter less than 1 minute, and CMfinder 10 minutes to process each dataset, while it takes RaveNnA

4.8 hours to scan each motif. We could save considerable computation time by running expensive algorithms only on good candidates.

As shown in Table 5.1, the ranks for most known ncRNAs improve at each successive step of the pipeline, as more supporting evidence is found. Starting from FootPrinter motifs, CMfinder improves the alignment and identifies consensus secondary structure, while genome scans locate many more motif instances, typically providing still better alignments and additional clues to their functions.

To measure the quality of our automatically constructed motif models, we compared them to the Rfam alignments for the same families. Rfam's covariance models are built from hand-curated "seed" alignments/structure annotations. These in turn are used to build Rfam's "full" alignments by automatically searching RFAMSEQ, a high quality, non-redundant subset of EMBL, and automatically aligning all hits.

For 14 Rfam families in Table 5.1 for which we found good matching motifs, we selected the top CMfinder two motifs from each family, and performed full genome scans on RFAMSEQ, the same sequence database used to construct the Rfam full alignment. To reduce computation time, we did not scan Eukaryote genomes, and the Rfam hits from these genomes were excluded from the following analysis. (This treatment affects only a few eukaryotic Cobalamin and Lysine hits, all believed to be Rfam errors or bacterial contamination in the genome sequences, plus a few THI hits, which are real.) For each motif, we selected scan hits at an E-value cut-off of 100, reconstructed the motif alignments using CMfinder, and removed the low scoring instances ($\leq$ 20 bits). We compared these predicted motifs to corresponding Rfam full alignments, which serve as the gold standard in this test. Table 5.2 shows the accuracy of our motifs in membership prediction, motif coverage, and secondary structure prediction. Secondary structures were compared at the base pair level, and only the base pairs with at least one end falling into the overlapped regions are counted. For both predicted motifs and Rfam full alignments, we removed non-canonical base pairs from each sequence. Of the two motifs chosen for each family, we report the one with better results.

For membership prediction, we achieved an average of 84% sensitivity and 91% specificity. The overlapped regions between predicted motif members and corresponding Rfam

members account for 81% of the length of the predicted members, and 82% of the length of Rfam members. In the overlapped regions, the secondary structure prediction has 75% sensitivity and 77% specificity. These results suggest our predicted motif models are very accurate compared to Rfam models, which are learned from the hand curated seed alignments.

For many riboswitch families, the main differences between our motif models and Rfam models are located in boundary regions. Our predicted motifs tend to include the transcription terminator (if present), which is a stable hairpin followed by a stretch of U's (e.g., Lysine, S_box, T-box). Although transcription terminators are functionally important, the Rfam riboswitch models do not include them. On the other hand, CMfinder tends to miss the closing helix of large multi-loop structures (e.g., Cobalamin, *ykoK*). Most other differences are local perturbations such as small shifts or extra base pairs.

As shown in Table 5.2, we achieved over 80% membership sensitivity for all families except *yybP-ykoY*, *gcvT* and Cobalamin. The predicted *yybP-ykoY* motif differs from Rfam's motif mainly at the multi-loop closing helix. Cobalamin and *gcvT* are two riboswitches with poor sequence conservation (46% and 51% average sequence identity, respectively). While our motifs from the initial full genome scan may be too specific, sensitivity increases significantly with only a small loss in specificity after another iteration of RaveNnA scan and refinement (data not shown).

For *ykkC-yxkD* and T-box, we predicted many more members than Rfam. The predicted *ykkC-ykxD* motif includes the transcription terminator, which caused false positives in our full genome scans. These false positives, however, all have much less significant E-values than the true positives, hence are relatively easy to eliminate by inspection. In contrast, for T-box we believe most "false positives" (with respect to Rfam) are actually real. Out of 291 members not included in the Rfam full alignment, 127 are upstream of and on the same strand as aminoacyl-tRNA synthetase genes, where most T_box leaders are found, and the others are largely in poorly annotated regions.

Motifs not in Rfam

We examined the best scoring motif in each of the top 200 motif clusters (after RaveNnA scan). Of these 200 motifs, 116 were deemed unlikely to represent novel ncRNAs: they have covariance model scores ≤ 40 bits, single hairpin structures, and most were shorter than 30 nucleotides. (Many of these 116 are nevertheless biologically relevant. Many correspond to transcription terminators of upstream genes. Others contain known inverted repeat motifs targeted by DNA binding proteins.) Of 84 remaining motifs, 20 correspond to Rfam families, and 11 to hypothetical transposons. The remaining 53 are candidates for novel ncRNAs. Literature review suggests that many of these candidates are functional. We manually removed the redundant candidates with the same functional roles, and present the rest in Table 5.3.

Several candidates turn out to be known regulatory elements that have been described previously in the literature, including:

- **PyrR attenuator:**

  Upstream of CDD 28178 we predicted a PyrR RNA binding site (87), which regulates *pyr* operon transcription by switching between alternative antiterminator versus anti-antiterminator plus terminator structures. The motif we predicted corresponds to the anti-antiterminator plus terminator structure, which is stabilized upon binding to PyrR. It includes 69 instances in 31 Firmicutes species, with 2 copies per species on average: one copy upstream of the *pyrP* or *pyrR* gene, and one copy upstream of *pyrB*.

- **6S:**

  This ncRNA binds to σ^{70} RNA polymerase holoenzyme to globally regulate gene expression in response to the shift from exponential growth to stationary phase. Although 6S has been known in *Escherichia coli* and close relatives for over 35 years (15), the corresponding Rfam model (RF00013 6S/SsrS RNA) is confined to γ-proteobacteria, and its Firmicutes homologs were only identified recently by experimental (145) and computational (9) means; see also (161). We have discovered 6S

in Firmicutes independently in this study. The motif we predict is a partial 6S that includes the most conserved core.

- **Inverted Repeats:**

  It is difficult to determine whether a motif with inverted repeats functions at the DNA or RNA level without considering its genomic context. Based on the literature, three single hairpin inverted repeat motifs in Table 5.3 appear to be known DNA binding sites for regulatory proteins: *hrcA* binding sites (rank 44), *blaI/mecI* binding sites (rank 140), and hypothetical CadC binding sites (rank 50). (All three are longer and had significantly higher covariance model scores than the 116 removed inverted repeats mentioned above.) Details of these elements are discussed below:

  CIRCE (Controlling Inverted Repeat of Chaperone Expression) is a wide spread regulatory heat shock element in bacterial species (102) that binds to *hrcA*, present predominately upstream of *groE* and *dnaK* operons. We have predicted the exact consensus sequence upstream of both *groE* (CDD 10332) and *dnaK* (CDD 11135) operons. Previous biochemical assays show that it is not a potential stem-loop structure but the conserved sequence of CIRCE that is required for repression, while a dual regulatory role is also suggested of CIRCE both as a DNA reducing transcription and as an RNA element promoting the rapid turnover of the mRNA under normal growth conditions (102).

  CDD 8892 contains penicillinase repressor genes in *blaR1* and *mecR1* operons. The motif instances between the two operons contain significant differences while sharing some common signatures. Literature search reveals that the motif instances upstream of the mecR1 operon is a palindrome that acts as a repressor (132). The motif instance upstream of blaR1 is also a binding site, which contains two 18-bp palindromes with dyad symmetry, separated from each other by a 13-bp linker. Our motif also includes other regulatory signals, such as -10 and -35 promoter sequences and ribosome-binding sites. However, it is unlikely that this element functions as an RNA transcript.

  CDD 5638 contains homologs of CadD, a Cadmium resistance transporter. CadA,

another cadmium resistance protein, is known to contain a CadC binding site at nucleotide positions -7 to +14 relative to the transcription start site, which is an inverted repeat TCAAATA-AA-TATTTGA. The motif we predicted contains a similar sequence, TCAAAAATATTTTTGA. We hypothesize that this motif has a similar function.

Novel ncRNA candidates: Ribosomal protein leaders

To demonstrate how CMfinder predictions can accelerate the discovery and characterization of new RNA motifs, we present a detailed analysis of two conserved mRNA leader structures that are most likely involved in autoregulation of L19 and L13-S9 ribosomal protein expression. Five additional presumed ribosomal autoregulatory motifs are presented in the online supplement.

Many ribosomal protein (r-protein) operons regulate their own expression in *E. coli* (168; 104). Once enough of a specific r-protein encoded by an operon has been produced, i.e., all of its rRNA binding sites are saturated, excess copies of the protein bind to the $5'$ untranslated leader region of its mRNA and induce structural changes that compete with ribosome binding or stall initiating ribosome complexes. This general repression mechanism appears to apply to many r-protein operons, but the specific RNA structures recognized by orthologous r-proteins are generally not conserved between *E. coli* and other bacterial groups.

For example, the S15 mRNA leaders from *E. coli*, *Geobacillus stearothermophilus*, and *Thermus thermophilus* assume different, apparently unrelated RNA structures, that all seem to mimic the same rRNA binding site (137). Similarly, the mRNA binding site of S4 differs between *E. coli* and *Bacillus* species (60).

Within a bacterial division, the same regulatory structure may be used in many species. Thus, an mRNA leader structure recognized by L4 is conserved in many, but not all, γ-proteobacteria (1). Our comparative analysis using CMfinder is well suited to recognize r-protein mRNA leader motifs conserved at this taxonomic level. Indeed, it detects the only two r-proteins leader structures that have currently been characterized in Firmicutes (S4 and

S15). However, the structure predicted for S4 leaders by CMfinder agrees only partially with a previous phylogenetic analysis of this element based on fewer, exclusively *Bacillus*, species (60). After manually examining the regions aligned by CMfinder, we predict a consensus structure that is close to the Grundy and Henkin (60) model but has a different pseudoknot (see online supplement). The relatively poor performance of CMfinder on the S4 leaders may be partly due to the clustering of a subfamily of *Lactobacillus* sequences with a slightly different consensus structure from the *Bacillus* sequences. CMfinder performed better on the S15 leader (rank 842), accurately predicting the location and extent of the largest helix-2 feature (127). Here, it misses only the small adjacent helix-3, and an additional stem that overlaps the open reading frame.

CMfinder also predicts a novel regulatory RNA structure upstream of L19, encoded by the *rplS* gene, in *Bacilli*, *Lactobacilli*, *Clostridia* and *Fusobacteria* species (Table 5.3, rank 160). In *E. coli*, L19 is expressed as the last of four genes from a polycistronic mRNA (159). A similar gene order is conserved in some Firmicutes (approximately 2/3 of those with the RNA motif), and there is not an intrinsic transcription terminator between the orthologous upstream *trmD* gene and *rplS* in *B. subtilis*. However, the intergenic distance between *trmD* and *rplS* is typically greater than 100 bp in Firmicutes (142 nt in *B. subtilis*) compared to only 41 nt between *trmD* and *rplS* in *E. coli*. Putative promoter -35 and -10 hexamers occur within this intergenic region upstream of each predicted RNA structure (Fig. 5.3A), suggesting that L19 is expressed as a separate transcriptional unit from the upstream genes in Firmicutes.

The putative L19 autoregulatory mRNA structure is a small bulged hairpin (Fig. 5.3B). The length of the terminal P2 stem-loop varies, but the outer P1 helix always has exactly 7 base pairs. Most primary sequence conservation occurs in the asymmetric internal bulge and P1 stem. The original CMfinder results include some nonconserved sequences and a spurious stem-loop upstream that are not preserved in all examples. Within the conserved region, CMfinder identifies most of the pairing predicted in our manually refined model.

This RNA structure is always found close to the ribosome-binding site (RBS) of the L19 open-reading frame. If it is involved in typical r-protein autoregulation, then L19 binding might stabilize an alternate paired conformation wherein the 5′ side of P1 sequesters the

RBS to repress gene expression (Fig. 5.3C). Alternately, the predicted P1 stem might only be stable in the presence of L19, and when it forms, its proximity to the open-reading frame might prevent translation initiation. Ribosomal protein L19 binds to the large rRNA subunit at the 50S-30S interface. We were unable to identify any homology between the predicted mRNA leader structure and its 23S rRNA binding site in the *E. coli* ribosome (126), or homologous positions in the *B. subtilis* ribosome (16), that might suggest a simple regulatory model. It is possible that the predicted regulatory hairpin mimics the structure of the rRNA binding site, or participates in a more complex regulatory mechanism.

CMfinder predicts a second novel RNA structure (Fig. 5.4) upstream of the L13-S9 operon, encoded by the *rplM* and *rpsI* genes, in *Bacilli* and *Lactobacilli* species (Table 5.3, rank 157). There is a strong, near-consensus, promoter directly upstream of this motif that defines a conserved transcription start site. The L13-S9 structure is also a bulged hairpin, but it is larger than the L19 motif. There is striking conservation of 7 loop nucleotides (CCCCGGA) that are identical in all sequences. Additional conservation occurs in the bulge and within the P1 helix. CMfinder correctly predicts the P2 helix in this manually revised model, and it also identifies the core base pairs in the P1 helix, except in cases where an inserted stem loop occurs in the $3'$ side of the bulge.

S9 is a secondary small subunit binding protein, requiring prior S7 binding to associate with 16S rRNA (113). Most r-proteins involved in autoregulation are primary binding proteins that can bind directly to rRNA, so it seems most likely that L13, a protein that binds to 23S rRNA early in large subunit assembly, recognizes this leader structure. Here again, we were unable to identify any conservation between the rRNA contact sites of L13 and *E. coli* 23S rRNA or the corresponding sites in *B. subtilis* 23S that suggest a regulatory model. There is sometimes a significant distance between the putative regulatory RNA structure and the open-reading frame. Alternate pairings between the U-rich $5'$ side of P1 and a region overlapping the start codon can be devised for many sequences, so it is possible that this alternate conformation is enforced by L13- or S9-binding to the mRNA leader to prevent translation.

In the online supplement, we present the full manually refined structural alignments for the above two motifs plus five additional putative r-protein leader regulatory motifs: IF-3,

L10, L21, S4 and S10; cf Table 5.3. Based on our experiences with these putative mRNA leader structures, it should be straightforward to define many more candidates for r-protein autoregulatory structures in other bacterial groups with our pipeline. Such studies could illuminate how this form of regulation has been modified and preserved during evolution and would make genomic annotation of noncoding RNAs more comprehensive. Five of these seven new putative regulatory RNA elements are now included in the Rfam database (version 8.0).

5.3 ncRNA discovery in all bacterial groups

5.3.1 Method

We proceed to apply the same pipeline described above to find novel RNAs within all major bacterial groups, which include Firmicutes; α-proteobacteria; β-proteobacteria; γ-proteobacteria; δ- and ϵ-proteobacteria; Actinobacteria; Cyanobacteria; the Bacteroidetes/Chlorobi group, Chlamydiae/Verrucomicrobia group and Spirochaetes. Bacteria are split into groups to reduce meaningless comparisons. Phyla with few sequenced members are coalesced based on taxonomy to attempt to build critical mass. Phyla with only one or two sequenced members (e.g., Chloroflexi) are ignored.

The automated pipeline predicted roughly 10,000 motifs. We went through the motif rank list, manually selected promising candidates and investigated them. Our decision to select a motif was based on the motif scores (described in Section 4.2), as well as our qualitative evaluation. These evaluation criteria include evidence of covariation and variable-length or modular stems. Motifs that we find most convincing have both conserved sequence and structure. Conserved structure and particularly covarying paired nucleotides argues in favor of a functional RNA. Conserved sequence is also important both because functional RNAs usually have them and it is possible that some of these are protein binding motifs in which a homodimeric protein binds to a given DNA-based element in opposite strands. We especially favor covarying nucleotide positions with strong surrounding sequence conservation, indicating correct alignment.

We also examine the gene context of candidates. Probable cis-regulatory motifs were

consistently located upstream of homologous genes, or a set of genes with related functions, and often had features typical of known gene-control mechanisms. *Cis*–regulatory RNAs often have one of two noteworthy structural features: rho-independent transcription terminators, or stems that overlap the Shine-Dalgarno sequence (bacterial ribosome-binding site) (163). Rho-independent transcription terminators usually consist of a strong hairpin followed by four or more U residues. Regulatory RNA domains can control gene expression by conditionally forming the terminator stem. Similarly, conditionally formed stems can overlap the Shine-Dalgarno sequence, thereby regulating genes at the translational level.

Riboswitch candidates were motifs that were classified as an RNA and a *cis*–regulatory element, showed evidence of high conservation of nucleotides at some positions, exhibited a complex secondary structure (not just a hairpin) and were associated with genes that were judged likely to be controlled by a small molecule.

After selecting a motif, our main focus was to improve the alignment by finding additional stems, finding additional homologs or refining the alignment. Editing and annotation were done with RALEE (57), NCBI BLAST (2), Mfold (171), Rnall (146) (for rho-independent transcription terminators). Metabolic pathways were routinely analyzed based on KEGG (72). To find additional structured elements in alignments, we extended alignments on their $5'$ and $3'$ ends by 50-100 nucleotides, and realigned using CMfinder or inspected sequences manually. RaveNnA were also used to adjust sequence alignment whenever needed.

To help reject motifs with spurious structure, we performed homology searches without structure; sequence-based matches that do not conserve the structure strongly argue against a true RNA. However, such homologs may be missed by CMs, which assume the structure is conserved. Several motifs were rejected with this strategy. A common sources of false positives, which we screened manually, were repetitive elements. These appear many times per genome and show extremely high sequence conservation, but little structure conservation.

5.3.2 Results

We found 22 novel motifs that are putative conserved RNAs. These motifs are summarized at table 5.4. The structural diagrams of these candidates can be found at (152). Six of

these candidates are hypothesized to be riboswitches, and they will be described next.

- GEMM motif

  We found 322 instances of this motif in both Gram-positive and Gram-negative bacteria. It is common in δ-proteobacteria, particularly in Geobacter and related genera. Within γ-proteobacteria, it is ubiquitous in Alteromonadales and Vibrionales. It is also common in Firmicutes and Plantomycetes. Prominent pathogens with this motif include the causative agents of cholera and anthrax. Out of 309 instances where sequence data includes gene annotations, GEMM is in a 5' regulatory configuration to a gene in 297 cases, implying a *cis*–regulatory role.

  GEMM consists of two adjacent hairpins (paired regions) designated P1 and P2 (Figure 5.5). P1 is conserved in sequence and structure, with a highly conserved internal loop, and the hairpin loop is almost always a GNRA tetraloop. The P2 hairpin shows even more modest conservation than P1. When the P1 tetraloop is GAAA, a GNRA tetraloop receptor usually appears in P2, which is a well-known 11-nt motif, likely to be favored by GAAA loops (39). The sequence linking P1 and P2 is almost always AAA. Many instances of GEMM include a rho-independent transcription terminator hairpin. The $5'$ side of the terminator stem often overlaps (and presumably competes with) the 3' side of the P2 stem. If GEMM is a riboswitch, ligand binding could stabilize the proposed P1 and P2 structure, thus preventing the competing transcription terminator from forming. In this model, higher ligand concentrations will increase gene expression. A significant proportion of GEMM motifs are in a tandem arrangement. Such arrangements of regulatory RNAs tend to be involved in sophisticated regulatory control systems (90).

  Thorough analysis of gene context revealed that genes presumably regulated by GEMM display a wide range of functions, but most are related to the extracellular environment, the membrane, or motility. Therefore, this motif is termed as "GEMM" - Genes for the Environment, for Membranes, and for Motility. We have proposed three hypotheses for its ligand, if it is indeed a riboswitch. First, GEMM RNA association with chitin-related genes (chitin being a polymer of GlcNAc) suggests GlcNAc or a

derivative as a ligand. Second, GEMM RNA may be involved in a signal transduction cascade. For example, the regulated GGDEF signaling domains synthesize cyclic di-GMP. A ligand involved in signal transduction could explain why GEMM is involved in a variety of processes in different bacteria, if different bacteria use the signaling molecule for different purposes. The fact that many GEMM-associated genes themselves encode signal transduction domains could suggest a mechanism by which the abundance of the signal transduction proteins are regulated. Finally, a molecule involved in cell-cell communication is another possible ligand. Such a role could also explain why different species use GEMM in different contexts, particularly competence. To this date, we have experimentally validated that one of the hypothesized ligands indeed binds to GEMM (unpublished data), yet further study to unravel its functional role is still undergoing.

- SAH motif

  This motif was found at $5'$ UTR of genes related to SAH (S-adenosylhomocysteine) metabolism. This motif is highly conserved in sequence and structure (Figure 5.6), showing covariation within predicted stem regions, including modular and variable-length stems. This motif is present primarily in β- and γ-proteobacteria, especially the genus Pseudomonas.

  SAH is a part of the S-adenosylmethionine (SAM) metabolic cycle, whose main components include the amino acid methionine. SAH is a byproduct of enzymes that use SAM as a cofactor for methylation reactions. Typically, SAH is hydrolyzed into homocysteine and adenosine. Homocysteine is then used to synthesize methionine, and ultimately SAM. High levels of SAH are toxic to cells because SAH inhibits many SAM-dependent methyltransferases. Therefore cells likely need to sense rising SAH concentrations and dispose of this compound before it reaches toxic levels. The genes that the SAH motif associates with are S-adenosylhomocysteine hydrolase (ahcY), cobalamin-dependent methionine synthase (metH) and methylenetetrahydrofolate reductase (metF), which synthesizes a methyl donor used in methionine synthesis. The genetic arrangement of the SAH motif and its high degree of conservation are consis-

tent with a role in sensing SAH and activating the expression of genes whose products are required for SAH destruction. Indeed, biochemical and genetic evidence supports the hypothesis that this motif is an SAH-sensing riboswitch (J.X. Wang *et al.*, submitted).

- The SAM-IV motif

  We found a SAM (S-adenosylmethionine)-binding riboswitch in Streptomyces coelicolor and related species (Z. Weinberg *et al.*, to appear, RNA 2008). This is the fourth distinct set of mRNA elements to be reported that regulate gene expression via direct sensing of S-adenosylmethionine (SAM or AdoMet), and is referred to as "SAM-IV" (the previous three have been reported in (164; 23; 44)). Experimental evidence validated the binding of this motif with SAM and its regulatory role (Z. Weinberg, *et al.*, submitted).

  While all three previously known riboswitch families specifically recognize SAM, they have no apparent similarity in sequence or structure. While SAM-IV and SAM-I share similar ligand binding sites, they also have different scaffolds (see Figure 5.7) and distinct patterns of nucleotide conservation in many places. The P4 hairpin in the SAM-I core is absent in SAM-IV, but a different P4 hairpin is found outside of P1 in SAM-IV with dissimilar conserved nucleotide identities. SAM-IV is also predicted to form an additional pseudoknot that SAM-I lacks, and the P2 hairpin of SAM-IV is significantly different from the P2 hairpin in SAM-I.

  The similarity between SAM-I and SAM-IV is too low to automatically assume a common evolutionary origin. To account for such divergent evolution, we hypothesized that a SAM-I-like ancestor loses its P4 stem, and any negative effect later caused by the loss of P4 was compensated by the gain of a different P4 stem $3'$ to the P1 stem, leading eventually to a SAM-IV-like RNA.

  The discovery of multiple distinct SAM-binding riboswitches supports the view that RNA has sufficient structural sophistication to solve the same biochemical challenges in diverse ways. If RNAs can easily assume various structures with similar functions,

then we expect that the diversity of riboswitch RNA structures could be far greater than what has been found to date.

- The moco motif

  This RNA motif is located upstream of genes encoding molybdate transporters, molybdenum cofactor (Moco) biosynthesis enzymes, and proteins that utilize Moco coenzymes, with a total of 176 instances in Proteobacteria, Clostridia, Actinobacteria and Deinococcus-Thermus species.

  The principle secondary structure of the Moco motif consists of five stems, labeled P1 through P5 (see Figure 5.8). For most instances, the P4 stem contains a GNRA tetraloop, and the tetraloop receptor is centered in the P2 bulge. The tetraloop at the P4 is GAAA most of the time, with a few exceptions being GCAA, in which case, the corresponding tetraloop receptors are missing. The P3 stem seems optional, present in only some instances. The alignment of all instances reveals that while the aptamer region is highly conserved, different types of expression platforms are formed. In some cases, a possible expression platform encompass the ribosome binding site (RBS) for the downstream open reading frame (ORF) within the nucleotides forming the P1 stem, while in some other cases, a hypothetical intrinsic transcription terminator is downstream of the aptamer region. Moreover, some organisms carry more than one Moco RNA representative and manifest both types of expression platforms in the same organism. Genetic assays support that Moco RNA controls gene expression in response to Moco production, and this conserved RNA discriminates against closely related analogs of Moco.

- The COG4708 motif

  This motif is found upstream of COG4708 genes in some but not all Streptococcus and Lactococcus lactis species. COG4708 genes are predicted to encode membrane proteins. Although the COG4708 motif is highly constrained phylogenetically and has only six unique sequences, it shows covariation, modular stems and variable-length stems (see Figure 5.9). The motif has a pseudoknot that overlaps the putative

Shine-Dalgarno sequences of COG4708 genes, which suggests that the motif encodes a *cis*–regulator of these genes.

We hypothesized that the COG4708 motif is a preQ1-sensing riboswitch because some genes in COG4708 are associated with a previously characterized riboswitch (122). Preliminary experiments support this hypothesis (M. Meyer *et al.*, unpublished data). It is interesting to note that the COG4708 motif shares no similarity in sequence or structure with the previously characterized preQ1-sensing riboswitch, as in the case of the SAM riboswitch family. Discovery of this motif implies that the existence of significant structural variants of SAM riboswitches is not an isolated case. It may be a far more prevalent phenomenon than we previously thought, and it is very likely we will discover TPP-II, Lysine-II, etc. in the near future.

- The sucA motif

  The *sucA* motif is only found in the 5′ region of *sucA* genes. All detected instances of the sucA motifs are in β-proteobacteria in the order Burkholderiales. Although many nucleotides in the sucA motif are strictly conserved, those that are not show covariation and contain very few non-canonical base pairs (Figure 5.10). The motif has stems that overlap the putative Shine-Dalgarno sequence, so the sucA motif probably corresponds to a *cis*–regulatory RNA. Note that the exact position of the putative Shine-Dalgarno sequence is inconsistent among sucA motif instances, so is not well reflected in Figure 5.10. The relatively complex structure of the sucA motif suggests that it might be a riboswitch. However, it is difficult to evaluate its degree of sequence and structure conservation since the motif is not broadly distributed.

5.4 Discussion

In this study, we have presented a method for automatically finding *cis*–regulatory RNA motifs in prokaryotes. In a careful test with available sequenced Firmicutes, the method exhibited excellent rejection of negative controls (randomly permuted alignments) and excellent recovery of known, experimentally validated ncRNAs, including most riboswitches known in this bacterial group, as well as RNA elements such as 6S that have only recently

been recognized there. Careful inspection and refinement of several novel motifs in ribosomal protein leaders provides compelling evidence that they are indeed conserved structures involved in regulation of these important operons.

We have made more exciting discoveries when extending the search to all bacteria groups. This study produced 22 novel RNA-like motifs that fit a very stringent set of selection criteria. These include six riboswitch candidates, and five have been validated experimentally to this date. For several others, covariation and other evidence suggests that they are functional RNAs, and we have proposed reasonable hypotheses for many of the motifs. Thus, our pipeline greatly assists the discovery of novel RNAs, which in turn will contribute to our understanding of RNA biochemistry and bacterial gene regulation.

We attribute the power of this pipeline to two key characteristics—a relaxation of the constraints on sequence conservation imposed by most previous methods, and integration of motif inference with genome-scale search. Our method performs motif inference on regions that are not defined by sequence conservation: we search unaligned sequences upstream of homologous genes, instead of multiple sequence alignments constructed by sequence comparison tools. Additionally, both the RNA motif finding algorithm CMfinder and the RNA homology search algorithms RaveNnA/Infernal exploit structural information. Sequence conservation can be used as well, but is not required. Finally, automatic refinement of motifs to incorporate genome-scale search results has proven to be a powerful component of the pipeline (as in other contexts, such as PSI-BLAST (2)). The integration of these tools enables us to discover RNA motifs with low sequence conservation, and to expand the motif family with remote homologs. For example, the predicted motif for the glycine riboswitch (*gcvT* family) has only 35% average pairwise sequence similarity. Remote RNA homologs with appropriate gene context are particularly important as they are the strongest evidence, short of experiments, that a motif is functional, as well as providing clues to that function.

Future work will seek to strengthen this pipeline by improved exploitation of phylogeny in scoring and motif search. Phylogeny is crucial in all comparative genome analysis, without which the concept of conservation is meaningless. It is important in our work because the sequences upon which motif inference is performed are not evolutionarily equidistant, and the significance of conserved nucleotides and compensatory mutations are distance-

dependent. As we have discussed in Chapter 4, the classic phylogenetic likelihood model can be used to evaluate the significance level of structural conservation. Unfortunately, in our application neither an alignment nor an evolutionary tree are initially available, and, for our application, use of the corresponding species tree is questionable in the common case when there are multiple sequences per species. Incorporating phylogeny into motif search is another challenge.

Our pipeline is designed to discover structured RNAs that are widespread, highly conserved and structured, which it has achieved very successfully. The fact that we have found six novel riboswitch candidates with these characteristics, compared to 12 known riboswitch classes, suggests either that (1) there are relatively few unknown riboswitches, at least among the sequenced organisms, or (2) most as-yet-undiscovered riboswitches will not follow these characteristics. Although our pipeline has been successful at finding many novel RNAs, the latter possibility may motivate approaches that can eliminate some of the blind spots of our current approach. To tackle applications in which the ncRNAs are not located upstream of homologous genes, or there are relatively few sequenced genomes available, we need to modify the pipeline significantly. How to do so effectively remains an open question.

These opportunities for improvement notwithstanding, the approach described in this study has proven itself to be highly effective in discovering noncoding RNA elements in prokaryotes, and promises more discoveries to come.

Table 5.1: Motifs that correspond to Rfam families. "Rank": the three columns correspond respectively to ranks for refined motifs after genome scans ("RV"), CMfinder motifs before genome scans ("CM"), and FootPrinter results ("FP"). We used the same ranking scheme for "RV" and "CM", and here ranks actually refer to the ranks of the motif clusters. "Score": CMfinder composite motif score (after refinement). "#": The number of motif instances after genome scan ("RV"), and before genome scan ("CM"). "CDD Gene: Description": Conserved Domain Database PSSM-ID (accession), name and description of an examplery gene. "Rfam": Rfam accession and family name. The genome scan here refers to the "mini" scan (rather than full scan) described in the method section. [1]A few tRNAs partially outside the limits of the collected upstream regions evaded our masking procedure.

Rank			Score	#		CDD Gene: Description	Rfam
RV	CM	FP		RV	CM		
0	43	107	3400	367	11	9904 IlvB: Thiamine pyrophosphate-requiring enzymes	RF00230 T-box
1	10	344	3115	96	22	13174 COG3859: Predicted membrane protein	RF00059 THI
2	77	1284	2376	112	6	11125 MetH: Methionine synthase I	RF00162 S_box
3	0	5	2327	30	26	9991 COG0116: Predicted N6-adenine-specific DNA methylase	RF00011 RNaseP_bact_b
4	6	66	2228	49	18	4383 DHBP_synthase 3:	RF00050 RFN
7	145	952	1429	51	7	10390 GuaA: GMP synthase	RF00167 Purine
8	17	108	1322	29	13	10732 GcvP: Glycine cleavage system protein P	RF00504 gcvT
9	37	749	1235	28	7	24631 DUF149: Uncharacterised BCR	RF00169 SRP_bact
10	123	1358	1222	36	6	10986 CbiB: Cobalamin biosynthesis protein CobD/CbiB	RF00174 Cobalamin
20	137	1133	899	32	7	9895 LysA: Diaminopimelate decarboxylase	RF00168 Lysine
21	36	141	896	22	10	10727 TerC: Membrane protein TerC	RF00080 yybP-ykoY
39	202	684	664	25	5	11945 MgtE: Mg/Co/Ni transporter MgtE	RF00380 ykoK
40	26	74	645	19	18	10323 GlmS: Glucosamine 6-phosphate synthetase	RF00234 glmS
53	208	192	561	21	5	10892 OpuBB: ABC-type proline/glycine betaine transport systems	RF00005 tRNA[1]
122	99	239	413	10	7	11784 EmrE: Membrane transporters of cations and cationic drug	RF00442 ykkC-yxkD
255	392	281	268	8	6	10272 COG0398: Uncharacterized conserved protein	RF00023 tmRNA

Table 5.2: Motif prediction accuracy compared to Rfam. All comparisons are to the prokaryotic subset of Rfam full alignments. Under "Membership", the three numbers are: the number of overlapped sequences between our predictions and Rfam's ("#"), the sensitivity ("Sn") and specificity ("Sp") of our membership predictions. Under "Overlap", the numbers are the average length of overlap between our predictions and Rfam's ("nt"), the percentage length of the overlapped region in Rfam's predictions ("Sn"). and in ours ("Sp"), Under "Structure", the numbers are the average number of correctly predicted canonical base pairs (in overlapped regions) in the secondary structure ("bp"), and the sensivity ("Sn") and specificity ("Sp") of our predictions. [1]After another iteration of RaveNnA scan and refinement, the sensitivities of gcvT and Cobalamin increase to 76% and 98% respectively, while the specificity of gcvT remains the same, and specificity of Cobalamin dropped to 84%).

Rfam	Membership			Overlap			Structure		
	#	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174 Cobalamin	183	0.74[1],	0.97	152	0.75,	0.85	20	0.60,	0.77
RF00504 gcvT	92	0.56[1],	0.96	94	0.94,	0.68	17	0.84,	0.82
RF00234 glmS	34	0.92,	1.00	100	0.54,	1.00	27	0.96,	0.97
RF00168 Lysine	80	0.82,	0.98	111	0.61,	0.68	26	0.76,	0.87
RF00167 Purine	86	0.86,	0.93	83	0.83,	0.55	17	0.90,	0.95
RF00050 RFN	133	0.98,	0.99	139	0.96,	1.00	12	0.66,	0.65
RF00011 RNaseP_bact_b	144	0.99,	0.99	194	0.53,	1.00	38	0.72,	0.78
RF00162 S_box	208	0.95,	0.97	110	1.00,	0.69	23	0.91,	0.78
RF00169 SRP_bact	177	0.92,	0.95	99	1.00,	0.65	25	0.89,	0.81
RF00230 T-box	453	0.96,	0.61	187	0.77,	1.00	5	0.32,	0.38
RF00059 THI	326	0.89,	1.00	99	0.91,	0.69	13	0.56,	0.74
RF00442 ykkC-yxkD	19	0.90,	0.53	99	0.94,	0.81	18	0.94,	0.68
RF00380 ykoK	49	0.92,	1.00	125	0.75,	1.00	27	0.80,	0.95
RF00080 yybP-ykoY	41	0.32,	0.89	100	0.78,	0.90	18	0.63,	0.66
mean	145	0.84,	0.91	121	0.81,	0.82	21	0.75,	0.77
median	113	0.91,	0.97	105	0.81,	0.83	19	0.78,	0.78

Table 5.3: High ranking motifs not found in Rfam. For each we give its rank (after refinement), # of sequences containing the motif, the CDD dataset from which it was found, name and description of an examplery gene, and annotations. Two of the ribosomal protein leader motifs are described in the Results section, and 5 others in the online supplement.

Rank	#	CDD	Gene: Description	Annotation
6	69	28178	DHOase_IIa: Dihydroorotase	PyrR attenuator; (87)
15	33	10097	RplL: Ribosomal protein L7/L1	L10 r-protein leader; see suppl.
19	36	10234	RpsF: Ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: Dinucleotide-utilizing enzymes	6S RNA; (9)
27	27	9926	RpsJ: Ribosomal protein S10	S10 r-protein leader; see suppl.
29	11	15150	Resolvase: N terminal domain	
31	31	10164	InfC: Translation initiation factor 3	IF-3 r-protein leader; see suppl.
41	26	10393	RpsD: Ribosomal protein S4 and related proteins	S4 r-protein leader (60); see suppl.
44	30	10332	GroL: Chaperonin GroEL	HrcA DNA binding site; (102)
46	33	25629	Ribosomal_L21p: Ribosomal prokaryotic L21 protein	L21 r-protein leader; see suppl.
50	11	5638	Cad: Cadmium resistance transporter	(37)
51	19	9965	RplB: Ribosomal protein L2	S10 r-protein leader
55	7	26270	RNA_pol_Rpb2_1: RNA polymerase beta subunit	
69	9	13148	COG3830: ACT domain-containing protein	
72	28	4174	Ribosomal_S2: Ribosomal protein S2	S2 r-protein leader
74	9	9924	RpsG: Ribosomal protein S7	S12 r-protein leader
86	6	12328	COG2984: ABC-type uncharacterized transport system	
88	19	24072	CtsR: Firmicutes transcriptional repressor of class III	CtsR DNA binding site; (26)
100	21	23019	Formyl_trans_N: Formyl transferase	
103	8	9916	PurE: Phosphoribosylcarboxyaminoimidazole	
117	5	13411	COG4129: Predicted membrane protein	
120	10	10075	RplO: Ribosomal protein L15	L15 r-protein leader
121	9	10132	RpmJ: Ribosomal protein L36	IF-1 r-protein leader
129	4	23962	Cna_B: Cna protein B-type domain	
130	9	25424	Ribosomal_S12: Ribosomal protein S12	S12 r-protein leader
131	9	16769	Ribosomal_L4: Ribosomal protein L4/L1 family	L3 r-protein leader
136	7	10610	COG0742: N6-adenine-specific methylase	ylbH putative RNA motif; (8)
140	12	8892	Pencillinase_R: Penicillinase repressor	BlaI, MecI DNA binding site; (132)
157	25	24415	Ribosomal_S9: Ribosomal protein S9/S16	L13 r-protein leader; see Fig. 5.4
160	27	1790	Ribosomal_L19: Ribosomal protein L19	L19 r-protein leader; see Fig. 5.3
164	6	9932	GapA: Glyceraldehyde-3-phosphate dehydrogenase/erythrose	
174	8	13849	COG4708: Predicted membrane protein	
176	7	10199	COG0325: Predicted enzyme with a TIM-barrel fold	
182	9	10207	RpmF: Ribosomal protein L32	L32 r-protein leader
187	11	27850	LDH: L-lactate dehydrogenases	
190	11	10094	CspR: Predicted rRNA methylase	
194	9	10353	FusA: Translation elongation factors	EF-G r-protein leader

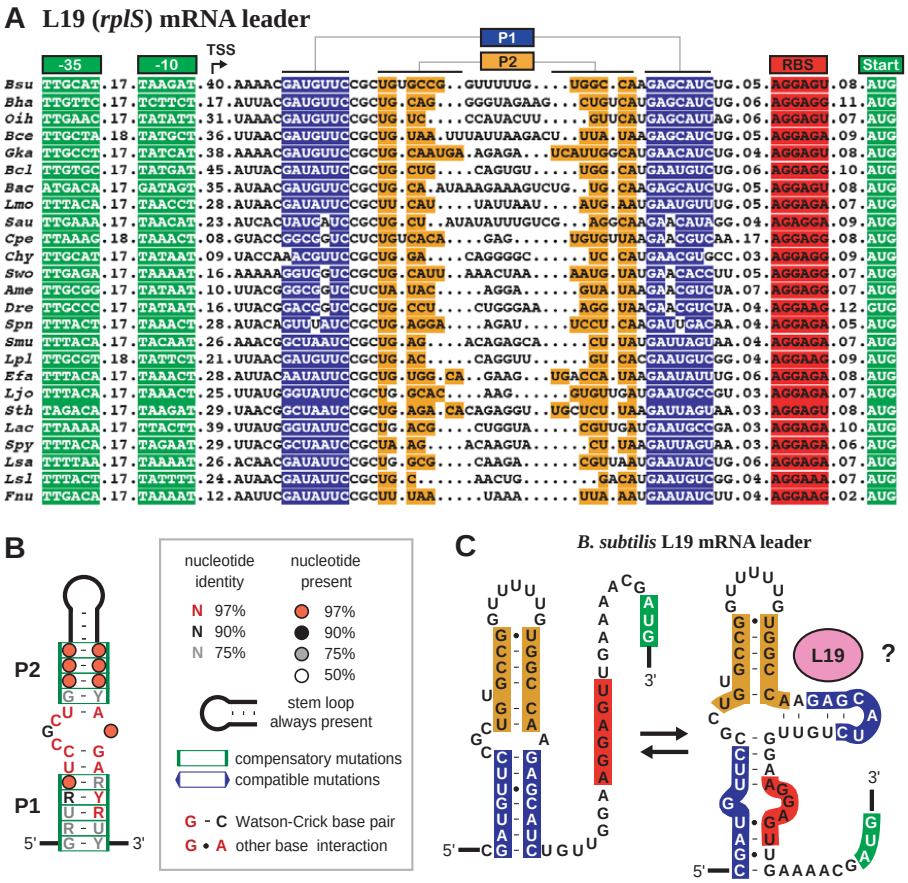


Figure 5.3: Putative autoregulatory structure in L19 mRNA leaders. (A) Sequence alignment of a conserved RNA structure found in the 5′ UTR of Firmicutes *rplS* genes. Possible promoter -35 and -10 boxes in genomic DNA are shown, followed by the putative mRNA leader with the predicted secondary structures (P1 and P2), ribosome binding sites, and start codons highlighted. Numbers represent inserted nucleotides that are not shown. The examples shown are representative of 34 total sequences in the complete alignment, available in the online supplement. Species abbreviations: *Bsu, Bacillus subtilis; Bha, Bacillus halodurans; Oih, Oceanobacillus iheyensis; Bce, Bacillus cereus; Gka, Geobacillus kaustophilus; Bcl, Bacillus clausii; Bac, Bacillus sp. NRRL; Lmo, Listeria monocytogenes; Sau, Staphylococcus aureus; Cpe, Clostridium perfringens; Chy, Carboxydothermus hydrogenoformans; Swo, Syntrophomonas wolfei; Ame, Alkaliphilus metalliredigenes; Dre, Desulfotomaculum reducens; Spn, Streptococcus pneumoniae; Smu, Streptococcus mutans; Lpl, Lactobacillus plantarum; Efa, Enterococcus faecalis; Ljo, Lactobacillus johnsonii; Sth, Streptococcus thermophilus; Lac, Lactobacillus acidophilus; Spy, Streptococcus pyogenes; Lsa, Lactobacillus sakei; Lsl, Lactobacillus salivarius; Fnu, Fusobacterium nucleatum.* (B) Consensus sequence and secondary structure. Pairs supported by compensatory (when both bases in a pair mutate between sequences in the alignment) and compatible (when only one base mutates but pairing is preserved, e.g. G-C to G-U) are boxed. (C) Structural model of the *B. subtilis* L19 mRNA leader, showing a possible alternate structure that could be stabilized by L19 binding to repress translation.

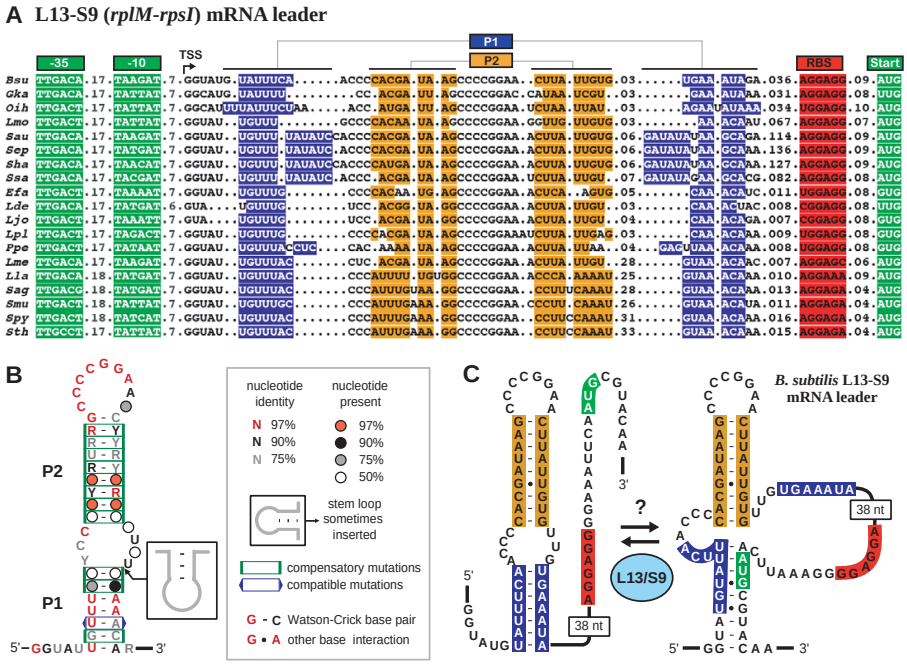


Figure 5.4: Putative autoregulatory structure in L13-S9 mRNA leaders. (A) Sequence alignment of a conserved RNA structure found in the 5′ UTR of Firmicutes *rplM-rpsI* operons. The examples shown are representative of 27 total sequences in the complete alignment, available in the online supplement. Details are as in the legend for Fig 5.3 with additional species abbreviations: *Sep, Staphylococcus epidermidis; Sha, Staphylococcus haemolyticus; Ssa, Staphylococcus saprophyticus; Lde, Lactobacillus delbruecki; Ppe, Pediococcus pentosaceus; Lme, Leuconostoc mesenteroides; Lla, Lactococcus lactis; Sag, Streptococcus agalactiae.* (B) Consensus sequence and secondary structure. (C) Structural model of the *B. subtilis* L13-S9 mRNA leader, showing a possible alternate structure that could be stabilized by L13 or S9 binding to repress translation.

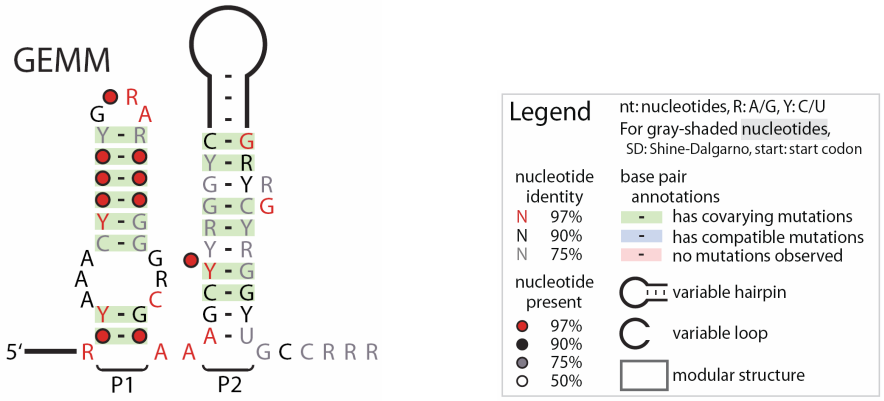


Figure 5.5: Consensus sequence and secondary structure model of the GEMM motif.

Table 5.4: Summary of putative structured RNA motifs. RNA = functions as RNA (as opposed to dsDNA), Cis = cis-regulatory, Switch = riboswitch. Evaluation: Y = certainly true, y = probably true, ? = possible, n = probably not, N = certainly not. Remaining columns are Phylum/class (phylum containing the motif, or class for Proteobacteria), M,V (M = has modular stems, which are stems that are only sometimes present, V = variable-length stems), Cov. = number of covarying paired positions, # = number of representatives, Non-cis = X/Y where X is number of representatives that are not in a 5' regulatory configuration to a gene and Y is the number of representatives within sequences that have annotated genes (some RefSeq sequences lack annotations).

Motif	RNA?	Cis?	Switch?	Phylum/class	M,V	Cov.	#	Non-cis
GEMM	Y	Y	y	Widespread	V	21	322	12/309
Moco	Y	Y	Y	Widespread	M,V	15	105	3/81
SAH	Y	Y	Y	Proteobacteria	M,V	22	42	0/41
SAM-IV	Y	Y	Y	Actinobacteria	V	28	54	2/54
COG4708	Y	Y	y	Firmicutes	M,V	8	23	0/23
sucA	Y	Y	y	β-proteobacteria		9	40	0/40
23S-methyl	Y	y	n	Firmicutes		12	38	1/37
hemB	Y	?	?	β-proteobacteria	V	12	50	2/50
(anti-hemB)		(n)	(n)				(37)	(31/37)
MAEB	?	Y	n	β-proteobacteria		3	662	15/646
mini-ykkC	Y	Y	?	Widespread	V	17	208	1/205
purD	y	y	?	ϵ-proteobacteria	M	16	21	0/20
6C	y	?	n	Actinobacteria		21	27	1/27
alpha-transposases	?	N	N	α-proteobacteria		16	102	39/99
excisionase	?	?	n	Actinobacteria		7	27	0/27
ATPC	y	?	?	Cyanobacteria		11	29	0/23
Cyano-30S	Y	Y	n	Cyanobacteria		7	26	0/23
lacto-1	?	?	n	Firmicutes		10	97	18/95
lacto-2	y	N	n	Firmicutes		14	357	67/355
TD-1	y	?	n	Spirochaetes	M,V	25	29	2/29
TD-2	y	N	n	Spirochaetes	V	11	36	17/36
coccus-1	?	N	N	Firmicutes		6	246	112/189
gamma-150	?	N	N	γ-proteobacteria		9	27	6/27

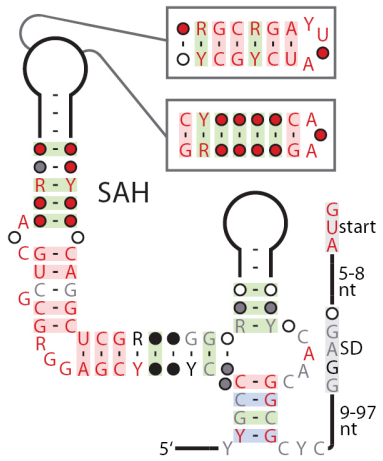


Figure 5.6: Consensus sequence and secondary structure model of the SAH motif. See Figure 5.5 for figure legends.

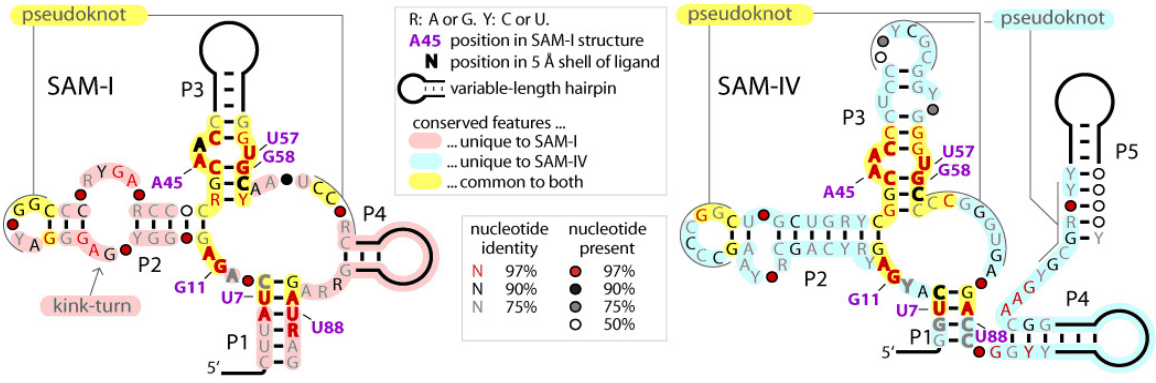


Figure 5.7: Comparison of the secondary structures of SAM-I and SAM-IV motifs

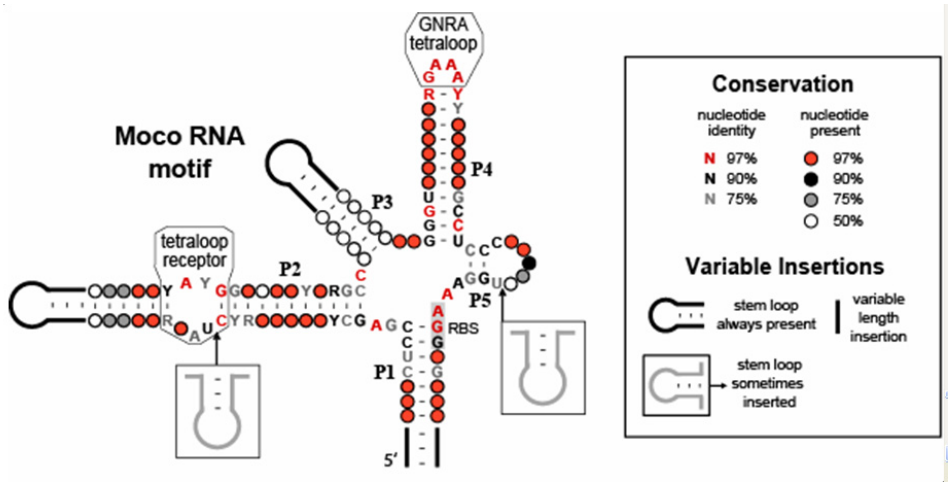


Figure 5.8: Consensus sequence and secondary structure model of the Moco RNA motif. Boxed nucleotides are predicted to be the ribosome binding site for the adjacent ORF in some Moco RNA representatives.

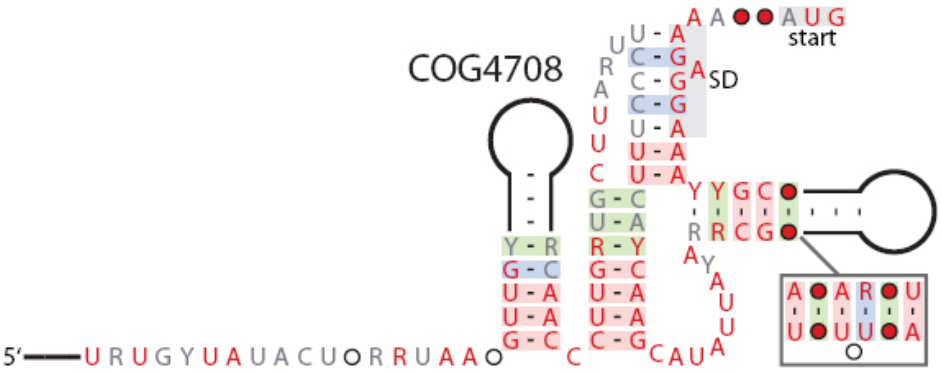


Figure 5.9: Consensus sequence and secondary structure model of the COG4708 motif. See Figure 5.5 for figure legends.

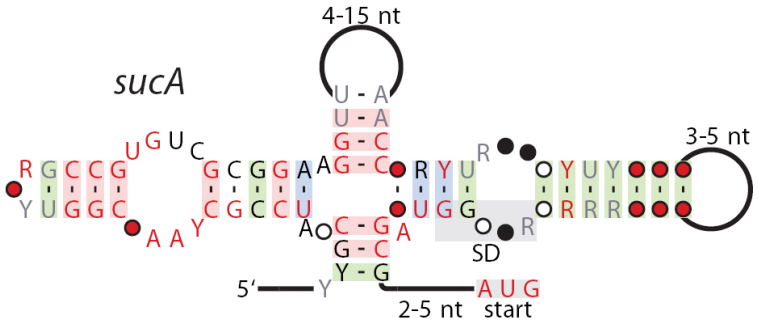


Figure 5.10: Consensus sequence and secondary structure model of the *sucA* motif. See Figure 5.5 for figure legends.

Chapter 6

NONCODING RNA DISCOVERY IN VERTEBRATES

6.1 Introduction

The computational prediction of ncRNAs in Eukaryotes based on comparative genomic analysis has become increasingly attractive as more genomes become available. Several important studies on ncRNA discovery in higher organisms have been reported in the last few years. For example, RNAz and Evofold have been used on Vertebrate and Drosophila genomes (151; 109; 143; 121; 138). These methods adopted a similar paradigm, which is to scan the given multiple sequence alignments for conserved RNA secondary structure signal. As we discussed previously, this strategy is subject to alignment errors, which are likely to occur as sequence conservation level decreases. Therefore, many previous scans were restricted to highly conserved regions of the multiple alignments to minimize effects of alignment errors. An additional concern is that these methods generally assume that an RNA structure, if present, is present in all sequences in the alignment, ignoring the possibility of gain or loss on some branches of the phylogeny. Finally, both RNAz and Evofold initially evaluate only global alignments within fixed width sliding windows, which further reduces sensitivity since a given placement of the window may include extraneous sequence flanking a given RNA structure, or may include only part of the structure, or both. Therefore, although these studies have successfully turned up thousands of candidates, it is very likely that they may have missed a significant fraction of real ncRNAs due to their inherent technical limitations.

As one of the first attempt to fill the gap of previous methods, Torarinsson et al. (143) used FOLDALIGN to search conserved RNA structures in (presumably) syntenic regions between human and mouse that are not aligned in the UCSC MULTIZ alignments. This study produced thousands of candidates, suggesting the existence of potentially a large set of ncRNAs that were totally ignored before. While this has been an important mile-

stone, FOLDALIGN has its own limitations. Derived from the classical Sankoff algorithm (124), FOLDALIGN (52; 64) is a dynamic programming algorithm that performs structure prediction and alignment simultaneously. The key weakness of this problem is its computational overhead, which makes it applicable only on relatively small datasets. In addition, FOLDALIGN is more effective for pairwise alignment than multiple alignment, therefore, cannot exploit the full potential of the available genomic sequences.

These observations led us to apply CMfinder as a complement to the RNAz/EvoFold scans, and as a more practical alternative to FOLDALIGN. Compared to previous studies, we do not rely on externally supplied alignments (except to indicate orthology), do not use a sliding window approach, and can ignore diverged sequences that do not appear to share the discovered RNA motif. We have applied CMfinder for ncRNA search within the ENCODE (ENCyclopedia Of DNA Elements) regions (22), which include 30MB, or roughly 1% of the human genome, as a pilot study prior to the full scan of the human genome.

In agreement with the previous studies, we found a large number of potential RNA structures, totaling 6,587 candidate regions with an estimated false discovery rate of 50%. More intriguingly, many of our predicted motifs may be better represented by alignments taking the RNA secondary structure into account than those based on primary sequence alone, often quite dramatically. For example, approximately one quarter of our motifs show alignment revisions in more than 50% of their positions, in comparison to the sequence-based MULTIZ alignments. Furthermore, while overlap with the candidates generated by the RNAz scan is much greater than would be expected by chance, our predictions are largely complementary to those from the RNAz/EvoFold scans, with 84% nonoverlapping candidate regions. These results broadly suggest caution in any analysis relying on multiple sequence alignments in less well-conserved regions, and strongly argue for taking RNA structure directly into account in any searches for these elements. This study is conducted in collaboration with Jan Gorodkin's lab from University of Copenhagen, and was reported at (142). Elfar Torarinsson contributed to most of the computational analysis, with other collaborators performing wet-lab verifications.

6.2 Results

6.2.1 The candidates

We scanned 56,017 (forward/reverse) multiple alignment blocks from the UCSC MULTIZ multiple alignment (.maf) files, one block at a time (155nts long on average). Since previous studies were presumed to be effective in well-conserved regions, we restricted analysis to alignment blocks which overlap neither exons nor the most conserved elements (as defined by the PhastCons Conserved Elements (Siepel et al. 2005) available at `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/phastConsElements17way.txt.gz`).

These alignments covered 8.68 Mb of human sequence (out of the total of 30 Mb in the ENCODE regions), and included 3.87Mb of repeat sequence. We included alignments in repeat regions in human because many of the known ncRNAs are found there. This resulted in 10,106 predicted motifs that met our cutoff criteria: a composite score above 5 and free energy below -5 kcal/mol. We estimated a false discovery rate of 50% by repeating the analysis on shuffled alignments. Composite score and energy distributions for randomized vs. original alignments are depicted in Figure 6.1, showing a slight shift in the distribution towards lower energy and higher score for our native predictions. Some of these predicted motifs overlap or are sense/antisense to each other. Considering these as a single candidate region we have 6,587 candidate regions. Our candidate regions average 80nt in length, collectively covering a total of 0.53 MB, or 6.1% of our human input sequence. Candidate regions are approximately twice as dense (per nucleotide) in non-repeat regions (0.38 MB or 7.9%) as in repeat regions (0.15 MB or 3.9% of the repeat input set).

6.2.2 Known ncRNAs

As noted by Washietl et al. (2007) the ENCODE regions are surprisingly poor in annotated ncRNAs. In fact, comparing to Rfam (58), the Functional RNA project (www.ncRNA.org), and the snoRNA and miRNA tracks that have been mapped to the human genome by the UCSC Genome Browser, we could only find one ncRNA which fully overlapped our input alignments. This was the miRNA hsa-miR-483 on chromosome 11 identified by Fu et al.

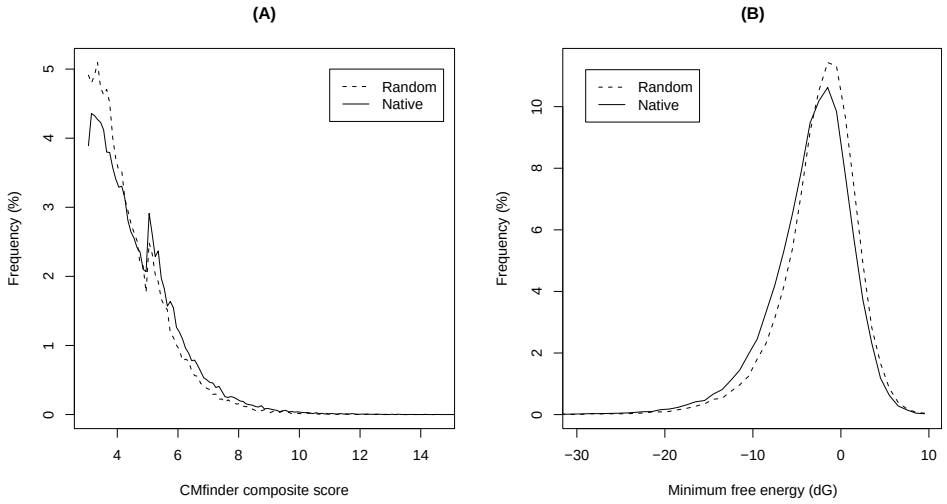


Figure 6.1: Score distribution of the full CMfinder motif set (A) composite score and (B) consensus minimum free energies, for the native and random (shuffled) sequences. There is a slight shift towards lower energy and higher score for our native data.

(43) in fetal liver in human. In addition miR-483 has been annotated in mouse and rat "by similarity" in miRBase (56; 54; 55). This miRNA was detected in our scan (composite score 8.6, energy -31.4) and was scored highly as a miRNA by RNAmicro (66), which we ran on all our predictions. Our prediction, in addition to human, rat and mouse, also includes dog, cow and rabbit. Hsa-miR-483 was also detected by RNAz but was not in the input set for EvoFold (150).

6.2.3 Tiling-array data and purifying selection

Using oligonucleotide tiling array techniques, transcription maps of TARs (transcriptionally active regions) (11) and Transfrags (transcribed fragments) (19) have been generated. We compared our predictions to TARs and Transfrags generated as a part of the ENCODE project, which used 11 human tissues (ENCODE Consortium, 2007 (22)). Note that these

maps were derived from RNA fragments longer than 200 nucleotides. TARs and Transfrags were only generated for the non-repetitive regions of the genome whereas we included the repeat regions, so candidates in repeat regions (25% of our total candidate regions) were ignored in calculating the following numbers. 16.9% of these candidate regions overlap TARs/Transfrags. At the nucleotide level, 11.8% of the bases in the predictions overlap a TAR or a Transfrag, compared to 7.0% of the input bases (that is, our whole non-repeat input data).

A recent study by Kapranov et al. (74) investigated the genomic origins and the associations of human nuclear and cytosolic polyadenylated RNAs longer than 200 nucleotides (lRNA) and whole-cell RNAs less than 200nt (sRNA). Comparing our candidate regions to these new transfrags, on the nucleotide level, 3.0% and 27.4% of our candidates were overlapped by short and long RNAs respectively, compared to 1.5% and 16.0% of the input bases. The increased overlap with TARs/Transfrags, sRNA and lRNA is highly significant with p-values of 10^{-40}, 10^{-24} and 10^{-86}, respectively. Still, one has to be cautious since, as noted by Washietl et al. in (150), array technologies are typically more sensitive on GC-rich regions, and as a result, TARs/Transfrags are GC-rich. With this in mind we divided our input data into five GC bins containing similar numbers of multiple alignment blocks (0-35%, 35-40%, 40-45%, 45-50%, 50-100%) and repeated our analysis. Within each of the five GC bins, there was no significant overlap with the tiling-array data. We also did the same analysis for the RNAz and EvoFold candidates that are contained in our input data, and came to the same conclusion for their candidates. Washietl et al. (150) pointed out that it is unclear if the GC bias for tiling-array data has biological reasons, and it is also unclear how secondary structure affects detection performance on tiling-arrays, considering several cases in which highly stable ncRNAs result in negative signal "holes" in tiling-array data (19).

Ponjavic et al. (112) studied the noncoding regions apparently under purifying selection, based on lack of indels and other evidence. Even though homologous ncRNAs allow indels, as we have observed in bacteria, the distribution of indels are not uniformly distributed, and some ncRNAs have far fewer indels than expected by chance. We compared our candidate regions to their set of Indel Purified Segments (IPSs) on human assembly hg18. For our two

Table 6.1: Pvalues of overlap between CMfinder candidates and multiple transcription related datasets, stratified by GC bins

Dataset	0-35	35-40	40-45	45-50	50-100	All
EST	7.71E-01	7.36E-01	5.86E-01	3.60E-01	5.18E-01	9.25E-09
TAR/Trans	6.42E-01	8.11E-01	2.59E-01	6.09E-01	2.99E-02	5.46E-40
lRNA	3.49E-01	6.55E-01	4.23E-01	4.55E-03	3.52E-01	4.08E-86
sRNA	2.13E-01	9.33E-01	1.75E-03	3.22E-01	1.83E-02	3.33E-24
IPS	4.30E-01	1.03E-01	7.91E-04	5.16E-08	2.70E-31	1.54E-33
EvoFold	5.54E-01	7.34E-01	1.48E-02	5.20E-06	6.76E-06	2.56E-06
RNAz	4.40E-23	2.17E-18	4.14E-29	3.20E-28	1.43E-40	3.72E-244
$3'$ UTR	4.35E-01	9.39E-01	6.58E-01	9.95E-02	9.53E-01	9.98E-01
$5'$ UTR	9.29E-01	9.90E-01	5.34E-01	3.80E-01	4.23E-07	1.28E-02
Introns	8.06E-01	8.21E-01	1.08E-01	6.61E-01	2.78E-02	4.88E-01

most GC-rich bins (where the majority of our candidate regions lie), there is a significant overlap to the IPSs ($P < 10^{-8}$ and $P < 10^{-31}$), indicating that many of our candidate regions are under indel purifying selection.

The pvalues for overlap with CMfinder candidates and other datasets suggestive of ncRNAs are shown in table 6.1, stratified based on GC content.

6.2.4 GENCODE

We also compared our candidate regions to the GENCODE annotations (63), whose overall goal is to identify all protein-coding genes in the ENCODE selected regions of the human genome. We found that 40% of our candidates are intergenic whereas 60% overlap some non-exonic part of a protein coding gene (see Table 6.2). We analyzed if introns, $3'$ UTRs or $5'$ UTRs were enriched for our candidate regions using the five GC bins. Only for the GC bin of 50-100%, there is significant enrichment of predicted candidate regions in 5' UTRs ($P < 10^{-7}$) (see Table 6.1). There are also 23 candidates that overlap with an exon, because we use the GENCODE annotation here, whereas our initial filtering was done with

Table 6.2: GENCODE overlaps

Sense	Antisense	Both	Intron	5′ UTR	3′ UTR
1721(43.7%)	1332(33.8%)	884(22.5%)	3274(83.1%)	551(14%)	89(2.3%)

UCSC known genes annotation.

6.2.5 RNAz and EvoFold

As mentioned earlier, a similar scan to ours was performed with the global, alignment-dependent programs RNAz and EvoFold (150). Note that they use the TBA (Threaded Blockset Aligner) non-repetitive multiple sequence alignments with up to 28 species as prepared by the ENCODE alignment group (93), whereas we used the MULTIZ alignments (with autoMZ driver) with up 17 species available at the UCSC genome browser. In both cases the alignments are prepared using the TBA/MULTIZ software (13). We used the latest assemblies (human hg18) whereas Washietl et al. (150) use earlier assemblies (human hg17) because the TBA ENCODE alignments are only available for hg17. We used hg18 because it was the latest assembly with genome wide multiple alignments available. Furthermore, the input alignments for RNAz and EvoFold were pre-processed according to different preferences of these programs (150).

To compare our predictions with those of RNAz and EvoFold, we used all their candidates (low and high confidence) that overlapped neither exons nor the PhastCons conserved elements (38% of their total predictions) (133), and compared them to our 4933 (75% of our total candidate regions) non-repetitive candidate regions. 6.7% of these candidate regions overlap with EvoFold predictions, with P-value of $2.56 \cdot 10^{-6}$, whereas 17.2% overlap with RNAz candidates with P-value of $3.72 \cdot 10^{-244}$ (see Figure 6.2). We suspected that the significance of these P-values are due to GC composition bias within both our predictions and predictions of Evofold and RNAz. Therefore, we re-calibrated the significance of overlap within each of the five GC bins, as defined in section 6.2.3. For the two most GC-rich GC bins (45-50% and 50-100%), the overlap with EvoFold was significant ($P < 10^{-5}$ in both

bins). The overlap with RNAz was significant in all five GC bins ($P < 10^{-22}$, $P < 10^{-17}$, $P < 10^{-28}$, $P < 10^{-27}$ and $P < 10^{-39}$, ordered by increasing GC percentage). In the regions that do not overlap exons, PhastCons conserved elements, or repeat regions, we add 3861 new candidates to the 6071 RNAz or EvoFold candidates. Furthermore, we predict 1654 candidates in regions that are in repeat regions in human (excluded by the RNAz scan (150)) and thereby add 5515 candidates to the 17,046 RNAz or EvoFold candidates in the ENCODE regions, corresponding to 32% of the total number of candidates.

EvoFold has a strong preference for TA-rich regions whereas RNAz prefers GC-rich regions since the Gibbs energy is important to RNAz. The CMfinder predictions are approximately normally distributed, centered on 53% GC content. Still, when considering that the background GC content is 43%, it is clear that CMfinder also prefers GC rich regions which tend to be more structurally stable.

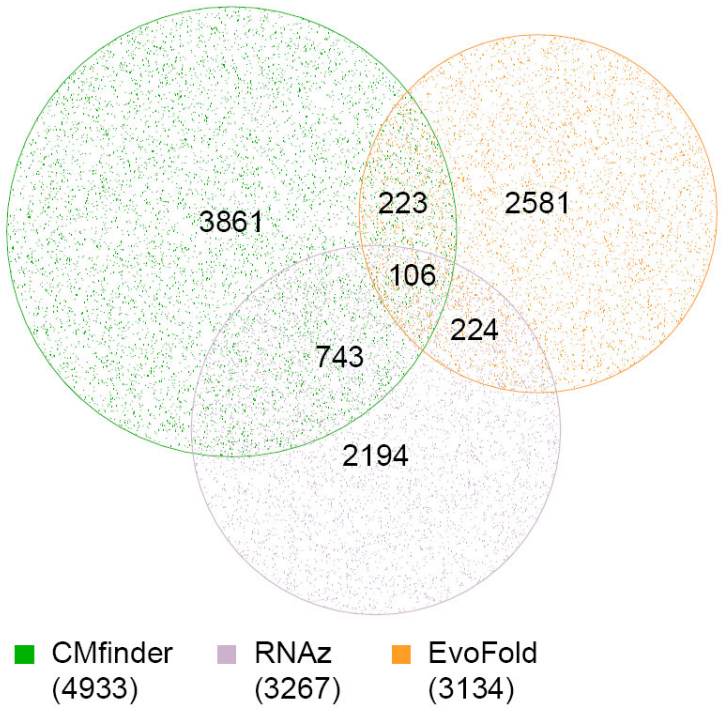


Figure 6.2: Overlap of predictions made by CMfinder, RNAz and EvoFold. Only predictions that are not highly conserved (defined by phastCons), outside exons and repeat regions are considered, as they are the common subset of the input to three programs. The total number for each program is indicated in parentheses below the label.

6.2.6 Candidate Database

All of our candidate regions are available in an online database (`http://genome.ku.dk/resources/cmf_encode`). The database includes a variety of additional annotations such as the overlaps described above, occurrences such as conserved tetraloop motifs and predicted microRNA using RNAmicro. The database also supports easy access to subsets of the candidates with different features. For example, one can easily retrieve all candidates overlapping TARs/Transfrags or all miRNA predictions. Furthermore, each candidate region is linked directly to the UCSC genome browser. Despite the relatively high false discovery rate, it is possible to use the information in our database to select higher confidence predictions through the "Database Search" link. For example, one can choose predictions that overlap with EvoFold/RNAz predictions and/or overlap TARs/Transfrags.

6.2.7 Re-aligning parts of the genomes

A benchmark study by Gardner et al. (45) compared the relative performances of structure- versus sequence-based methods when aligning pairs of known tRNAs. The study revealed a dramatic divergence in performance for sequences with identity below 60%; i.e., sequence-based methods were dramatically worse below this threshold. Note that Gardner et al. define pairwise sequence identity as *IDENTITIES/MIN(length(A), length(B))* for sequences A and B (personal communication), whereas we, dealing with multiple alignments, define this as *IDENTITIES/MAX(length(A), length(B))*. IDENTITIES is the number of identical positions in the alignment and the length is the gap-free length of the sequence. For example, the sequences ATGC and AG are 100% identical by the former definition, but only 50% identical by the latter. Applying our definition to Gardner et al.'s data lowers the pairwise sequence identities by 3% on average. We use a different definition because gap distribution can affect the performance of the alignment methods significantly, which is taken into account more effectively by our definition. Although Gardner et al.'s observation is based on pairwise alignments on tRNAs, it is reasonable to extrapolate it to multiple sequence alignment. It suggests that one should be careful when searching for structural ncRNAs in sequence-based alignments when the sequence similarity drops below a certain threshold,

as alignment errors accumulate and propagate via sequence-dependent methods. CMfinder considers both sequence and structure information and is therefore expected to perform better on regions with low sequence similarity. Considering that the original input alignments have 50% average pairwise sequence similarity, they tend to benefit from re-alignment by taking structure into account, when the RNA secondary structure is of importance. We calculated the extent of the re-alignment suggested by CMfinder, compared to the original sequence-based alignment. As expected, the degree of re-alignment correlates with sequence similarity (Pearson correlation of > 0.77) (see Figure 6.3). Approximately one quarter of the alignments show re-alignment in more than 50% of positions (see Methods).

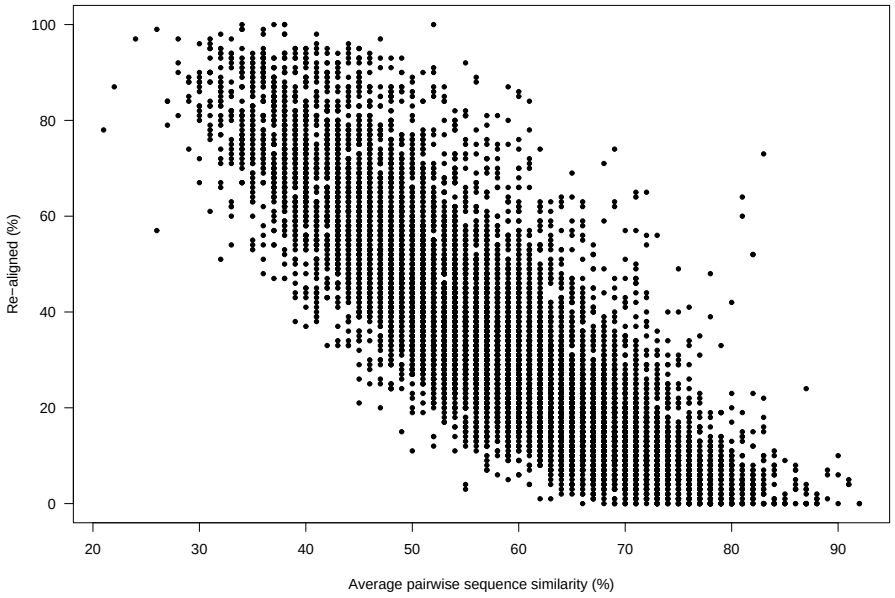


Figure 6.3: Average pairwise sequence similarity of the predicted motifs vs. the fraction that has been re-aligned compared to the original alignments.

Most of the known ncRNA families probably exhibit artificially high sequence similarities as most are discovered based on sequence-based methods. To demonstrate possible benefits of structure-aware alignment, we examined MULTIZ multiple alignment blocks identified

by Wang et al. (147) to contain matches to Rfam ncRNAs (58), with good matches to the Rfam model in all species in the same region of the alignment. In one example containing 10 mammals, with fairly high sequence identity (72%), neither EvoFold nor RNAz report a candidate there. However, CMfinder identifies a candidate (composite score > 5 and energy $<$ -5) in all 10 species in good general agreement with the H/ACA snoRNA known there (Rfam accession RF00402). See Figure 6.4 for the MULTIZ alignment and the CMfinder alignment. CMfinder's alignment of the region differs from the MULTIZ alignment in only 13% of positions, yet this change is sufficient to flip the RNAz prediction from negative ("RNA probability" 0.11, based on using their script to select 6 organisms) to strongly positive (probability 0.98). EvoFold did not predict anything for either alignment. While this is just one example, it does highlight the fact that even reasonably solid sequence-based alignments may not suffice for RNA discovery. Considering the high number of ENCODE region alignments with relatively low sequence similarities, it is reasonable to expect CMfinder, in many cases, to perform better on these alignments than sequence-alignment-dependent tools.

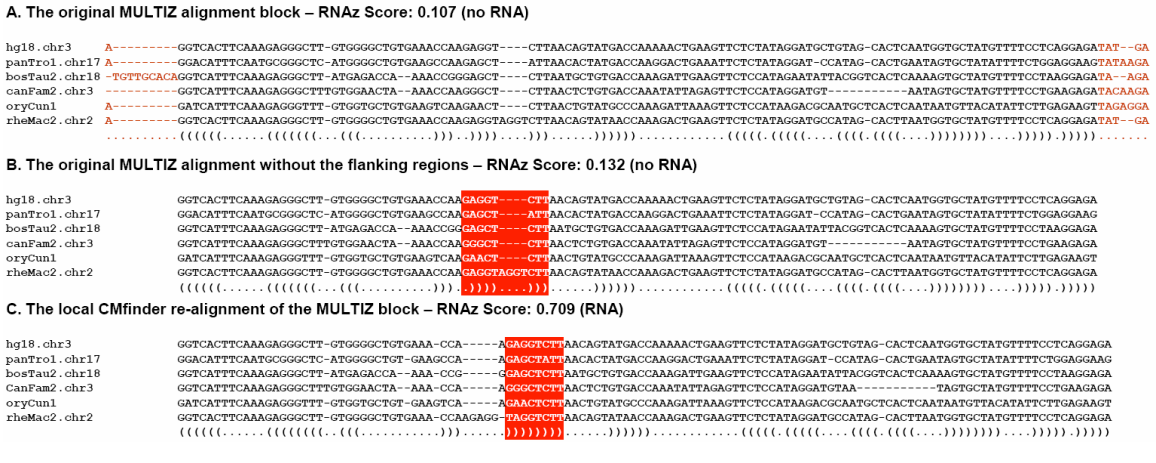


Figure 6.4: The results from running RNAz on three different alignments of RF00402 family members. (A) The original MULTIZ alignment block where we use RNAz scripts to select 6 sequences. (B) The flanking regions were removed to study the effect of that on the score. (C) The local CMfinder re-alignment of the whole MULTIZ block (0.13% is realigned). The most important changes between the alignments are highlighted in red.

It should be noted that RNAz and EvoFold remove individual sequences with more than 25% and 20% gaps, respectively, as compared to human. This is not necessary when using CMfinder because it takes structure into account when placing the gaps. CMfinder found motifs in 1408 and 673 individual sequences that would have been removed because of too many gaps by EvoFold and RNAz, respectively. In addition, RNAz is limited to 4-6 sequences, so they sample 6 sequences (repeated 3 times if there are more than 10 sequences in the alignment), optimizing the selected sequences to have sequence similarity as close to 80% as possible. EvoFold considers every sequence in the alignment, resulting in a lower score if any sequence is missing the motif. In contrast, although number of species is a factor in its composite score, CMfinder can ignore a sequence if it does not contain the motif and still report a high scoring motif for the rest of the sequences.

6.2.8 Experimental verification

An increasing number of ncRNAs are reported to be implicated in tissue-specific developmental and disease processes (24), yet the precise biological function of most ncRNAs remains elusive. To make a biological relevance of our prediction method probable, we initially generated a list of 11 high scoring ncRNA candidates. We chose to select high confidence predictions by setting a stricter score cutoff, composite score > 9 and energy $<$ -15, furthermore we chose a length cutoff of 60 and required more than five compensating base changes, indicating a possible evolutionary pressure to maintain the structure. We tested the expression of these 11 candidates in human RNA pools using strand specific primers (see Methods). We found that 8 out of 11 ncRNA candidates indeed could be detected in human RNA samples by reverse transcription PCR (RT-PCR) (ncRNA candidate #1, #2, #4, #7, #8, #9, #10 and #11; Figure 6.5A). Such expression may simply reflect random transcriptional noise, yet current literature suggests that mammalian ncRNAs exhibit highly tissue-specific expression profiles, which is likely to be indicative of specialized functions in the organism (116; 125). Hence, in order to expand our analysis and identify potential spatial and functional roles of our predicted set of ncRNAs, we performed an extensive expression analysis in 22 human tissues by RT-PCR totaling more than 250

separate reactions (see Methods). Our analysis demonstrated that 10 out of the 11 candidates are indeed expressed in one or more human tissues (Figure 6.5B). Interestingly, this analysis showed that 7 of 10 confirmed candidates exhibited a highly tissue-specific expression profile, whereas only two ncRNAs were more ubiquitously expressed (#10 and #11; Figure 6.5B). Hence, in agreement with the current consensus, we believe that the predicted ncRNAs may have highly defined biological roles. In addition, the highly differential expression patterns of the ncRNA candidates strongly suggest that the expression is real and not merely transcriptional noise, thus supporting the validity of our prediction method.

An interesting observation is that 9 out of 11 ncRNA candidates were detected in brain (Figure 6.5B). In fact, a similar enrichment of ncRNA expression in brain versus other tissues has previously been demonstrated in mouse (116) and involvement of ncRNAs in function and development of human central nervous system (CNS) have recently been identified (18; 41; 111; 136). Further, an RNAz screen of porcine EST sequences revealed that developmental brain tissue seems to contain more ncRNAs than nonbrain tissues (128). In order to examine the expression profile of our CNS-expressed candidates in more detail, we performed RT-PCR analysis on human RNA purified from total brain, fetal brain, cerebellum, hippocampus and spinal cord (Figure 6.5C). Candidate #11 was expressed in all the investigated nervous tissues (Figure 6.5C), as observed in the other tissues. Candidate #8, on the other hand, showed a more restricted expression profile, detected only in fetal brain, and at very low level in hippocampus of adult brain (Figure 6.5C). Hence, even within a single organ, the predicted ncRNA candidates appear to have highly specialized expression profiles, which is suggestive of a distinct biological function.

To expand our analysis, Northern blot analysis was performed for the most highly expressed ncRNA candidates on human RNA from 15 different tissues (Figure 6.5D). In general, detection of ncRNAs by Northern blotting has proven very difficult as the majority of ncRNAs are low abundance transcripts (125). We were able to detect bands for candidates #6 (Figure 6.5D), and its expression was confirmed to be strictly brain-specific. The 2.8-kb-long transcript is located within a 4-kb-long intron of synapsin III (SYN3) along with five more non-overlapping CMfinder-predicted motifs on the same strand. In Figure

6.5D, we have removed four tissues because of a high level of background noise, interfering with the results.

Next, we investigated the precise genomic locations of the ncRNAs; five candidates (#1,#2, #6, #9 and #10) are located within intronic sequences of known genes. Overall, we find a good correlation between our ncRNA expression analysis and database searches for the predicted host mRNA; for instance ncRNA #6 is located within an intron of Synapsin 3 (SYN3), which is neuron-specific and predominantly expressed in the brain (73). This expression profile is well confirmed by both our RT-PCR and Northern blot analysis showing a clear brain specific expression of #6. Furthermore, candidate #9 is located within an intron of the GRM8 (glutamate receptor metabotropic 8) precursor encoding a G-protein-coupled metabotropic glutamate receptor expressed in the central nervous systems (32). Again, our RT-PCR analysis confirms ncRNA #9 expression in most compartments of the brain and in spinal cord (Figure 6.5B and 6.5C). Finally, candidate #10 is located within the primary TIMP3 RNA transcript that encodes an inhibitor of matrix metalloproteinases (Genbank acc. NM_000362). TIMP3 mRNA is rather broadly expressed predominantly in brain, kidney and lung (83). which correlates well with the expression patterns of candidate #10 as evaluated by our RT-PCR analysis (Figure 6.5B). In conclusion, we find by both RT-PCR and Northern blot analysis that predicted ncRNA candidates are expressed in a highly tissue-specific manner which is likely indicative of specialized biological functions and thus supports the validity of our prediction method.

6.3 Methods

6.3.1 Data

The multiple alignments from the ENCODE regions were obtained from the UCSC genome browser, more specifically, the multiple alignments of 16 vertebrate genomes with the human genome (assembly hg18, Mar. 2006). We post-processed these alignments to remove all alignments blocks that overlapped with exons of known genes (`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz`) or the highly conserved elements defined by PhastCons (`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/`

database/phastConsElements17way.txt.gz) in human. Furthermore, we made an additional set with the reverse complementary sequences of each sequence in the alignment. GENCODE, TARs and Transfrags data were obtained from UCSC's table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and converted from assembly hg17 to hg18 using their liftOver software (http://genome.ucsc.edu/cgi-bin/hgLiftOver). sRNA and lRNA data were obtained at http://transcriptome.affymetrix.com/publication/hs_whole_genome. EvoFold and RNAz candidates were obtained at http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ENCODE .

6.3.2 False discovery rate

In order to estimate the false discovery rate (FDR), we shuffled all of our input alignments and ran CMfinder on them. The alignments were shuffled as described by Washietl and Hofacker (148) resulting in random alignments of the same base composition, sequence conservation and gap patterns. This method retains a coarse grained pattern of local conservation (columns are binned into two classes based on mean pairwise similarity using threshold 0.5, and only columns within the same class are shuffled with each other). This does not conserve the dinucleotide frequencies, which have an effect on the Gibbs free energies due to stacking interactions. Since the Gibbs free energy plays a role in our scoring of the candidates, this has an unknown effect on our estimated false discovery rate.

6.3.3 Running CMfinder

We ran CMfinder (version 0.2) separately on each alignment block in the MULTIZ alignment as well as the reverse complement of each such block. When running CMfinder we output up to 5 single stem predictions (size range 30-100bp) and 5 double stem predictions (size range 40-100bp). This corresponds to running CMfinder with the options "-n 5 -m 30 -M 100" and then with the options "-n 5 -s 2 -m 40 -M 100". Then we tried to combine the motifs using the greedy heuristics described in Section 2.2.2.

We ranked all CMfinder motifs using a heuristic scoring function adapted from the

bacteria study:

$$r = sp \cdot \sqrt{lc/sid} \cdot (bp/len)$$

See Chapter 4 for explanation of the involved variables, and interpretation of the formula. We changed the scoring function used in the bacteria study because it overly favors the long motifs. This is less of a problem in bacteria due to the general great sequence divergence among the bacterial intergenic regions, and the fact that homologous sequences around a structural motif is likely to be functionally relevant as well. However, in vertebrates, conserved intergenic regions that are not ncRNAs are prevalent. So we adjust our scoring function to favor compact RNA structures even more. This function is still referred to as the composite score. After systematically studying various cutoffs, we chose to focus on candidates with a composite score over 5 and free energy below -5, which resulted in a large number of candidates with a reasonable false positive rate. The energy is computed as the average energy of each sequence in the alignment as calculated by RNAfold (68), constrained to the secondary structure annotated by CMfinder.

6.3.4 pvalue calculation

To calculate the pvalues of the overlap between our predictions and other data sources, we counted the number of candidate regions whose center nucleotide overlaps the data we are testing against, i.e. RNAz candidates. We construct the following null model: each candidate is considered a dart thrown randomly onto the genome. If the target RNAz candidates cover a fraction p of the ENCODE nucleotides in multiple alignment blocks (our input data), then it is a simple binomial model: each of the N darts has probability p of hitting a TAR. We can approximate the pvalue using the normal approximation to the binomial distribution. The expected number of hits μ is np, with a standard deviation $\sigma = \sqrt{np(1-p)}$. The pvalue can then be calculated by using these numbers to look it up in a normal approximation table, for example, using the pnorm function in R $pnorm(observed, \mu, \sigma, lower.tail = F)$. We also calculated the pvalues using the leftmost and rightmost nucleotide, instead of the center nucleotide. This gives very similar results, although, when comparing to RNAz and EvoFold the pvalues were a bit worse, probably

because they are global and use window lengths, whereas CMfinder is local, therefore an overlap with our candidates' central nucleotide to RNAz and EvoFold candidates seems more likely.

6.3.5 Re-alignment calculation

To quantify how much has been realigned by CMfinder in a given motif compared to the original multiple alignment (see Figure 6.3), we calculate the following quantities. Let sp be the number of sequences in the CMfinder alignment, and m the number of matched positions in that alignment, i.e., the number of quadruples (s, t, i, j) with $1 \leq s < t \leq sp$ and such that position i of sequence s is aligned with position j of sequence t. Let v be the number of those matches that are re-aligned relative to the MULTIZ alignment, i.e., the number of quadruples as above for which position i of s is matched to position j of t in the CMfinder alignment, but not in the MULTIZ alignment (either, i and j are aligned to nucleotides in different positions or to gaps). The overall re-alignment fraction we report is $\frac{v}{m}$. For example if we have two multiple alignments, A and B, of four sequences which are all 10bp long, we will compare all six possible sequence pairs (all pair-combinations of the four sequences). Say we have 6 columns that are aligned differently in alignment A and B between sequences 1 and 3 and that the rest is aligned alike. Then we would say that $\frac{6}{6*10} = 10\%$ of alignment B is re-aligned compared to alignment A.

6.3.6 Experiments

The tissue specific expression profiles of 11 candidate ncRNAs were determined by RT-PCR using purified total RNA from 22 different human tissues (adrenal gland, bone marrow, brain (whole, fetal, cerebellum, and hippocampus), kidney, liver (fetal), lung, prostate, salivary gland, skeletal muscle, spleen, testis, thymus, thyroid gland, trachea, uterus, colon and small intestine). cDNAs were generated by reverse transcription (RT) using M-MLV SuperScript III Reverse Transcriptase (Invitrogen, Carlsbad CA). For Northern blot analysis of ncRNA expression, Nylon membranes with pre-blotted human RNA samples (15μg/tissue) (Zyagen, San Diego CA) were hybridized at 37C in Ultrahyb hybridization buffer (Ambion, Austin

TX) with 80 nt. end-labeled probes antisense to the predicted ncRNAs. See (142) for other experimental details of RT-PCR and Northern blot.

6.4 Discussion

Non-coding RNAs are receiving increasing attention in genome science. This study conducted the first large-scale search for structural ncRNAs in several vertebrate genomes using a local structural motif finding algorithm, and identified several thousand novel candidate ncRNAs. Our work complements a previous pairwise scan for local structural RNA elements in corresponding unaligned regions of the human and mouse genomes (143) by extending it to multiple genomes and including a wider range of sequence similarities. Furthermore, except to indicate orthology, the scan was not dependent on sequence-based pre-aligned genomic regions, contrary to the case with RNAz and EvoFold scans (150), allowing us to increase the number of ncRNAs candidates in the ENCODE regions by 32%. With a growing number of sequenced genomes, and with improving genome alignment methods that are capable of capturing orthology among phylogenetically diverse species, analysis of syntenic yet diverse regions becomes more feasible (92). Alignments of increasingly diverse regions often mean decreasing average pairwise sequence similarity. This is problematic for sequence-based alignment methods. When searching for structural ncRNAs, one can therefore benefit from disregarding the alignments, and re-aligning the regions considering sequence and structure. Indeed it has been shown, for pairwise alignments of tRNAs, that it is preferable also to consider structure when aligning these if sequence similarity is below 60% (45).

Although useful, we don't feel that any of the methods we used to date constitute the last word on this topic, with several remaining challenges. The first issue is on data collection. In this study, we limited the motif search within an alignment block, which compromises the performance if the true ncRNAs extend across alignment blocks. This is quite likely if the sequence conservation drops to a certain level that causes major mis-alignment of sequence segments. We would like to expand the motif search to wider regions, probably anchored by highly conserved regions. The enhanced CMfinder 0.3 can handle large datasets, and seems appropriate for this application. The main obstacle in doing this currently is to determine

the optimal strategy for identifying the anchors, and for merging/dividing alignment blocks.

Secondly, we expect many functionally important ncRNA motifs to be repeated in the genome, e.g., cis-regulatory elements controlling several genes in a common pathway, or multiple members of as-yet unknown RNA families. There has been limited work to date attempting to identify or cluster repeated motifs predicted by genome-scale RNA discovery approaches (143; 160; 121; 160). It is unclear how these methods scale to large genomes. The CMfinder-based approach we have described in this thesis potentially provides an alternative to the clustering approaches. Since each of our RNA motifs is described by a covariance model, in principle, we could use each to scan the genome for additional instances. In this study, we have scanned the selected candidates on the input sequences. However, since ENCODE only comprise 1% of the genomes, and we removed the highly conserved regions, we did not have much success in discovering additional copies of the motifs, except those falling in the repeat regions. However, if we extend the search to full genomes, we will have a much better chance at finding multiple occurrences of the motifs. CM scans on the full genomes are very expensive, even with the HMM filtering technique. An alternative is to perform CM scans only on candidate regions. If two motifs overlap, the CM of one motif tends to pick up a good hit in the other one. If we limit the scan within the candidate sequences, we can afford to perform the CM scans for a much larger set of candidates.

Thirdly, the analysis in this study is based on a heuristic scoring function. In Chapter 4, we have described a new scoring function with better statistical properties, but we have yet to investigate how well this method behaves in practice, including experimental validation of the top candidates, overlap with known transcripts, etc.

Our next immediate step is to extend the search to the full human genome. The preliminary phase, including CMfinder search and scoring candidates (with the new method pscore), has already finished. This phase has taken over 100 CPU years, producing over one million raw candidates (without filtering or clustering). We will use the false discovery rate estimate from Chapter 4 to guide the selection among these candidates, which we hope will help us to skim the most valuable predictions. We plan to repeat the analysis in this study. The mere size of our prediction set poses big challenges for even a simplistic analysis, and

scalability issue is our top concern. Therefore, a precise run time estimate is a prerequisite for our choice of any method. For example, the CM scan on the full human genome is very slow despite its merits, and we need to determine for how many top candidates we can afford to scan.

Finally, ncRNA validation remains a challenging task. The popular methods, such as tiling arrays, RT-PCR and northern blot, all have only limited success. With the prevalence of new sequencing technology, it has been becoming more feasible to simply sequence all ncRNA transcripts, potentially leading to big breakthroughs in the field. There is also a big need for high-throughput methods, computational and experimental, to identify a potential function for the tens of thousands of candidates that have resulted from scans like this.

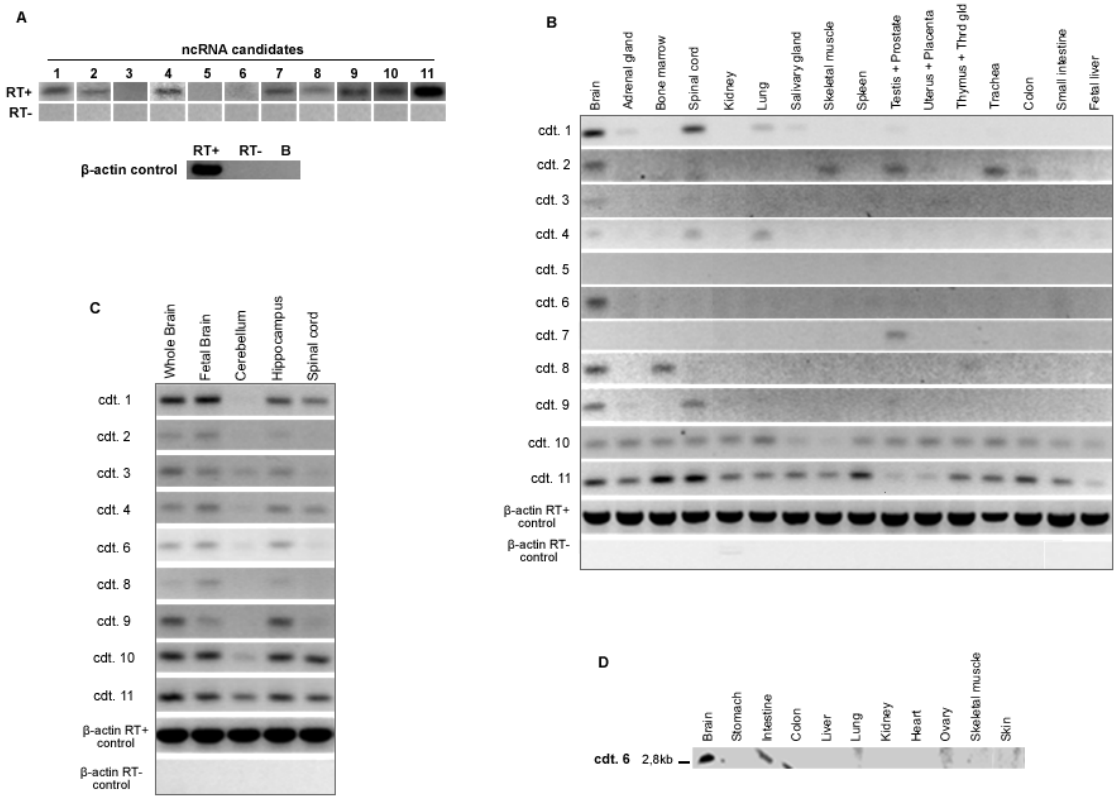


Figure 6.5: Expression of predicted ncRNA candidates by RT-PCR and Northern blot analysis. (A) Strand-specific RT-PCR analysis of ncRNA candidates on human RNA pools (see Methods). β-actin was used as control yielding PCR products in the presence of reverse transcriptase (RT+), but not in its absence (RT-). (B) Tissue-specific expression of ncRNA candidates as evaluated by RT-PCR analysis of human RNA samples. The same β-actin controls as for A were used. (C) Expression of ncRNA candidates within the human CNS as evaluated by RT-PCR analysis. The same β-actin controls as for A and B were used. (D) Expression of ncRNA candidates #6 as evaluated by Northen blotting of human RNA samples from 15 tissues.

BIBLIOGRAPHY

[1] T. Allen, P. Shen, L. Samsel, R. Liu, L. Lindahl, and J. M. Zengel. Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. *Journal of Bacteriology*, 181:6124–6132, 1999.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. In *Nucleic Acids Research* (3), pages 3389–3402.

[3] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[4] Ilka M Axmann, Philip Kensche, Jorg Vogel, Stefan Kohl, Hanspeter Herzel, and Wolfgang R Hess. Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biology*, 6(9):R73, 2005.

[5] J.P. Bachellerie, J. Cavaill, and A. Httenhofer. The expanding snoRNA world. *Biochimie*, 84:775–790, Aug 2002.

[6] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. In *Proceedings of the Third Internation Conference on Intelligent Systems for Molecular Biology*. AAAI, 1995.

[7] Timothy L. Bailey and Charles Elkan. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, CA, 1994. AAAI Press.

[8] Jeffrey E. Barrick, K. A. Corbino, W. C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J. K. Wickiser, and Ronald R. Breaker. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proceedings of the National Academy of Sciences of the Unites States of America*, 101(17):6421–6426, 2004.

[9] Jeffrey E Barrick, Narasimhan Sudarsan, Zasha Weinberg, Walter L Ruzzo, and Ronald R Breaker. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, 11(5):774–784, May 2005.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[11] P. Bertone, V. Stolc, T.E. Royce, J.S. Rozowsky, A.E. Urban, X. Zhu, J.L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306:2242–2246, Dec 2004.

[12] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14:708–715, Apr 2004.

[13] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14:708–715, Apr 2004.

[14] M. Blanchette and M. Tompa. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Research*, 31(13):3840–2, 2003.

[15] G.G. Brownlee. Sequence of 6S RNA of *E. coli*. *Nat New Biol*, 229(5):147–149, February 1971.

[16] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(2), 2002.

[17] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V.B. Bajic, S.E. Brenner, S. Batalov, A.R. Forrest, M. Zavolan, M.J. Davis, L.G. Wilming, V. Aidinis, J.E. Allen, A. Ambesi-Impiombato, R. Apweiler, R.N. Aturaliya, T.L. Bailey, M. Bansal, L. Baxter, K.W. Beisel, T. Bersano, H. Bono, A.M. Chalk, K.P. Chiu, V. Choudhary, A. Christoffels, D.R. Clutterbuck, M.L. Crowe, E. Dalla, B.P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C.F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T.R. Gingeras, T. Gojobori, R.E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T.K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S.P. Krishnan, A. Kruger, S.K. Kummerfeld, I.V. Kurochkin, L.F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K.C. Pang, W.J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J.F. Reid, B.Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S.L. Salzberg, A. Sandelin, C. Schneider, C. Schnbach, K. Sekiguchi, C.A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S.L. Tan, S. Tang, M.S. Taylor, J. Tegner, S.A. Teichmann, H.R. Ueda, E. van Nimwegen, R. Verardo, C.L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S.M. Grimmond, R.D. Teasdale, E.T. Liu, V. Brusic, J. Quack-

enbush, C. Wahlestedt, J.S. Mattick, D.A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, and Y. Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, Sep 2005.

[18] J. Cavaill, P. Vitali, E. Basyuk, A. Httenhofer, and J.P. Bachellerie. A novel brain-specific box C/D small nucleolar RNA processed from tandemly repeated introns of a noncoding RNA gene in rats. *J. Biol. Chem.*, 276:26374–26383, Jul 2001.

[19] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D.K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D.S. Gerhard, and T.R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308:1149–1154, May 2005.

[20] P. Clote, F. Ferr, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11:578–591, May 2005.

[21] B. Conne, A. Stutz, and J.D. Vassalli. The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat Med*, 6(6):637–41, 2000.

[22] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007.

[23] K.A. Corbino, J.E. Barrick, J. Lim, R. Welz, B.J. Tucker, I. Puskarz, M. Mandal, N.D. Rudnick, and R.R. Breaker. Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biology*, 6(8), 2005.

[24] F.F. Costa. Non-coding RNAs: new players in eukaryotic biology. *Gene*, 357:83–94, Sep 2005.

[25] Alex Coventry, Daniel J. Kleitman, and Bonnie Berger. MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the Unites States of America*, 101(33):12102–7, 2004.

[26] I. Derre, G. Rapoport, and T. Msadek. CtsR, a novel regulator of stress and heat shock response, controls *clp* and molecular chaperone gene expression in gram-positive bacteria. *Molecular Microbiology*, 31(1):117–131, January 1999.

[27] R.D. Dowell and S.R. Eddy. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7(400), 2006.

[28] Robin D. Dowell and Sean R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71, 2004.

[29] M. Dsouza, N. Larsen, and R. Overbeek. Searching for patterns in genomic data. *Trends in Genetics*, 13(12):497–498, 1997.

[30] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, 1998.

[31] L Duret, C Chureau, S Samain, J Weissenbach, and P Avner. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312 (5780):1653–1655, 2006.

[32] R.M. Duvoisin, C. Zhang, and K. Ramonell. A novel metabotropic glutamate receptor expressed in the retina and olfactory bulb. *J. Neurosci.*, 15:3075–3083, Apr 1995.

[33] Sean R. Eddy. Infernal 0.55. `http://www.genetics.wustl.edu/eddy/infernal/`, 2003.

[34] Sean R. Eddy. *Infernal User's Guide*, 2003–06. `ftp://selab.janelia.org/pub/software/infernal/Userguide.pdf`.

[35] Sean R. Eddy and Richard Durbin. RNA Sequence Analysis Using Covariance Models. In *Nucleic Acids Research* (36), pages 2079–88.

[36] Sean R. Eddy and Richard Durbin. RNA Sequence Analysis Using Covariance Models. *Nucleic Acids Research*, 22(11):2079–88, 1994.

[37] G. Endo and S. Silver. CadC, the transcriptional regulatory protein of the cadmium resistance system of *Staphylococcus aureus* plasmid pI258. *Journal of Bacteriology*, 177(15):4437–41, August 1995.

[38] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[39] Jianchi Feng, Chunming Bi, Brian S. Clark, Rina Mady, Palak Shah, , and Jhumku D. Kohtz1. The evf-2 noncoding RNA is transcribed from the dlx-5/6 ultraconserved region and functions as a dlx-2 transcriptional coactivator. *Genes Dev.*, 20(11):1470–1484, 2006.

[40] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, and D.H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the Unites States of America*, 83(24):9373–9377, 1986.

[41] P.J. French, T.V. Bliss, and V. O'Connor. Ntab, a novel non-coding RNA abundantly expressed in rat brain. *Neuroscience*, 108:207–215, 2001.

[42] J.R. Fresco, B.M. Alberts, and P. Doty. Some molecular details of the secondary structure of ribonucleic acid. *Nature*, 188:98–101, Oct 1960.

[43] H. Fu, Y. Tie, C. Xu, Z. Zhang, J. Zhu, Y. Shi, H. Jiang, Z. Sun, and X. Zheng. Identification of human fetal liver miRNAs by a novel method. *FEBS Lett.*, 579:3849–3854, Jul 2005.

[44] R.T. Fuchs, F.J. Grundy, and T.M. Henkin. The S(MK) box is a new SAM-binding RNA for translational regulation of SAM synthetase. *Nat. Struct. Mol. Biol.*, 13:226–233, Mar 2006.

[45] P.P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, 33:2433–2439, 2005.

[46] Daniel Gautheret and André Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of Molecular Biology*, 313:1003–1011, 2001.

[47] Mark Gerstein, Erik L. L. Sonnhammer, and Cyrus Chothia. Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4):1067–78, 1994.

[48] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, and A. Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15:1451–1455, Oct 2005.

[49] Walter Gilbert. Origin of life: The rna world. *Nature*, 319, 1986.

[50] Vadim N. Gladyshev, Gregory V. Kryukov, Dmitri E. Fomenko, and Dolph L. Hatfield. Identification of trace element-containing proteins in genomic databases. *Annu Rev Nutr*, 24:579–596, 2004.

[51] J. Gorodkin, L. J. Heyer, and G.D. Stormo. Finding the most significant common sequences and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 25(18):3724–32, 1997.

[52] J. Gorodkin, S. L. Stricklin, and G. D. Stormo. Discovering common stem-loop motifs in unaligned RNA sequence. *Nucleic Acids Research*, 29(10):2135–44, 2001.

[53] P. Green, B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.*, 33:514–517, Apr 2003.

[54] S. Griffiths-Jones. miRBase: the microRNA sequence database. *Methods Mol. Biol.*, 342:129–138, 2006.

[55] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34:D140–144, Jan 2006.

[56] S. Griffiths-Jones, H.K. Saini, S. van Dongen, and A.J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154–158, Jan 2008.

[57] Sam Griffiths-Jones. RALEE - RNA alignment editor in emacs. *Bioinformatics*, 21(2):257–259, 2005.

[58] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441, 2003.

[59] Giorgio Grillo, Flavio Licciulli, Sabino Liuni, Elisabetta Sbisà, and Graziano Pesole. PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Research*, 31(13):3608–3612, 2003.

[60] F. J. Grundy and T. M. Henkin. Characterization of the *Bacillus subtilis rpsD* regulatory target site. *Journal of Bacteriology*, 174(21):6763–6770, November 1992.

[61] Silviu Guiasu. *Information Theory with Applications*. McGraw-Hill, 1977.

[62] M. Hamada, K. Tsuda, T. Kudo, T. Kin, and K. Asai. Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, 22:2480–2487, Oct 2006.

[63] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigo. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, 7 Suppl 1:1–9, 2006.

[64] J.H. Havgaard, R. Lyngsø, G.D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–1824, 2005.

[65] M.W. Hentze and L.C. Kuhn. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proceedings of the National Academy of Sciences of the Unites States of America*, 93:8175–82, 1996.

[66] J. Hertel and P.F. Stadler. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22:197–202, Jul 2006.

[67] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary Structure Prediction for Aligned RNA sequences. *Journal of Molecular Biology*, 319(5):1059–66, 2002.

[68] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structure. *Chemical Monthly*, 125:167–88, 1994.

[69] Ian Holmes. A probabilistic model for the evolution of rna structure. *BMC Bioinformatics*, 5(166), 2004.

[70] Ian Holmes. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, 6(73), 2005.

[71] Y. Ji, X. Xu, and G. D. Stormo. A graph theoretical approach to predict common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20(10), 2004.

[72] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:354–357, 2006.

[73] H.T. Kao, B. Porton, A.J. Czernik, J. Feng, G. Yiu, M. Hring, F. Benfenati, and P. Greengard. A third member of the synapsin gene family. *Proc. Natl. Acad. Sci. U.S.A.*, 95:4667–4672, Apr 1998.

[74] P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermller, I.L. Hofacker, I. Bell, E. Cheung, J. Drenkow,

E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T.R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, Jun 2007.

[75] K. Katoh, K. Kuma, T. Miyata, and H. Toh. Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform*, 16:22–33, 2005.

[76] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30:3059–3066, Jul 2002.

[77] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, Y. Ono, A. Kojima, Y. Kimura, T. Komori, and K. Asai. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, 35:D145–148, Jan 2007.

[78] Robbie J. Klein and Sean R. Eddy. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(1):44, 2003.

[79] Peter S Klosterman, Andrew V Uzilov, Yuri R Bendana, Robert K Bradley, Sharon Chao, Carolin Kosiol, Nick Goldman, and Ian Holmes. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, 7:428, 2006.

[80] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[81] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, July 2003.

[82] S.Y Le, J.H Chen, and J.V. Maizel. Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic Acids Research*, 17:6143–6152, 1989.

[83] K.J. Leco, R. Khokha, N. Pavloff, S.P. Hawkes, and D.R. Edwards. Tissue inhibitor of metalloproteinases-3 (TIMP-3) is an extracellular matrix-associated protein with a

distinctive pattern of expression in mouse cells and tissues. *J. Biol. Chem.*, 269:9352–9360, Mar 1994.

[84] RC Lee, RL Feinbaum, and V Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 2003.

[85] C. Liu, B. Bai, G. Skogerb, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao, and R. Chen. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, 33:D112–115, Jan 2005.

[86] P.E. Lobert, N. Escriou, J. Ruelle, and T. Michiels. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11560–11565, Sep 1999.

[87] Y. Lu, R.J. Turner, and R.L. Switzer. Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. *Proceedings of the National Academy of Sciences of the Unites States of America*, 93(25):14462–14467, December 1996.

[88] R.B. Lyngs, M. Zuker, and C.N. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15:440–445, Jun 1999.

[89] M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, and R.R. Breaker. Riboswitches Control Fundamental Biochemical Pathways in *Bacillus subtilis* and Other Bacteria. *Cell*, 113(5):577–86, 2003.

[90] Maumita Mandal, Mark Lee, Jeffrey E. Barrick, Zasha Weinberg, Gail Mitchell Emilsson, Walter L. Ruzzo, and Ronald R. Breaker. A Glycine-dependent Riboswitch that Uses Cooperative Binding to Control Gene Expression in Bacteria. *Science*, 306:275–279, 8 October 2004.

[91] Aron Marchler-Bauer, John B Anderson, Praveen F Cherukuri, Carol DeWeese-Scott, Lewis Y Geer, Marc Gwadz, Siqian He, David I Hurwitz, John D Jackson, Zhaoxi

Ke, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Benjamin A Shoemaker, Vahan Simonyan, James S Song, Paul A Thiessen, Roxanne A Yamashita, Jodie J Yin, Dachuan Zhang, and Stephen H Bryant. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Research*, 33(Database issue):D192–D196, January 2005.

[92] E.H. Margulies, C.W. Chen, and E.D. Green. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.*, 22:187–193, Apr 2006.

[93] E.H. Margulies, G.M. Cooper, G. Asimenos, D.J. Thomas, C.N. Dewey, A. Siepel, E. Birney, D. Keefe, A.S. Schwartz, M. Hou, J. Taylor, S. Nikolaev, J.I. Montoya-Burgos, A. Lytynoja, S. Whelan, F. Pardi, T. Massingham, J.B. Brown, P. Bickel, I. Holmes, J.C. Mullikin, A. Ureta-Vidal, B. Paten, E.A. Stone, K.R. Rosenbloom, W.J. Kent, G.G. Bouffard, X. Guan, N.F. Hansen, J.R. Idol, V.V. Maduro, B. Maskeri, J.C. McDowell, M. Park, P.J. Thomas, A.C. Young, R.W. Blakesley, D.M. Muzny, E. Sodergren, D.A. Wheeler, K.C. Worley, H. Jiang, G.M. Weinstock, R.A. Gibbs, T. Graves, R. Fulton, E.R. Mardis, R.K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D.B. Jaffe, J.L. Chang, K. Lindblad-Toh, E.S. Lander, A. Hinrichs, H. Trumbower, H. Clawson, A. Zweig, R.M. Kuhn, G. Barber, R. Harte, D. Karolchik, M.A. Field, R.A. Moore, C.A. Matthewson, J.E. Schein, M.A. Marra, S.E. Antonarakis, S. Batzoglou, N. Goldman, R. Hardison, D. Haussler, W. Miller, L. Pachter, E.D. Green, and A. Sidow. Analyses of deep mammalian sequence alignments and constraint predictions for 1 *Genome Res.*, 17:760–774, Jun 2007.

[94] D. H. Mathews and D. H. Turner. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2):191–203, 2002.

[95] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. Incorporating chemical modification constraints into a dynamic programming algo-

rithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the Unites States of America*, 101:7287–7292, 2004.

[96] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[97] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–19, 1990.

[98] John P. McCutcheon and Sean R. Eddy. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Research*, 31(14):4119–4128, 2003.

[99] A.S. Mironov, I. Gusarov, R. Rafikov, L.E. Lopez, K. Shatalin, R.A. Kreneva, D.A. Perumov, and E. Nudler. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111 (5):747–756, 2002.

[100] B. Morgenstern, A. Dress, and T. Werner. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the Unites States of America*, 93:12098–103, 1996.

[101] A. Nahvi, N. Sudarsan, M.S. Ebert, X. Zou, K.L. Brown, and R.R. Breaker. Genetic control by a metabolite binding mRNA. *Chem Biol*, 9(9):1043, 2002.

[102] Franz Narberhaus. Negative regulation of bacterial heat shock genes. *Molecular Microbiology*, 31(1):1–8, January 1999.

[103] E.P. Nawrocki and S.R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, 3:e56, Mar 2007.

[104] F. C. Neidhardt, J. L. Ingraham, and R. C. Curtiss, editors. *Regulation of ribosome synthesis in Escherichia coli and Salmonella: Cellular and Molecular Biology.* ASM Press, 1996.

[105] S. Neph and M. Tompa. MicroFootPrinter: a Tool for Phylogenetic Footprinting in Prokaryotic Genomes. *Nucleic Acids Research*, 34, 2006.

[106] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences of the Unites States of America*, 77:6309–6313, 1980.

[107] R. Nussinov, G. Piecznik, J.R. Grigg, and D.J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 1978.

[108] K.C. Pang, S. Stephen, M.E. Dinger, P.G. Engstrm, B. Lenhard, and J.S. Mattick. RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, 35:D178–182, Jan 2007.

[109] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S. Lander, Jim Kent, Webb Miller, and David Haussler. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Computational Biology*, 2006.

[110] J.S. Pedersen, I.M. Meyer, R. Forsberg, P. Simmonds, and J. Hein. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Research*, 32(16):4925–4923, 2004.

[111] K.S. Pollard, S.R. Salama, N. Lambert, M.A. Lambot, S. Coppens, J.S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, A.D. Kern, C. Dehay, H. Igel, M. Ares, P. Vanderhaeghen, and D. Haussler. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443:167–172, Sep 2006.

[112] J. Ponjavic, C.P. Ponting, and G. Lunter. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, 17:556–565, May 2007.

[113] T. Powers, L. M. Changchien, G. R. Craven, and H. F. Noller. Probing the assembly of the 3' major domain of 16S ribosomal RNA. Quaternary interactions involving

ribosomal proteins S7, S9 and S19. *Journal of Molecular Biology*, 200(2):309–319, 1988.

[114] K.D. Pruitt, T. Tatusova, and D.R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(1):501–504, 2005.

[115] E Puerta-Fernandez, JE Barrick, A Roth, and RR Breaker. Identification of a large noncoding RNA in extremophilic eubacteria. *Proceedings of the National Academy of Sciences of the Unites States of America*, 103(51):19490–19495, 2006.

[116] T. Ravasi, H. Suzuki, K.C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M.C. Frith, M.M. Gongora, S.M. Grimmond, D.A. Hume, Y. Hayashizaki, and J.S. Mattick. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, 16:11–19, Jan 2006.

[117] Elena Rivas. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics*, 6(63), 2005.

[118] Elena Rivas and Sean R. Eddy. A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots. *Journal of Molecular Biology*, 285(5):2053–2068, February 1999.

[119] Elena Rivas and Sean R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

[120] Elena Rivas and Sean R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

[121] D. Rose, J. Hackermller, S. Washietl, K. Reiche, J. Hertel, S. Findei, P.F. Stadler, and S.J. Prohaska. Computational RNomics of Drosophilids. *BMC Genomics*, 8:406, 2007.

[122] A. Roth, W.C. Winkler, E.E. Regulski, B.W. Lee, J. Lim, I. Jona, J.E. Barrick, A. Ritwik, J.N. Kim, R. Welz, D. Iwata-Reuyl, and R.R. Breaker. A riboswitch selec-

tive for the queuosine precursor preQ1 contains an unusually small aptamer domain. *Nat. Struct. Mol. Biol.*, 14:308–317, Apr 2007.

[123] G. Ruvkun. Molecular biology. Glimpses of a tiny RNA world. *Science*, 294 (5543):797–799, 2001.

[124] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45:810–25, 1985.

[125] Y.T. Sasaki, M. Sano, T. Ideue, T. Kin, K. Asai, and T. Hirose. Identification and characterization of human non-coding RNAs with tissue-specific expression. *Biochem. Biophys. Res. Commun.*, 357:991–996, Jun 2007.

[126] B. S. Schuwirth, M. A. Borovinskaya, C. W. Hau, W. Zhang, A. Vila-Sanjurjo, J. M. Holton, and J. H. Cate. Structures of the bacterial ribosome at 3.5 A resolution. *Science*, 310:827–834, 2005.

[127] L.G. Scott and J.R. Williamson. Interaction of the *Bacillus stearothermophilus* ribosomal protein S15 with its 5'-translational operator mRNA. *Journal of Molecular Biology*, 314:413–422, 2001.

[128] S.E. Seemann, M.J. Gilchrist, I.L. Hofacker, P.F. Stadler, and J. Gorodkin. Detection of RNA structures in porcine EST data and related mammals. *BMC Genomics*, 8:316, 2007.

[129] A.V. Seliverstov, H. Putzer, M.S. Gelfand, and V.A. Lyubetsky. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology*, 5(54), 2005.

[130] M.J. Serra and D.H. Turner. Predicting thermodynamic properties of RNA. *Methods Enzymol*, 259:242–261, 1995.

[131] B. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Computer Applications in the Biosciences*, 6:309–18, 1988.

[132] V.K. Sharma, C.J. Hackbarth, T.M. Dickinson, and G.L. Archer. Interaction of native and mutant MecI repressors with sequences that regulate *mecA*, the gene encoding penicillin binding protein 2a in methicillin-resistant staphylococci. *Journal of Bacteriology*, 180(8):2160–2166, April 1998.

[133] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, Aug 2005.

[134] Michael Sipser. *Introduction to the Theory of Computation*. PWS Publishing, 1997.

[135] F. Sleutels, R. Zwart, and D.P. Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415:810–813, 2001.

[136] M. Sone, T. Hayashi, H. Tarui, K. Agata, M. Takeichi, and S. Nakagawa. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J. Cell. Sci.*, 120:2498–2506, Aug 2007.

[137] M. Springer and C. Portier. More than one way to skin a cat: Translational autoregulation by ribosomal protein S15. *Nature Structural Biology*, 10:420–422, 2003.

[138] A. Stark, M.F. Lin, P. Kheradpour, J.S. Pedersen, L. Parts, J.W. Carlson, M.A. Crosby, M.D. Rasmussen, S. Roy, A.N. Deoras, J.G. Ruby, J. Brennecke, E. Hodges, A.S. Hinrichs, A. Caspi, B. Paten, S.W. Park, M.V. Han, M.L. Maeder, B.J. Polansky, B.E. Robson, S. Aerts, J. van Helden, B. Hassan, D.G. Gilbert, D.A. Eastman, M. Rice, M. Weir, M.W. Hahn, Y. Park, C.N. Dewey, L. Pachter, W.J. Kent, D. Haussler, E.C. Lai, D.P. Bartel, G.J. Hannon, T.C. Kaufman, M.B. Eisen, A.G. Clark, D. Smith, S.E. Celniker, W.M. Gelbart, and M. Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450:219–232, Nov 2007.

[139] 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 Research*, 22(22):4673–4680, November 1994.

[140] J.L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

[141] I. Tinoco, O.C. Uhlenbeck, and M.D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230:362–367, Apr 1971.

[142] E. Torarinsson, Z. Yao, E.D. Wiklund, J.B. Bramsen, C. Hansen, J. Kjems, N. Tommerup, W.L. Ruzzo, and J. Gorodkin. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.*, 18:242–251, Feb 2008.

[143] Elfar Torarinsson, Milena Sawera, Jakob H Havgaard, Merete Fredholm, and Jan Gorodkin. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Research*, 16(7):885–889, July 2006. Erratum Genome Research 16:1439, 2006.

[144] H. Touzet and O. Perriquet. CARNAC: folding families of noncoding RNAs. *Nucleic Acids Research*, 142, 2004.

[145] A.E. Trotochaud and K.M. Wassarman. A highly conserved 6S RNA structure is required for regulation of transcription. *Nat Struct Mol Biol*, 12(4):313–319, April 2005.

[146] X.F. Wan and D. Xu. Intrinsic terminator prediction and its application in *synechococcus sp. wh8102. International Journal of Computer Science and Technology*, 20:465–482, 2005.

[147] Adrienne X. Wang, Walter L. Ruzzo, and Martin Tompa. How Accurately Is ncRNA Aligned within Whole-Genome Multiple Alignments? *BMC Bioinformatics*, 8, 2007.

[148] S. Washietl and I.L. Hofacker. Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics. *Journal of Molecular Biology*, 342:19–30, 2004.

[149] S. Washietl, I.L. Hofacker, and P.F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the Unites States of America*, 102:2454–2459, 2005.

[150] S. Washietl, J.S. Pedersen, J.O. Korbel, A.R. Gruber, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Reiche, C. Stocsits, and A. Tanzer. Structured RNAs in the ENCODE Selected Regions of the Human Genome. *Genome Research*, 17:852–864, 2007.

[151] Stefan Washietl, Ivo L Hofacker, Melanie Lukasser, Alexander Huttenhofer, and Peter F Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnology*, 23(11):1383–1390, November 2005.

[152] Z. Weinberg, J.E. Barrick, Z. Yao, A. Roth, J.N. Kim, J. Gore, J.X. Wang, E.R. Lee, K.F. Block, N. Sudarsan, S. Neph, M. Tompa, W.L. Ruzzo, and R.R. Breaker. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.*, 35:4809–4819, 2007.

[153] Zasha Weinberg and Walter L. Ruzzo. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. In *Bioinformatics* (154), pages i334–i341. ISMB 2004.

[154] Zasha Weinberg and Walter L. Ruzzo. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, 20(suppl_1):i334–i341, 2004. ISMB 2004.

[155] Zasha Weinberg and Walter L. Ruzzo. Faster Genome Annotation of Non-coding RNA Families Without Loss of Accuracy. In *RECOMB04: Proceedings of the Eighth*

Annual International Conference on Computational Molecular Biology (156), pages 243–251.

[156] Zasha Weinberg and Walter L. Ruzzo. Faster Genome Annotation of Non-coding RNA Families Without Loss of Accuracy. In *RECOMB04: Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, pages 243–251, San Diego, CA, March 2004. ACM.

[157] Zasha Weinberg and Walter L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. In *Bioinformatics* (158), pages 35–39.

[158] Zasha Weinberg and Walter L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, 22(1):35–39, January 2006.

[159] P. M. Wikstrom, L. K. Lind, D. E. Berg, and G. R. Björk. Importance of mRNA folding and start codon accessibility in the expression of genes in a ribosomal protein operon of *Escherichia coli*. *Journal of Molecular Biology*, 224:949–966, 1992.

[160] S. Will, K. Reiche, I.L. Hofacker, P.F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3:e65, Apr 2007.

[161] Dagmar K Willkomm, Jens Minnerup, Alexander Huttenhofer, and Roland K Hartmann. Experimental RNomics in Aquifex aeolicus: identification of small non-coding RNAs and the putative 6S RNA homolog. *Nucleic Acids Res*, 33(6):1949–1960, 2005.

[162] Wade C. Winkler and Ronald R. Breaker. Genetic control by metabolite-binding riboswitches. *Chembiochem.*, 4(10):1024–1032, 2003.

[163] W.C. Winkler and R.R. Breaker. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.*, 59:487–517, 2005.

[164] W.C. Winkler, A. Nahvi, N. Sudarsan, J.E. Barrick, and R.R. Breaker. An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat. Struct. Biol.*, 10:701–707, Sep 2003.

[165] Z. Yao, J. Barrick, Z. Weinberg, S. Neph, R. Breaker, M. Tompa, and W.L. Ruzzo. A Computational Pipeline for High- Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. *PLoS Comput Biol*, 3:e126, Jul 2007.

[166] Zizhen Yao, Zasha Weinberg, and Walter L. Ruzzo. CMfinder supplementary website. `www.bio.washington.edu/yzizhen/CMfinder/`.

[167] Zizhen Yao, Zasha Weinberg, and Walter L Ruzzo. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, February 2006.

[168] J. M. Zengel and L. Lindahl. Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*. *Prog Nucleic Acid Res Mol Biol*, 47:331–370, 1994.

[169] C. Zhu, H. Byrd, and J. Nocedal. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

[170] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

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

[172] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using themodynamic and auxilary information. *Nucleic Acids Research*, 9:133–48, 1981.

VITA

Zizhen Yao was born in Wuhan, China. She earned a Bachelor of Science degree in Computer Science from Wuhan University. She then came to USA for graduate study, and earned a Master of Science degree in Computer Science from University of California, Riverside. In 2008, she earned a Doctor of Philosophy degree, still in Computer Science, from University of Washington. Currently, she settles in Seattle.