uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

u Ottawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Mohammad Anwar

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.C.S. (Master of Computer Science)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Implementation and Evaluation of Scoring Schemes for the Automated Discovery of Nucleic Acid Structures

TITRE DE LA THÈSE / TITLE OF THESIS

Marcel Turcotte

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Nathalie Japkowicz

Evangelos Kranakis

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Implementation and evaluation of scoring schemes for the automated discovery of nucleic acid structures

Mohammad Anwar

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
for the M.Sc. degree in Computer Science

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-18393-9
Our file Notre référence
ISBN: 978-0-494-18393-9

Abstract

With recent experimental evidence, it has been shown that RNA (ribonucleic acid) plays a greater role in various cellular functions than previously thought. With the increasing number of known RNA families a need arises to develop computational techniques to analyze RNA sequences. An array of evolutionary related RNA sequences believed to contain signals at both the sequence and structure levels can be exploited to detect motifs common to all or a portion of those sequences. Finding these similar structural features can provide substantial information as to which parts of the sequence are functional.

Recently, Nguyen (M.A.Sc thesis, Electrical Engineering, University of Ottawa, 2004) introduced a novel approach for discovering consensus secondary structure motifs in a set of unaligned RNA sequences. The algorithm has been implemented in a software system called Seed. The aim of this thesis is to devise, implement and evaluate (3) scoring schemes for the software system. The first scoring scheme is based on the sum of the thermodynamics free energy, based on the nearest neighbor model. We then present a general framework for evaluation of RNA structures using statistical regression analysis. The third scoring scheme to be validated is based on the framework of minimum description length principle.

We implemented and validated the above scoring schemes on four different data sets having varying range of complexity. The first two were derived from selected members of UTRdb database where the coding region is flanked by two untranslated regions (5' UTR and 3' UTR). The others were assembled using a subset of the sequences from Masoumi and Turcotte (IJBRA, 1(2), 230–245, 2005). By three measures, positive predicted value, sensitivity and Matthews correlation coefficient, our methods performed well on the data sets and showed significant ranking statistics. Also, our first method compares favorably with state-of-the-art tool, RNAprofile. For small motifs, the scoring methods are able to rank motifs with high PPV/sensitivity, often 100%. The top ranked motifs were used as input constraints for MFOLD, a widely used tool for RNA secondary structure determination. They showed improvements in both PPV and sensitivity measurements of the foldings made.

Acknowledgements

I would like to thank my supervisor Marcel Turcotte for his guidance, patience and dedication to teaching. Without his endless help and support, I can not image how I could achieve this success. I am especially grateful for his suggestions in selecting my research topic, and helping me with difficulties. Also he greatly helped me improve my writing skills, which I believe, will benefit me throughout my future career. And finally, thanks to my parents, brothers and numerous friends who endured this long process with me, always offering support and love.

MOHAMMAD ANWAR

University of Ottawa
July 2006

To my parents and grandparents.

Contents

List of Tables

List of Figures

List of Abbreviations

A	Adenine. A purine base found in DNA and RNA
Base pair	The hydrogen bounded structure formed by two complementary nucleotides
C	Cytosine. A pyrimidine base found in DNA and RNA
Consensus sequence	A sequence that represents the most common nucleotide at each position in two or more homologue sequences
DNA	Deoxyribonucleic Acid
Fold	Often used synonymously with the term structural motif
G	Guanine. One of the two purine nucleotides found in DNA and RNA
Hydrogen bond	Interaction between molecules resulting from from the slight separation of charges that results from polar covalent bonds
HSL3	Histone 3' UTR stem-loop structure
IRE	Iron responsive element
IUPAC	International Union of Pure and Applied Chemistry
Motif	Short conserved region of DNA/RNA/protein
MDL	Minimum description length
Nucleotide	A purine or pyrimidine base attached to a five carbon sugar, to which a mono-, di-, or triphosphate is also attached. The mono metric unit of DNA and RNA. Also known as 'base'
NN	Nearest neighbor
PPV	Positive predicted value. Fraction of the predicted base pairs that are also present in the reference structure
Purine	One of the two types of nitrogenous base found in nucleotides comprises Adenine and Guanine
Pyrimidine	One of the two types of nitrogenous base found in nucleotides comprises Cytosine and Uracil (or Thymine)

RNA	Ribonucleic Acid
Sensitivity	Fraction of the base pairs from the reference structure that are correctly predicted
T	Thymine. One of the two pyrimidine nucleotides found in DNA
tRNA	Transfer RNA
U	Uracil. One of the two pyrimidine nucleotides found in RNA
UTR	Untranslated Region. Regions of mature RNA that do not code for proteins, including 5' UTR and 3' UTR. UTR contains information for regulation of translation and mRNA stability

Chapter 1

Introduction

While the main role of DNA (deoxyribonucleic acid) is the storage of genetic information, RNA (ribonucleic acid), a versatile biopolymer has many different biological roles to fulfill. The widely recognized roles include carrying and encoding genetic information as messenger RNA (mRNA) and transfer RNA (tRNA). In addition to these, RNA has been implicated in several other biological events. Recent developments have shown that RNAs also have important regulatory functions. Examples include microRNAs that regulate the expression of genes by binding to the 3' untranslated regions of specific mRNAs. They exercise control over those RNAs that code for proteins. The function of these regulatory elements depends on the presence of motifs that are conserved both in structure and more loosely in sequence.

In the work by [37, 38], a multi-staged methodology has been implemented to explore the space of secondary structure motifs in a set of unaligned sequences. Having developed the algorithm, evaluation scheme(s) for measuring the quality of motifs generated is of paramount importance. This thesis stands as an extension of [37], by incorporating different scoring schemes to rank the motifs.

In this chapter, we first give background information on RNA, followed by a brief introduction to secondary structure prediction and the methods available today in Section 2 and 3 respectively.

Figure 1.1: Gene expression: the production of a protein from the genetic material, DNA. The DNA forms a template for the transcribed messenger RNA which is then translated into a protein.

A summary of the work by [38] is given in Section 4, and we give motivations to support why this work is important. The performance measures used to evaluate the predictions of our scoring models are detailed in Section 5 followed by a summary of the contributions in Section 6. The last section gives a brief outline of this thesis.

1.1 Background information

Gene expression occurs in two major stages – transcription and translation. In transcription, the gene is copied to produce an RNA molecule (a primary transcript) with essentially the same sequence as the gene. In case of eukaryotes, the primary transcripts are processed by splicing to remove introns and forming a mature transcript or messenger RNA that contains only exons. This mature transcript is then used in the process of protein synthesis via translation. This process is shown in Figure 1.1.

Differential gene expression is achieved by regulating transcription and translation. Transcriptional regulation is mostly effected through transcription factors binding to specific recognition sequence motifs located upstream of the coding region of the regulated gene. These transcription elements can either enhance or suppress the synthesis of mRNA. Gene expression is therefore controlled by the availability and activity of different transcription factors.

Recent research shows RNA to be an active and central player in the molecular genetics – a molecule that contains genetic information and acts as a biological catalyst. One of the key roles in the cellular processes is performed by regulatory molecule called microRNAs (miRNAs).

Table 1.1: Different functions of RNA.

Type of RNA	Function of RNA
ribosomal - rRNA	mRNA translation
transfer - tRNA	mRNA translation
messenger - mRNA	protein translation/regulatory
heterogeneous nuclear - hnRNA	intermediates of mRNAs
small cytoplasmic - scRNA	signal recognition particle, tRNA process
small nuclear - snRNA	mRNA processing, poly A, histone 3' process
small nucleolar - snoRNA	rRNA processing/maturation/methylation
regulatory RNAs	regulation of transcription and translation

This RNA molecule regulates the expression of protein genes by targeting a complementary region of their mRNAs. MicroRNAs constitute one of the most abundant class of regulatory molecules, and are key to many developmental process [4]. Post-transcriptional regulation of gene expression often involves secondary structure elements located in the untranslated regions of mRNAs [36].

These untranslated regions share structural features that are correlated with their biological function. Finding these similar structural features in a group of RNA sequences believed to share the same function can provide substantial information predicting which parts of the sequence are functional.

Apart from the vital roles played by RNA, there are other types of RNA which are involved in other important cellular functions. Table 1.1 lists the main types of RNAs and their functions. The structure is more important in performing their respective cellular functions.

Several databases have been designed lately to provide a comprehensive perspective and understanding of RNA structures and their functions. RNAdb is a collection of more than 800 unique empirically studied non-coding RNAs including many associated with diseases and/or development process [39]. Sequence and functional elements of the 5' and 3' untranslated regions of eukaryotic mRNAs are collected in the UTRdb and UTRsite databases [42]. Finally,

Rfam culls a large collection of multiple sequence alignments and covariance models covering many common non-coding RNA families [14]. These databases are available online and provide criteria based searches.

1.1.1 Terminology

RNA (ribonucleic acid) molecules are long polymer composed of a chain of four different nucleotides – Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). An official notation has been devised to describe ambiguous positions, e.g. R denotes a purine (one of the two types of nitrogenous base comprising Adenine and Guanine). See Appendix A for the list of all IUPAC codes. The two ends of a RNA sequence are distinguished by having a free phosphate group, denoted by the 5' at one end, and a free hydroxyl group, denoted by 3' at the other. The formation of a phosphodiester bond[1] between a phosphate group of one nucleotide to the deoxyribose sugar of another exhibits directionality in the sequence. The linear arrangements of nucleotides, so called primary structure of RNA (and DNA), proceeds in the 5' to 3' direction. The base pair between nucleotides form hydrogen bonds and act to thermodynamically stabilize the molecule in the cell. The most common hydrogen bonds are between C and G, and between A and U. They are called Watson-Crick (WC) pairs. C:G pairs are more stable, because they involve three hydrogen bonds, as opposed to A:U pairs, which involve two hydrogen bonds. Another frequent observed base pair is the G:U base pair, which allows uracil (U) to base pair with guanine (G).

Regions in the RNA fold together through base pair interactions to make intricate secondary and tertiary structures that are essential for correct biological function. The secondary structure is closely related to the function of the molecule and is conserved in evolution, despite potential sequence divergence. RNA secondary structure is generally divided into helices (contiguous base pairs), and various kinds of loops (unpaired nucleotides surrounded by helices). Commonly

[1]The covalent chemical bond that connects the phosphate group of one nucleotide to the deoxyribose sugar of another.

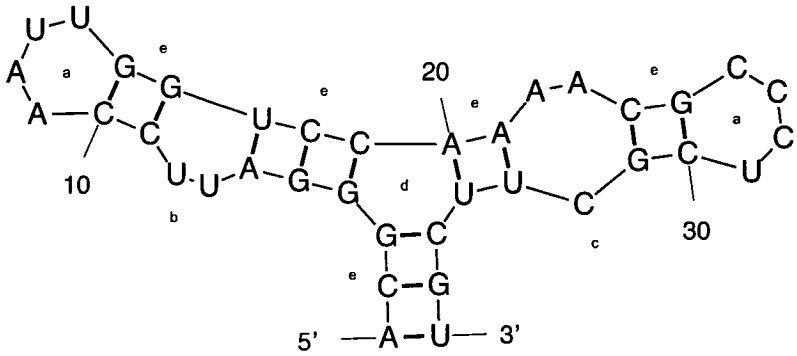

Figure 1.2: Decomposition of RNA structure in hairpin loops (a), bulges (b), internal loops (c), multi-loops (d) and stems (e).

described secondary structure elements include duplexes, single stranded regions, hairpins, internal and multi-loops (see Figure 1.2).

A stem is a double stranded (paired) region. A hairpin loop is where RNA folds back on itself. An internal loop is where a short unpaired region exists between two stems. If the internal loop is asymmetrical and only one strand forms a loop, while the other continues directly from one stem to the other, it is referred to as a bulge. In a multi-branched loop, several stems come together. RNA secondary structures can also include pseudoknots; a long-range interaction, where a loop pairs with another unpaired region.

Secondary structure of RNA for a given sequence is usually represented by the list of base pairs taking place between nucleotides, with the following constraints:

1. No nucleotide takes part in more than one (base) pair;
2. Base pairs never cross: if nucleotide in position i of the sequence is paired with nucleotide at position j, where $i < j$, and nucleotide position $k > i$ is paired with l, then either $i < j < k < l$ or $i < k < l < j$, but never $i < k < j < l$;
3. Loop minimum size: if nucleotide in position i of the sequence is paired with nucleotide

at position j, where $i < j$, then $|j - i| > 4$;

4. Pairing rules: only Watson-Crick (A:T, G:C) and Wobble (G:U) base pairs are allowed.

These constraints allow the secondary structure to be represented in a very space efficient way. One of them is the dot-bracket notation in which a structure on a sequence of length n is represented by a string of equal length consisting of matching brackets and dots. A base pair between base i and j is represented by a '(' at position i and a ')' at position j, unpaired bases are represented by dots. The structure shown in Figure 1.2 is denoted as

```
((((((..((....))))))((..((....)).)))))
```

1.2 Sequence and structure similarity

1.2.1 Sequences related by common ancestry

Evolutionary related sequences are known to adopt a similar structure. The term **homology** is used to designate sequences that are related by common ancestry. There are two processes that produce homologous sequences. Speciation is the process by which new species arise. Following a speciation event, the molecular sequences of each organism are evolving independently. **Othologous** sequences are corresponding sequences in different species arising from a common ancestor. The degree of sequence similarity between orthologs is related to the evolutionary distance of their species. That is orthologous sequences from closely related organisms, such as rodents, are highly similar. Whilst the similarity of sequences from distant organisms, such as *Saccharomyces cerevisiae* (baker's yeast) and *Fugu rubripes* (Japanese puffer fish), is tenuous. Othologous sequences are known to have a similar structure and generally identical function. The ribosome is a large complex consisting of several RNAs and proteins. It plays an essential role in the translation of messenger RNAs to proteins. Ribosomal RNAs are found in all kingdoms of life. Their structure is highly conserved and their function is identical.

The second process generating homologous sequences is gene duplication. One of the mechanisms by which organisms evolve is gene duplication. During recombination, homologous chromosomes are aligned to exchange genetic information. If the chromosomes break at slightly different locations, information will be duplicated on one of the chromosomes and deleted on the other. On a chromosome containing a duplicated segment, one copy of the gene fulfills the original function of the gene whilst the other copy is free to evolve (suffer mutations). Free of selective pressure one copy will drift and may eventually acquire a new generally related but distinct function. Sequences that are the result of gene duplication events are called **paralogs**. Paralogs are known to have a similar structure but a distinct, although related, function. Transfer RNA molecules are a family of RNAs that play an important role in the translation of messengers RNAs to proteins. These adapter molecules are used to translate triplet codons into the 20 amino acids. All the organisms have several tRNA genes, each one encoding a specific amino acid. At least some of these genes are believed to be paralogs [54]. All the tRNAs have common structural features.

1.2.2 Mechanistically related sequences

Sequences that are not related by common ancestry can independently acquire similar characteristics through **convergent evolution**. These similar structures are called **analogous structures** or **homoplasies**. For example, within eukaryotic cells, groups of genes are being translated at the same location within the cytoplasm. This creates gradients within the cell that are important for several developmental processes. Specific sequence elements, or zip codes, within the untranslated regions of the genes, are used to route (transport) messenger RNAs to these locations. Such genes are generally not evolutionary related and the extent of their similarity is limited to short (analogous) stem-loop structures [26, 25].

1.2.3 Nucleic acids databases

To facilitate understanding of nucleic acid structures, several databases has been assembled. The compilation of tRNA sequences by Sprinzl and colleagues comprises some 5,800 tRNA gene sequences from 111 organisms [47]. Most tRNA sequences are known to fold into a similar cloverleaf structure. The Comparative RNA Web (CRW) site regroups sequence and structure information of ribosomal RNAs (5S, 16S and 23S rRNA), transfer RNAs and catalytic introns (group I and II) [9, 10]. As of July 2006, considering ribosomal RNAs only, CRW contains 767 5S sequences, 10,149 16S sequences and 758 23S sequences. Each of these RNA families consists of homologous sequences. More recently, databases of regulatory elements have been put together. Often, these RNA families consists of analogous sequences. Examples of databases of regulatory RNAs include the noncoding regulatory RNAs database [48] as well as RNAdb [40]. These databases, as well as other nucleic acid databases, are available on Internet.

1.3 Prediction methods

Since the function of all biomolecules, including nucleic acids, is closely related to their (secondary) structure, biologists frequently need to determine this information. Because experimental approaches are time consuming, there is a need for computational tools capable of making accurate predictions. Figure 1.3 summarizes the choices that are available. Whenever a homologous sequence with known structure is available, and the sequence similarity is sufficient for making a reliable alignment, structural information can be transferred from one sequence to another accurately. Often no such sequence is available. However, if a database search can identify a set of sequences that are closely related, a multiple sequence alignment can be built and analyzed to identify pairs of columns where compensatory changes are observed (covariation analysis). The recent determination by X ray crystallography of the 30S and 50S ribosomal RNAs demonstrated the reliability of comparative sequence analysis for the determination of RNA secondary structure [18]. Sometimes 1) no homologous sequences are

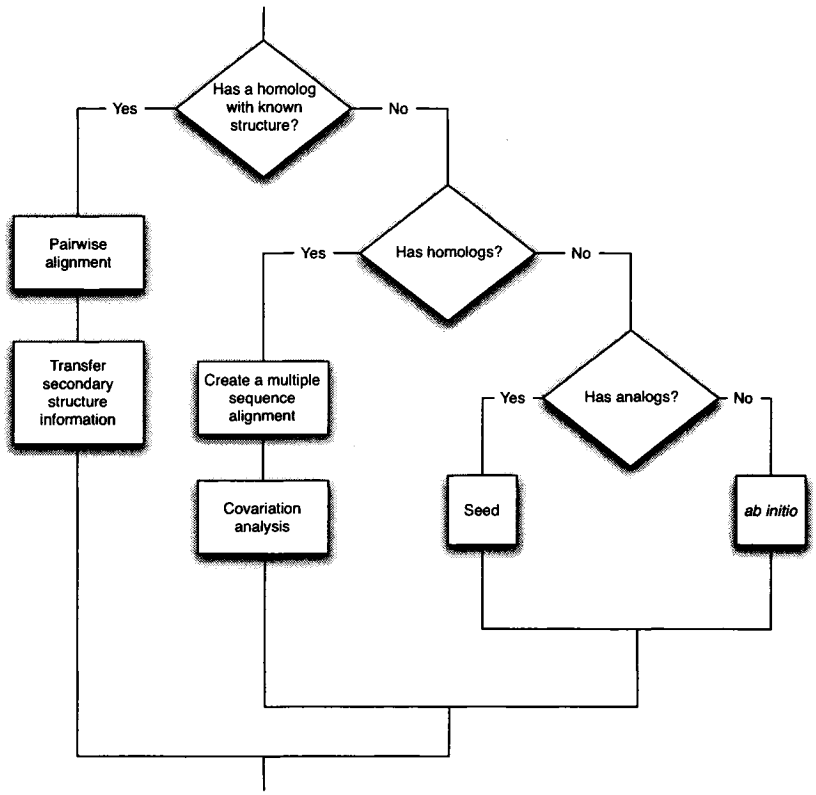

Figure 1.3: Secondary structure prediction.

available or, 2) the similarity of the sequences is so low that structural information would be needed for building the correct alignment or, 3) the sequences of interest are analogous and cannot be aligned on their entire length. If this is the case then Seed is the method of choice. Finally, if no homologous or analogous sequences are found, the biologist must rely on *ab initio* methods, such as MFOLD.

A group of methods fall under the category of predicting structure through energy minimization. Here, the free energy is assigned as the sum of independent contributions using nearest neighbor model [5]. MFOLD [33, 55], RNAFold [22, 57] and PairFold [44] are the examples in which this approach has been implemented.

There are algorithms that address the prediction of a consensus structure for a set of aligned RNA sequences. They make use of sets of evolutionary related sequences in order to infer a conserved secondary structure. The Pfold [27] algorithm derives a phylogenetic tree from the alignment using a maximum likelihood method over the possible trees calculated from the substitution (distance) matrices. The substitution matrices are based on large sets of published alignments of tRNAs and large sub-unit ribosomal RNAs. The tree is then used together with the alignment to infer the most likely structure. The dynamic programming algorithm of Alifold uses thermodynamics stability and compensatory mutations to generate a consensus secondary structure of a multiple RNA sequence alignment [21].

Often the functional elements or motifs are not conserved across the entire sequence, instead they are restricted to parts of the sequence. The UTR elements such as IRE and SEICS are two such examples. Also, the alignment may not be readily available due to low similarity in the sequences. In order to detect these conserved motifs, algorithms like FOLDALIGN [13] and RNAProfile [41] have been developed. These algorithms are not constrained by the initial alignment. FOLDALIGN uses a greedy strategy to create multiple alignments of the input sequences. A dynamic programming approach is used to perform the pairwise alignment. Starting with the pairwise alignment, the multiple alignment is progressively constructed by aligning a new sequence to an already existing alignment. The time complexity of FOLDALIGN restricts the usage for longer sequences. The RNAProfile algorithm outputs the most conserved regions (subsequences) of a set of unaligned sequences according to a similarity measure that accounts for both sequence and secondary structure. The initial step consists of selecting candidate regions from the input sequences that are later analyzed. These regions are derived by dynamic programming with a thermodynamic energy model. The algorithm proceeds by aligning each region in the first sequence with the regions in the second. The resulting alignments are converted into position specific profiles that represent the frequency of each nucleotide at each position in the alignment. These profiles are ordered based on the alignment scores and only a fixed number of best-scoring profiles are kept. After this, the candidate regions from the third

sequence are aligned to each profile and a new set of high scoring profiles is saved. This procedure continues until the candidate regions from all of the sequences have been aligned. The output is the highest scoring profiles after all sequences have been processed. The algorithm is expected to work for sequences less than 100 nucleotides long as it has been shown in [11] that nearest neighbor energy parameters works well for shorter RNA sequences or when the contact distance between the base pairs is less than 100 nucleotides.

To further address the requirement for computational tools to analyze RNA sequences, various methods have been developed that search for RNAs with given predefined models in a set of sequences. For example, RNAMotif [29] uses descriptors to describe rules that the motif instances must follow in order to be retrieved from the sequence database. The RSEARCH [24] algorithm is a local alignment search tool thats allows the user to search a database of sequences for homologous of a single RNA sequence with known secondary structure.

1.4 Seed software system

Seed is an exploratory tool to help studying the secondary structure landscape of RNA molecules. Its input consists of k sequences that are assumed to fold into a common secondary structure. Since the algorithm is intended to be mainly used for the study of analogous sequences, the common secondary structure motifs that are found often represent a small portion of each sequence (the algorithm does not seek to find a common global fold). Furthermore, because the assumption that all the input sequences are folding into one common structure could be wrong, the algorithm reports all the motifs that are found in a subset of the input sequences. The user controls the size of the subsets.

Seed was originally developed by Truong Nguyen [37, 38]. That work showed that the approach presented below 1) can be used to exhaustively search a constrained search space and that 2) this search space contains secondary structure motifs that are closely related to the native structures. However, the thesis proposed only the simplest scheme, information content, to

rank the search space. Herein, data sets for which known structures exist were selected and used to develop scoring schemes for ranking secondary structure motifs such that the ones that resemble the native structures more closely appear at the top of the list.

Exact solutions to the consensus secondary structure prediction problem require exponential resources (time and space) with respect to the number of input sequences. Our group is developing heuristics requiring polynomial amounts of resources. However, due to the high exponent, the search space is further reduced in a way that prevents these algorithms from finding common local structures. Therefore, these algorithms cannot be applied to analogous sequences. Seed is being developed specifically for tackling such problems.

The mandatory assumption for using Seed is that the majority of the input sequences will be sharing a common secondary structure motif. If this is the case then the search space of all the secondary structure motifs inducible from any of these sequences (i.e. the ones containing the shared motif) must contain an instance of the shared motif. Based on this premise, this algorithm elects one of the input sequences as the seed sequence and uses it to induce a secondary structure motifs search space. The algorithm traverses this search space selecting motifs that are also found in the remaining sequences. Selecting an outlier sequence as the seed would prevent the algorithm from finding a shared motif. Here, all the input sequences are known to contain at least one instance of the motif. The effect of selecting an outlier as the seed is beyond the scope of this thesis.

Below is a detailed description of the steps of the Seed algorithm. Its output is a list of motifs all matching at least some subset of the input sequences. The size of the subsets is controlled by the user. For all the experiments presented here, the motifs that are found match at least 70 % of the input sequences; in other words, given 10 input sequences, motifs will be found that match at least 7 input sequences. Each motif is simply a list of stems, $s_1, s_2 \ldots s_m$, where each s_i consists of a left and a right hand side. A left hand side is a fixed length sequence motif consisting of specific (**A**, **C**, **G** or **U**) or generic nucleotides (the letter **N** designates any nucleotide). The

right hand side is constrained to be the reverse complement of its corresponding left hand side. Let s_{il} and s_{ir} denote the left and right hand side of the ith stem of a motif. The following constraints are imposed on the stems:

- Minimum distance. In the current implementation of Seed, the two parts of a stem are separated by a fixed number of nucleotides denoted d and determined by the inference process. Assuming that s_{il} is a sequence motif x nucleotides long matching at position y, the 5' end of s_{ir} will be at position $y + x + d$. A user defined parameter controls the minimum value of d. For all the experiments presented here, the minimum value for d is 4;

- A motif consists of a list of stems and no two stems can overlap. Without loss of generality, let us assume that $s_{i\bullet}$ matches the seed sequence at a position before that of $s_{j\bullet}$, where the notation $s_{i\bullet}$ is used to denote either s_{il} or s_{ir}, similarly for $s_{j\bullet}$. Let $s_{i\bullet}$ be an x nucleotides long sequence motif matching at position y and $s_{j\bullet}$ be a z nucleotides long sequence motif. The no overlap constraint says that $y \geq x + d$. This must be true for all pairs $s_{i\bullet}$ and $s_{j\bullet}$;

- A motif consists of a list of stems and the distance between all the stems is fixed. Without loss of generality, the distances can be specified pairwise. In the implementation, each stem represents the distance to the next stem in the list. With no overlaps, the relationship between two stems, s_i and s_j, is either **embedded** or **adjacent**.

1.4.1 Schematic illustration of the algorithm

The algorithm consists of three main steps: the selection of the seed sequence, the construction of the most specific motif, and the general-to-specific search of the motif space. We present here a schematic illustration of the different steps. In this example, there are seven input sequences.

First step of the algorithm

The sequence RD0260 has been selected to be the seed sequence and is used to construct a set of building blocks that will be combined in the third step of the algorithm.

```
>RD0260 (*)
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA
>RD0500
GCCCGGGUGGUGUAGUGGCCCAUCAUACGACCCUGUCACGGUCGUGACGCGGGUUCAAAUCCCGCCUCGGGCGCCA
>RD1140
GGCCCCAUAGCGAAGUUGGUUAUCGCGCCUCCCUGUCACGGAGGAGAUCACGGGUUCGAGUCCCGUUGGGGUCGCCA
>RD2640
GGGAUUGUAGUUCAAUUGGUCAGAGCACCGCCCUGUCAAGGCGGAAGAUGCGGGUUCGAGCCCCGUCAGUCCCGCCA
>RE2140
GCCCCCAUCGUCUAGAGGCCUAGGACACCUCCCUUUCACGGAGGCGACAGGGAUUCGAAUUCCCUUGGGGGUACCA
>RE6781
UCCGUCGUAGUCUAGGUGGUUAGGAUACUCGGCUCUCACCCGAGAGACCCGGGUUCGAGUCCCGGCGACGGAACCA
>RF6320
GUCGCAAUGGUGUAGUUGGGAGCAUGACAGACUGAAGAUCUGUUGGUCAUCGGUUCGAUCCCGGUUUGUGACACCA
```

Second step of the algorithm

First, all the complementary regions from the seed sequence are enumerated. The user has control over the minimum length of the complementary regions, distances, number of allowed mismatches as well as the range (in number) of complementary regions to be discovered across the given set of sequences. See Appendix B for the complete list of Seed parameters and their description.

Second, using sequence information from the seed sequence, all the base pair instantiations are computed. More specifically, all the base pairs of every stem motifs are replaced by the actual base pair that occurs in the seed sequence in a combinatorial manner. The algorithm has now generated a list of one stem motifs that are generic, specific or a mixture of generic and specified pairs. This list is filtered to retain the motifs that are also found in the remaining sequences (minimum support).

a) Constructing all the stems inducible from the seed sequence and

```
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((.....)))

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((((((((.....)))))))))
```

⋮

```
                                                    NNNNNNNNNNNNNNNN
                                                    (((.....)))

                                                        NNNNNNNNNN
                                                        (((...)))
```

b) Instantiation of all its base pairs.

```
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((.....)))

NCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGN
(((.....)))

GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNU
(((.....)))

NNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNUNN
(((.....)))

GCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGU
(((.....)))

GCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNUGU
(((.....)))
```

⋮

Third step of the algorithm

In this step, the multi-stem motifs are constructed by selecting and composing two motifs. Motifs that are not structurally possible are discarded. The illustration shows

a) Composition using nested motifs and

```
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA

                GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC
                (((.....................................)))
                                  +
                      NNNUNNNNNNNNGNNN
                      (((((......)))))
                                  =
                GNNNNNNNUNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNC
                (((..(((((......)))))...................)))
```

b) Composition using adjacent motifs.

```
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA

                              CNUNNNNNGNG
                              (((.....)))
                                    +
                                                GNNUNNNNGNNC
                                                ((((....))))
                                    =
                              CNUNNNNNGNGNNNGNNUNNNNGNNC
                              (((.....)))...((((....))))
```

At the end of the run, Seed outputs a list of motifs that contains a mixture of single and multi-stem motifs, that are structural, partially or fully instantiated. For the above example, Seed outputs a total of 5,518 motifs. Here, we study various methods for ranking the motifs such that motifs corresponding to native folds are ranked highest.

1.4.2 Motif description

RNA molecules fold into secondary and tertiary structures that are responsible for their diverse functional activities. Many of these RNA structures are assembled from a collection of RNA structural motifs. These basic building blocks are used repeatedly, and in various combinations, to form different RNA structures. This accounts to define their unique structural and functional properties.

At the basic level, RNA secondary structure consists of nucleotides that are either paired or unpaired, where pairing includes Watson-Crick or Wobble base pairs. In general, most base pairings are adjacent with other pairings that constitutes a stem (helix). The combination of one of more stems are separated by unpaired, single stranded nucleotides that form a RNA structure. For example, the given structure below shows a tRNA motif. It consists of an outer stem that is nested with 3 adjacent stems. Also, the adjacent stems are interspersed by unpaired nucleotides.

```
(((((((..((((.........)))).(((((.......)))))).....(((((.......)))))))))))
```

Various filtering method have been proposed to select the initial motifs that are later combined to form secondary structures. However, the number of potential secondary structures for a given RNA sequence is exponential with respect to its length [56]. This is a large number even for short sequences. Since Seed generates an ensemble of motifs, a need arises to differentiate the biologically relevant motifs from the rest. The algorithm used in Seed is exhaustive and independent of any scoring scheme. This provides an ideal environment to evaluate and study scoring methods to identify native folds. Having good scoring functions to eliminate invalid motifs can be implemented to prune the search space and hence accelerate the running time of the algorithm.

As previously mentioned, RNA molecules are involved with a vast number of cellular functions.

It is compelling to consider more possible roles that remain to be discovered. Developing computationally accurate structure prediction methods will help us better understand the molecules ability to function. The work described in this thesis investigates different scoring schemes to identify putative functional regions in RNA sequences. We test the functions for their ability to correctly identify (known) structures with high positive predicted value/sensitivity.

1.5 Performance measure

In this section we explain the different measures used to calculate the accuracy of the ranked consensus structures by the different scoring methods proposed.

1.5.1 Matthews correlation coefficient, PPV, sensitivity

In order to get a quantitative measure of the accuracy of the ranked secondary structures, we chose to use sensitivity and positive predicted value (PPV) using three measures of accuracy. These measures have been traditionally used in the field of bioinformatics. We call **references**, the secondary structures that were obtained from curated databases. We define as **true positives** (TP) the base pairs that are occurring in both structures, reference and predicted, **false positives** (FP), the base pairs that are occurring in the predicted structure but not in the reference one, and **false negatives** (FN), the base pairs that are occurring in the reference structure but not in the predicted one.

In [33] a predicted base pair is considered to be consistent with the **reference** structure if the base pair is slipped by one position on one side of the helix. So the base pairs are correct if they are within one base of a trusted base pair. In other words, pairs i to j, $(i + 1)$ to j, $(i - 1)$ to j, i to $(j + 1)$, or i to $(j - 1)$ are considered to be consistent for comparative structure analysis (see Figure 1.4).

For our analysis, we use a strict comparison. Given the predicted and referenced structure, we

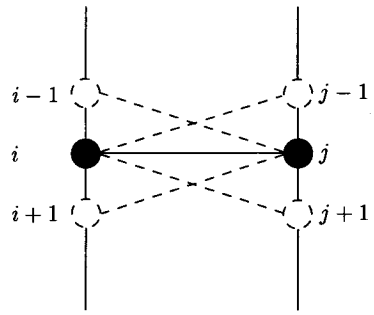

Figure 1.4: Comparison of structures. The solid line represents base pairs present in both structures while the dashed line indicates a base pair formation with an offset.

count the number of pairs that are correctly predicted. That is, if we predict base pair i, j, it is correct if and only if base pair i, j is present in the referenced structure. No offset/slippage is allowed.

The **positive predictive value** (PPV) is defined as the fraction of the predicted base pairs that are also present in the reference structure, TP/(TP + FP). The **sensitivity** is defined as the fraction of the base pairs from the reference structure that are correctly predicted, TP/(TP + FN). Finally, we also measured the **Matthews correlation coefficient**, as defined by Gorodkin, Stricklin and Stormo [13]:

$$\sqrt{\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \times \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}}$$

When a given motif matched the input sequence more than once, the performance indices of the match having the highest PPV are reported.

1.5.2 Distance

The measures described above quantify the agreement between the predicted and referenced structures with respect to the position of the base pair in the sequence under consideration. The above measures require the predicted and true structure to occur at the same location. A

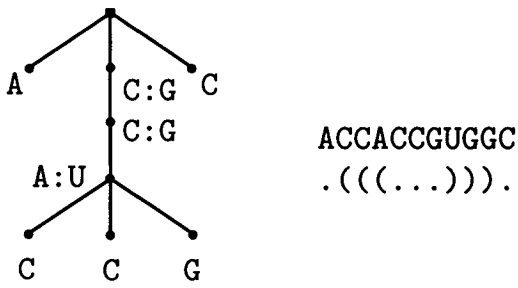

Figure 1.5: Tree representation for a 9 nucleotide long RNA molecule

structure that would be mostly correct, except that few bases force it to match with an offset gets no credit.

To obtain a purely structural comparison, we consider to measure the dissimilarity between the secondary structures. A simple measure would be the base pair distance, that is, the number of base pairs that have to be opened or closed to transform one structure into the other. However this comparison restricts both the structures to be of same length. Another measure to compare structures is by encoding the secondary structure by ordered trees, followed by the computation of edit distance. In the tree representation, secondary structure is represented as rooted tree, where the internal nodes represent the base pairs and the leafs represent the unpaired nucleotide. The root node is a virtual node that is added to avoid getting a forest representation instead of a tree representation. Figure 1.5 shows an example of a tree representation of a secondary structure. The edit distance defines a metric of the number of insertion, deletion and replacements of nodes in the tree. This approach is a part of the Vienna RNA package [22].

1.6 Thesis contributions

The research objectives were to develop scoring schemes for evaluating (ranking) RNA secondary structures. The main contributions of this thesis in the order they are discussed:

1. We first introduced several objective functions, combining free energy of motifs. A thorough analysis of the performance of these functions is performed on sets of UTR elements and more complex structures. We show that the thermodynamic based scoring functions are effective to discriminate native folds from the rest. This work has been published in the BMC Bionformatics journal [1];

2. We then selected a subset of these functions and formulated statistical models using techniques from regression analysis to derive a scoring model. We have observed that the models can generally identify motifs with high measures of sensitivity and positive predicted value to known motifs. This work has been published at the CCECE conference (Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada, 2006) [3];

3. We implemented and validated a scoring scheme based on the framework of Minimum Description Length principle. Here, the secondary structure prediction problem is seen as finding regularities in data. In Chapter 4, we propose a simple encoding system for consensus RNA secondary structure motifs. The performance of this scoring scheme is analyzed on the same data sets mentioned above. We found that the scoring method produces competing structures in comparison to the above methods. As far as it is known to the author, this work represents the first MDL based methodology for evaluating RNA secondary structure motifs. A manuscript presenting this work has been published in the International Conference on Bioinformatics and Computational Biology, Las Vegas, Nevada, USA, 2006 [2].

4. Finally, we performed a rank based evaluation of the models described above. This is the first application of ranking statistics proposed by Rosset *et al.* [43] in the field of bioinformatics.

1.7 Reading guidelines

The remainder of the thesis describes the data set used and preliminary experiments carried out for this work (Chapter 2), presents the scoring schemes for ranking the motifs, the results obtained and their discussion (Chapter 3, 4 & 5), and concludes with a comparison of the methods (Chapter 6) and future directions for this work.

Chapter 2

Experimental setup

In order to analyze and compare the performance of the proposed scoring methods, we used four different data sets. In this chapter, we discuss the data sets and the parameter settings used to run Seed for each data set.

2.1 Data set

2.1.1 Single stem motifs

This class of conserved structure elements exert their function as regulatory motifs in the untranslated region (UTR) of the primary transcript or the mature mRNA. The structural patterns of a histone 3' UTR stem-loop structure (HSL3) and an iron responsive element (IRE) were used in this study. Their specifications was extracted from the UTRdb database release 19 [42].

Histone 3' UTR stem-loop (HSL3)

This highly conserved stem-loop structure contains a six base stem and a four base loop. This stem-loop structure plays a different role in the nucleus and in the cytoplasm. In the nucleus it

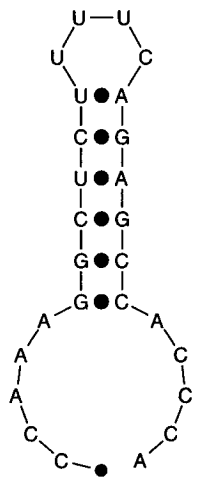

Figure 2.1: Consensus secondary structure for Histone 3 family.

is involved in pre-mRNA processing and nucleocytoplasmic transport, whereas in the cytoplasm it enhances translation efficiency and regulates histone mRNA stability. All the histone mRNAs analyzed so far do not have G in the four base loop. The 5' flanking sequence consensus is CCAAA and the 3' flanking sequence consensus is ACCCA or ACCA with cleavage occurring after the A. Figure 2.1 shows the graphical representation of the the histone 3 family. Also shown below is the dot-bracket notation of the family.

```
.....((((((....))))))....
CCAAAGGYYYUHHUHARRRCCACCCA
```

See Appendix A for the IUPAC representations of the single letter codes. Table 2.1 shows the characteristics of the HSL3 test data set used for our study.

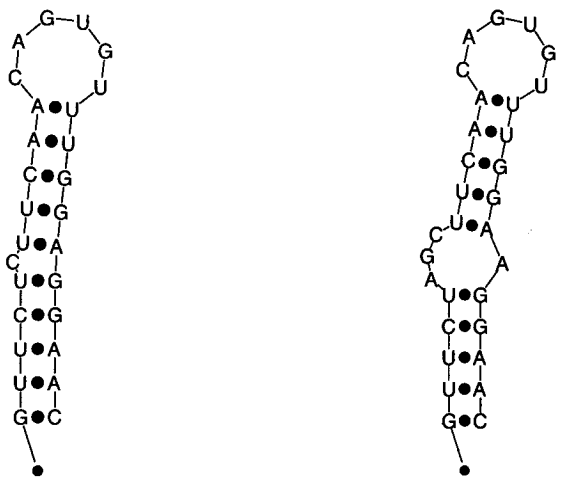

Figure 2.2: Consensus secondary structures for IRE family.

Iron responsive element (IRE)

The Iron Responsive Element is a particular hairpin structure located in the 5' UTR or in the 3' UTR of various mRNAs coding for proteins involved in cellular iron metabolism. The interaction of the IRE with regulatory proteins modulates the translation of the mRNA according to the amount of iron present in the cell. Two alternative IRE consensus structures have been discovered. Some IREs present an unpaired nucleotide (usually cytosine) along the stem, whereas in others the cytosine nucleotide and two additional bases seem to oppose one free 3' nucleotide, as shown below.

```
(((((.(((((......)))))))))).
NNNNNCNNNNNCAGWGHNNNNNNNNNNN
(((((...(((((......)))))).)))))
NNNNNNNCNNNNNCAGWGHNNNNNNNNNNN
```

Table 2.2 shows the characteristics of the IRE test data set used for our study. A graphical representation of the second structure is shown in Figure 2.2.

2.1.2 Multi-stem motifs

This class of motifs consists of more complex secondary structures involving higher number of stems of variable length. Examples included in this study are tRNA and 5S rRNA, which actively participate in the protein synthesis. For HSL3 and IRE, the motifs sought represents only a small fraction of the input sequences (local motif). In the following examples the motif encompasses most of the sequence (global motif). The secondary structure descriptions of the tRNAs entries were extracted from the compilation of Sprinzl *et al.* [45, 46]. For 5S, the descriptions were extracted from the Comparative RNA Web Site [17, 8, 10].

tRNA

Transfer RNA, or tRNA is the information adapter molecule. It is the direct interface between amino-acid sequence of a protein and the information stored in DNA. It is a key participant in decoding the information in DNA (RNA). tRNAs are small molecules having a length of 74–93 nucleotides. The secondary structure includes four short double-helical elements and three loops (D, anti-codon, and T loop). The following is an accepted secondary structure validated using comparative sequence analysis.

```
(((((((..((((.........)))).(((((.......)))))).....(((((.......)))))))))))
```

Table 2.3 shows the characteristics of the tRNA test data set used for our study. Figure 2.3 shows the secondary structure for the tRNA RD0260.

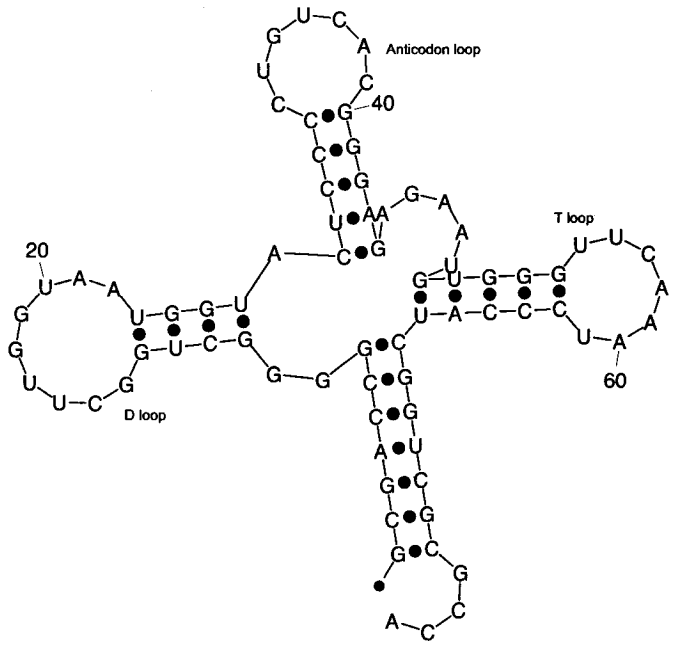

Figure 2.3: Consensus secondary structures for tRNA family.

5S rRNA

The ribosomal 5S RNAs are approximately 120 nucleotides long and are an integral part of the large ribosomal subunit. The structure of most known 5S molecules forms a consensus structure consisting of five helical regions (1 – 5), two hairpin loops (C and D), and two internal loops (B and E). An internal loop (loop A) is formed at the junction of three helices, namely 1, 2, and 5. Figure 2.4 shows the consensus secondary structure. Below is the dot-bracket notation for the same structure used. Note that the notation continues on the next line.

```
((((((((((.....((((((((....((((((.............))))..)))...)))))))..))
  .(((((....(((((((((...)))))))))....)))))...)))))))))).
```

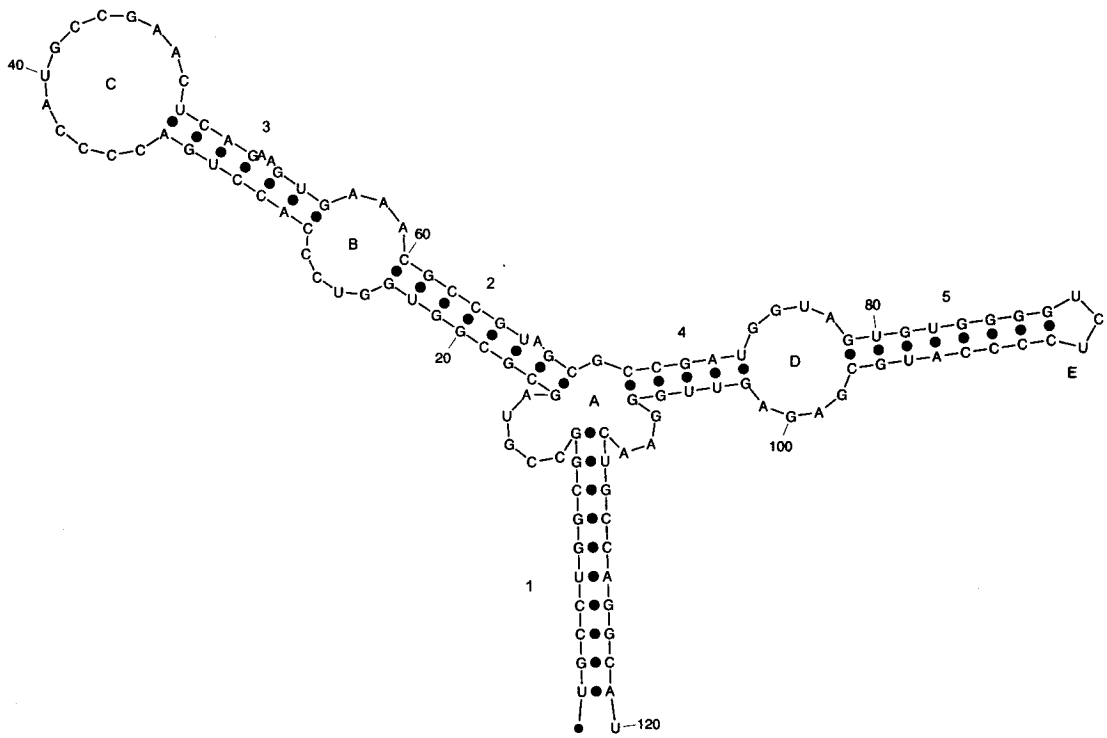

Figure 2.4: Consensus secondary structure for 5S rRNA family

2.2 Preliminary experiments

To create the input data (secondary structure motifs) for the scoring methods, a total of six experiments were conducted: HSL3 01, IRE 01, HSL3 02, IRE 02, tRNA and 5S. All the input sequences selected contain exactly one copy of the consensus structure. However, in general the possibility that certain sequences have been erroneously included in the input cannot be excluded. Secondly, the input perhaps consists of more than one fold family. Hence, a user defined level of noise is tolerated. The support is defined as the fraction of the input sequences containing a given motif; in the experiments presented here, we set the support to 70%. The parameter settings of Seed for all the experiments are detailed in Table 2.6. For all the experiments, the minimum total number base pairs was 5, option **skip_keep_longest_stem** was used, G:U base pair were allowed, no mismatch was allowed, and no time limit was set. The range was set to 1 for tRNA and 5S only. See Appendix B for the list of parameters. The statistics for each experiment is detailed in Table 2.5.

In the first experiment, HSL3 01, strict parameters were used to verify the presence of native folds in the search space. More precisely we searched for a consensus structure containing exactly only one stem segment made of at least six base pairs. The maximum distance for any two nucleotides involved in a base pair was set to 30. Seed discovered 65 motifs. The second data set consists of 14 5' UTR sequences that are known to contain an IRE motif. We searched for a structure consisting of one or two segments of at least three base pairs each; notice that one or two stem-loop structures could be formed with these options. The maximum distance for the nucleotides of a base pair was set to 30. For this experiment (IRE 01), a total of 32 structures were found.

Next, we reran the first two experiments with less restrictive parameters, suitable for finding stem-loop structures in general. In particular, no restrictions were imposed for the maximum number of segments that could be used to form a motif, except that the maximum distance for the elements of pair should be 30 nucleotides or less. The algorithm found 357 motifs for HSL3

02 and 110 motifs for IRE 02.

We also studied structures that are more complex. The next data set consist of 7 tRNA sequences representing diverse level of difficulties for MFOLD. In order to discover multiple stem structures, we placed no restrictions on the maximum distance between elements of a base pair. Also, the length of the unpaired regions in matched sequences was allowed to vary by one nucleotide for added flexibility. The size of the output was reduced by reporting only the motifs consisting of three or more segments (complementary regions). Seed found 5,518 motifs. The last experiment was carried out on 7 5S ribosomal RNA sequences. The sequences represent diverse levels of difficulties for MFOLD. The same options as above (tRNA experiment) were used. A total of 364,505 motifs were found by the algorithm.

Table 2.1: Description and characteristics of the HSL3 test data set showing the length of the sequence and the motif start-end location in the sequence.

UTR ID	Description	Length (bp)	Motif Location
3HSA054868	*Homo sapiens* cDNA FLJ33901 fis	1955	23,45
3HSA041812	3' UTR in Human DNA sequence from clone RP1-221C16 on chromosome 6	1251	39,61
3HSA027954	3' UTR in Human DNA sequence from clone RP1-221C16 on chromosome 6	91	32,54
3HSA034695	3' UTR in *Homo sapiens* histone 1, H1c	87	23,45
3HSA079397	3' UTR in *Homo sapiens* histone 1, H3h	838	27,49
3HSA082131	3' UTR in *Homo sapiens* H2A histone family	1146	45,67
3HSA047510	3' UTR in *Homo sapiens* histone gene complex 1	1225	35,57
3HSA083260	3' UTR in *Homo sapiens* histone 1, H4i	917	29,51
3HSA083338	3' UTR in *Homo sapiens* histone 1, H2ag	1844	46,68
3HSA083659	3' UTR in *Homo sapiens* histone 1, H4f	85	18,40
3HSA048427	3' UTR in *Homo sapiens* histone 1, H2ac	1211	39,61
3HSA049188	3' UTR in *Homo sapiens* histone 2, H2aa	136	63,85
3HSA084501	3' UTR in *Homo sapiens* histone 4, H4	1592	29,51
3HSA086570	3' UTR in *Homo sapiens* histone 1, H3d	106	31,53
3HSA086915	3' UTR in *Homo sapiens* histone 1, H2am	694	33,55
3HSA087013	3' UTR in *Homo sapiens* histone 1, H3d	941	31,53
3HSA089561	3' UTR in *Homo sapiens* histone 1, H3e	1896	26,48
3HSA058723	3' UTR in Human DNA sequence from clone RP1-97D16 on chromosome 6p21.3-22.2	82	22,44
3HSA058724	3' UTR in Human DNA sequence from clone RP1-97D16 on chromosome 6p21.3-22.2	99	40,62
3MMU017942	3' UTR in *Mus musculus* H2A histone family	935	42,64
3MMU040716	3' UTR in *Mus musculus* histone 2, H3c2	527	26,48
3MMU043604	3' UTR in *Mus musculus* histone 3, H2ba	231	24,46
3MMU045939	3' UTR in *Mus musculus* histone 1, H2bg	220	25,47
3MMU046704	3' UTR in *Mus musculus* histone 3, H2bb	1194	37,59
3MMU004991	3' UTR in *Mus musculus* histone H3.1-D (H3-D)	58	33,55
3MMU004994	3' UTR in *Mus musculus* histone H3.2-616 (H3-616)	51	26,48
3MMU004995	3' UTR in *Mus musculus* histone H3.2-616 (H3-616)	60	34,56
3DRE005245	3' UTR in *Danio rerio* similar to H2A histone family	171	23,45

Table 2.2: Description and characteristics of the IRE test data set showing the length of the sequence and the motif start-end location in the sequence.

UTR ID	Description	Length (bp)	Motif Location
5DLE000003	5' UTR in *Delphinapterus leucas* erythroid-specific 5-aminolevulinic acid synthase (ALS2) mRNA	58	4,26
5HSA021933	5' UTR in *Homo sapiens* mRNA	2188	236,258
5HSA033035	5' UTR in *Homo sapiens* mRNA	314	87,113
5HSA060296	5' UTR in *Homo sapiens* solute carrier family 40	430	203,229
5HSA072191	5' UTR in *Homo sapiens* ferritin, light polypeptide, mRNA	203	34,56
5HSA073036	5' UTR in *Homo sapiens* ferritin, heavy polypeptide 1, mRNA	210	37,59
5HSA079314	5' UTR in *Homo sapiens* ferritin, light polypeptide, mRNA	192	34,56
5MMU018600	5' UTR in *Mus musculus* ferritin light chain 1, mRNA	187	15,37
5MMU025452	5' UTR in *Mus musculus* RIKEN cDNA 5730461K03 gene, mRNA	283	232,257
5MMU027798	5' UTR in *Mus musculus* protein phosphatase 1, regulatory (inhibitor) subunit 1B, mRNA	356	246,271
5MMU032372	5' UTR in *Mus musculus* solute carrier family 40 (iron-regulated transporter), member 1, mRNA	300	72,98
5RNO004780	5' UTR in *Rattus norvegicus* cell adhesion regulator (CAR1) mRNA	306	80,106
5RNO005974	5' UTR in *Rattus norvegicus* 5-aminolevulinate synthase (ALS2) mRNA	58	19,41
5RNO007816	5' UTR in *Rattus norvegicus* cDNA clone MGC:72644	209	39,61

Table 2.3: Description and characteristics of the tRNA test data set showing the length of the sequence. For the given sequences, the pairwise sequence identity varies from 41.8–68.8%

ID	Description	Length (bp)
RD0260	Asp *Phage T5 (Virus) RD0500*	77
RD0500	Asp *Haloferax volcanii* (Archae)	76
RD1140	Asp *Mycoplasma Capric* (Eubacteria)	77
RD2640	Asp *Mycoplasma Capric* (Chloropast)	77
RE2140	Glu *Synechocystis sp.* (Eubacteria)	76
RE6781	Glu *Hordeum vulgare* (Chloroplast)	76
RF6320	Phe *Schizosaccharomyces pombe* (Cytoplasm, Fungi)	76

Table 2.4: Description and characteristics of the 5S test data set showing the length of the sequence. For the given sequences, the pairwise sequence identity varies from 49.2–99.1%

ID	Description	Length (bp)
V00336	*Escherichia coli*	120
X02627	*Agrobacterium tumefaciens*	120
X04585	*Rhodobacter capsulatus*	119
M24839	*Bacillus stearothermophilus*	117
X67579	*Saccharomyces cerevisiae*	118
AJ251080	*Geobacillus stearothermophilus*, strain NCA 1503	117
M25591	*Geobacillus stearothermophilus*	117

Table 2.5: Details of the execution of Seed for all the 6 experiments (**Expr**) showing the different/total number of motifs discovered (**Motifs**), the number of matches made (**Matches**), the **space** and **time** (in seconds) taken for each run. The run time is composed of 1) construction of tree, 2) motif generation (search space) and 3) general to specific search of search space. **Length** shows the minimum and maximum length of the sequences in the corresponding data set. Suffix of 01 and 02 indicate a different setup used on the same data set. Finally, **Sequences** shows the number of sequences present in the data set. These experiments were carried out on a Sun Fire V20z computer, 2 processor server, AMD Opteron 248 (2.2 GHz), 8 Gigabytes, Solaris 9, a single processor was used.

Expr	Sequences	Length	Motifs	Matches	Space	Time
HSL3 01	28	51–1,955	2/65	2,016	1.83 Mbytes	< 1s
IRE 01	14	58–2,188	29/32	42,462	0.39 Mbytes	5s
HSL3 02	28	51–1,955	232/357	1,945,328	1.37 Mbytes	5m 21s
IRE 02	14	58–2,188	102/110	167,076	0.46 Mbytes	25s
tRNA	7	76–77	2,010/5,518	3,407,012	9.40 Mbytes	6m 11s
5S	7	117–120	24,645/364,505	152,741,463	0.52 Gbytes	7h 40m

Table 2.6: Parameters settings for each experiment (**Expr**): **stem_min_len** is the minimum length of complementary regions that are reported in the first step of the algorithm, **min_num_stem** specifies the minimum number of complementary segments each motif should have to be part of the output, **max_num_stem** is a stopping criteria, the algorithm does not extend motifs beyond this threshold, finally, **stem_max_separation** is the maximum distance between the start and the end positions of a helix.

Expr	stem_min_len	min_num_stem	max_num_stem	stem_max_separation
HSL3 01	6	1	1	30
HSL3 02	3	1	-	30
IRE 01	3	1	2	30
IRE 02	3	1	-	30
tRNA	3	3	-	-
5S	3	3	-	-

Chapter 3

Thermodynamics based scoring functions

In Chapter 1, we presented a brief introduction to algorithms that predict consensus secondary structures from a set of sequences or a single RNA sequence. One of the most popular methods for prediction from a single sequence is through free energy minimization. A review of the developments of this paradigm can be found at [22, 32, 55]. While the algorithms predict satisfactory results for shorter sequences, the accuracy of these methods are tied to the accuracy of the folding routines. To address the limitations of free energy minimization, we introduce herein, several objective functions based on thermodynamic principles and evaluate their potential to discriminate native folds from the rest. This work has been published in the BMC Bionformatics journal [1];

3.1 Thermodynamics of RNA

To understand the full mechanisms of action of RNA, the structure of RNA needs to be understood. Often, the secondary structure of RNA sequence is solved before its tertiary structure because of two reasons. First, there are more accurate methods for determining the secondary structure of an RNA sequence. Secondly, the secondary structure helps in designing the con-

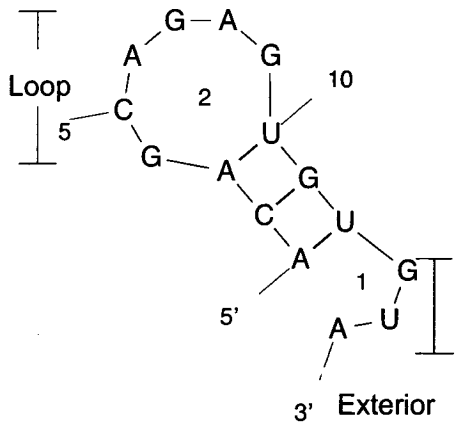

Figure 3.1: Nearest neighbor model takes into account the interactions between the nucleotides in the region labeled Loop and base pairs in the stem. The nucleotides in the region labeled Exterior, however, cannot pair with any of the nucleotides in the loop region without forming a pseudoknot. Thus, pairwise interactions between these bases and those in the loop are ignored by nearest neighbor energy rules.

structs for tertiary structure determination. This section discusses one of the methods for RNA secondary structure prediction: RNA folding thermodynamics, and how this can be incorporated as a scoring method to evaluate the motifs.

3.1.1 Free energy minimization

Thermodynamic principles suggest that the structure with the lowest free energy should be the most stable, and hence, the active fold. Thus, various methods for RNA secondary structure prediction rely on free energy minimization using nearest neighbor parameters for predicting the stability of an RNA secondary structure. Programs based on this technique include MFOLD [33, 55], RNAStructure [32] and RNAFold [22, 57].

3.1.2 Nearest neighbor model

Nearest neighbor model is the most common model used for predicting the stability of RNA. It assumes that the thermodynamic stability of each base pair depends only on the most adjacent pairs; the total free energy is the sum of each contribution. Terms that affect the thermodynamic stability have been studied in [44]. The base composition, unpaired terminal nucleotide and dangling ends have significant effect on the stability. However, detailed study for loops have not been as thoroughly investigated as done for base paired regions. Hairpin loops are the ones that have been studied extensively as compared to bulge and internal loops. Also, pairwise interactions that cross base paired stem regions are not considered in this model, see Figure 3.1. Not considering pseudoknots speeds up the calculations considerably.

Figure 3.2 shows the calculation of the nearest neighbor stability in a small RNA molecule. Factors such as helical stacking, loop initiation, and unpaired nucleotide stacking contribute to the total conformational free energy. Favorable free energy increments are less than zero. The thermodynamic values for a helical region are calculated as the sum of the adjacent stacked pairs. For example, the consecutive C:G base pairs contribute -3.3 kcal/mol. The free energy for loop regions have unfavorable increments called loop initiation energies. For example, the hairpin loop of four nucleotides has an initiation energy of 4.9 kcal/mol. These increments are different for hairpin, bulge and internal loops, see Table 3.1. Unpaired nucleotides in (internal) loops can provide favorable energy increments as either stacked nucleotides or as mismatched pairs. The mismatches G-A, U-U, and others adjacent to C:G pairs are assumed to make the loop more stable by -2.7, -2.5, and -1.5 kcal/mol, respectively. While mismatch adjacent to A-U pairs make the loop stable by -2.2, -2.0, and -1.0 kcal/mol, respectively. Dangling ends, such as 3'-most G in the figure, provide -0.2 towards the stabilization. The total free energy amounts to -5.5 kcal/mol. Similarly, using the nearest neighbor parameters presented in [44] the thermodynamic property of any RNA sequences can be predicted.

The two main methods of determining the thermodynamic parameters involve analyzing the

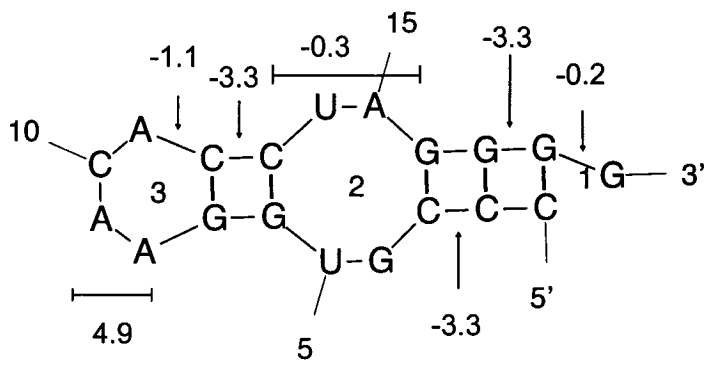

Figure 3.2: Prediction of conformational free energy for a conformation of **5'-CCCGUGGAACACCUAGGGG-3'**. Each contributing free energy increment is labeled. The total free energy is the sum of each increment.

melting curves (optical spectroscopy) of RNA molecules and by calorimetry [44]. In optical spectroscopy, the ultraviolet absorption changes are measured between the ordered conformation and disrupted coil. Calorimetric methods measure the heat lost or gained due to structural change. The most convenient method of bringing change in the ordered structure and coil form is heating. The parametric information obtained by calorimetry for a particular RNA can be used for prediction of thermodynamics values of other RNA molecules with similar sequences [53].

3.1.3 Limitations of minimum free energy

Although the predictions of RNA secondary structures based on thermodynamic data have been on an average about 70% successful for sequences less than 100 nucleotides long [11], the performance of thermodynamics based methods is limited by thermodynamic models and parameters. The predictions give rise to numerous lowest free energy conformation making it difficult to distinguish the native conformation from the rest. There are several reasons as to why free energy based models may fail.

Table 3.1: Free energy parameters for loop initiations. * indicates parameters not derived from experimental measurements (reproduce from [44])

Size (nucleotides)	Hairpin	Bulge	Internal
1		3.9	
2		3.1	0.8
3	4.1	3.5	5.1
4	4.9	4.2*	4.9
5	4.4	4.8	5.3
6	5.0	5.0*	5.7
7	5.0	5.2*	5.9*
8	5.1*	5.3*	6.0*
9	5.2	5.4*	6.1*

Firstly, not all of the factors that determine the energy of a fold are understood, and computational time limitations would make it infeasible to include all of such influences. Secondly, many situations are highly approximated, especially the calculation of multi-loop structures. No tertiary interactions are included in the model, no pseudoknotted structures and no interactions with cellular environment are considered.

Longer sequences would require more instances of parameters to be considered. As a result, this will make any model based on free energy introduce more errors. The resulting prediction is partly due to the simplified or inexact model, partly to the algorithm, which is simplified for time complexity reasons.

3.1.4 Vienna RNA package

The Vienna RNA package is a software system that can be used to predict secondary structure of RNA. It is available both as a web interface and by compilation on Unix-based machines [22]. It uses a dynamic programming approach and the current set of thermodynamic parameters. Three kinds of dynamic programming algorithms are provided for structure prediction: the minimum free energy algorithm of [57], which yields a single optimal structure, the partition

function algorithm of [35], which calculates base pair probabilities in the thermodynamic ensemble, and the suboptimal folding algorithm of [52], which generates all suboptimal structures within a given energy range of the optimal energy.

This package consists of a collection of routines for the prediction and comparison of RNA secondary structures. One of our scoring method uses the C code library, RNAlib, for the calculation of free energy and distance between RNA secondary structures.

3.2 Evaluation of motifs

The purpose of a scoring function is to distinguish the biologically relevant motifs amongst an ensemble of motifs, that is, to approximate the biological meaning of the motif in terms of a mathematical function. Different scoring functions have been proposed, based on information theory, background knowledge of nucleotide frequency. Others evaluate the quality of a pattern based on some statistical significance [49, 50].

The Seed algorithm outputs a list of motifs discovered, each followed by the respective match(es) to the input sequences. See Appendix C for an example. There are a number of observations that we can expect from the output generated.

- A motif describes an ensemble of sequences. It is not always defined by a unique sequence of bases and variation is permitted. Typically the variation is represented by generic base pair, N:N' which indicates A:U, U:A, C:G, G:C, G:U or U:G base pair;
- The precise location of the motif in the sequence may not be important, as the length of the input sequences vary;
- In case of regulatory motifs, multiple occurrences of a motif in the same sequence may be important as each occurrence may give a molecule a greater chance of finding the binding site;

- The motif will be common to most of the sequences. This is made sure with the minimum support parameter.

We expect the fold of the molecule to be primarily determined by the bases involved. As a result, understanding the factors determining the stability of an RNA structure is important in predicting the correct structure. There are several factors contributing to the stability of a RNA structure. They are,

- The number of G:C vs A:U and G:U base pairs (stable structures contain higher energy bonds);
- Number of base pairs in stem (longer stems result in more bonds);
- The number of base pairs in a hairpin loop;
- The number of unpaired bases (unpaired bases decrease the stability).

An effort to circumvent the limitations of the nearest neighbor model is to have a linear combination of the free energy of multiple input sequences, when folded to a common structure [34, 30, 31]. As the number of input sequences increases, it is unlikely that the common structure(s) will fold into a bad minimum free energy. Although the results of using more sequences are promising, the prohibitive time/space complexity of these approaches are restricted to few sequences (< 4) that are less than hundred nucleotides long.

Considering the above observations, contributing factors and limitations to the model, we introduce several simple functions combining the free energy of all or some of the matches of a given motif. We also introduce a function that consists of the sum of the information content contributions from unpaired and paired regions. These functions are *TLeft*, *NTLeft*, *TFirst*, *NTFirst*, *TSum*, *NTSum*, *TBest*, *NTBest*, *TWorst*, *NTWorst*, *TAvg*, *NTAvg* and *TInfo*.

3.2.1 Scoring functions

Seed is a framework for finding conserved RNA motifs in a set of unaligned sequences. Since the algorithm is exhaustive and does not involve any scoring scheme, it allows for the independent study of functions for ranking the motifs *a posteriori*.

Information content has often been used in the context of sequence pattern discovery [50]. Accordingly, we include a function, *TInfo*, that consists of the sum of the information content contributions from unpaired and paired regions. Shannon uncertainty (H) was calculated for each loop position and was subtracted from the maximum uncertainty possible, to give the information content (in bits). $H = -\sum P_i \log_2 P_i$ summed over each base (i = A, U/T, G, C), where the observed nucleotide frequencies of each base i from the input sequences are used to estimate P_i. A nucleotide in a stem is base paired to its partner, which increases the information content relative to an unpaired nucleotide in a loop. There are sixteen possible two nucleotide pairs giving a maximum uncertainty of 4 bits. Out of these pairs, there are four Watson-Crick base pairs (A:T, T:A, G:C, C:G). For a Watson-Crick base pair, the uncertainty is 2 bits making the information content 4 bits − 2 bits = 2 bits. Assuming equiprobable events, if G:U (and U:G) base pair are considered, the uncertainty is reduced to 2.8 bits; the information content of this pairing is 4 bits − 2.8 bits = 1.2 bits. The resulting loop and stem information contents were added to calculate the total information content.

Each motif matches at least **min_support**$\times k$ sequences, and up to k sequences; where **min_support** is a user defined parameter. Also, certain motifs will occur more than once in any sequence. Thus, there are several possible approaches to calculate the free energy score for a given motif and set of matches. Furthermore, the secondary structure expression matcher traverses the suffix array of the input sequences, rather than the input sequences themselves. Consequently, the matches are reported in lexicographic order. The execution time can be slightly reduced by taking into account the free energy of the first reported match, rather than finding the best or leftmost one.

For a given motif, we define *TLeft* as the sum of the free energy of its leftmost occurrence in each sequence. *TFirst* has a slightly different definition. *TFirst* is the sum of the free energy of the first occurrence reported by the matcher in each sequence. This function was considered since it can slightly reduce the execution time. *TSum* is the sum of the free energy of all the occurrences in all the matching sequences. *TBest* is the sum of the lowest free energy match in each sequence. *TAvg* is the sum of the average free energy of all the matches per sequence. *TWorst* is the sum of the highest free energy match from each sequence. Finally, normalized variants of these scores are obtained by dividing each score by the number of matched sequences or total number of matches for *TSum*; the resulting scores are noted with the letter N followed by the name of the base score. In particular, given a set of k sequences and m matches, the functions are calculated as

$$TSum = \sum_i \sum_j m_{i,j} \qquad TLeft = \sum_i m_{i,j}\ ,\ j = 1$$

$$TAvg = \sum_i \frac{\sum_j m_{i,j}}{|j|} \qquad TBest = \sum_i \min_j m_{i,j}$$

$$TWorst = \sum_i \max_j m_{i,j} \qquad TFirst = \sum_i f_j(m_{i,j})$$

where $m_{i,j}$ is the free energy of jth match occurring in the ith sequence and function f returns the free energy of the match that comes first in the lexicographic order. Figure 3.3 illustrates the variety of scenarios of *TSum*, *TBest* and *TWorst* scoring functions.

Seed produces enormous amounts of secondary structures motifs. The functions defined capture different features of an RNA structure. We used two criteria to evaluate their value as candidates for ranking function. We measured the correlation coefficient for each function against the Matthews correlation coefficient (MCC, see Chapter 1 for a definition of these performance measures). Clearly, a perfectly correlated function would allow selecting the consensus structure with the best MCC. However since the primary objective of the ranking functions is to order the consensus structures from best to worst, rather than modeling the free energy, we also used the ranking based evaluation measures recently proposed by Rosset *et al.* [43]. Rank-

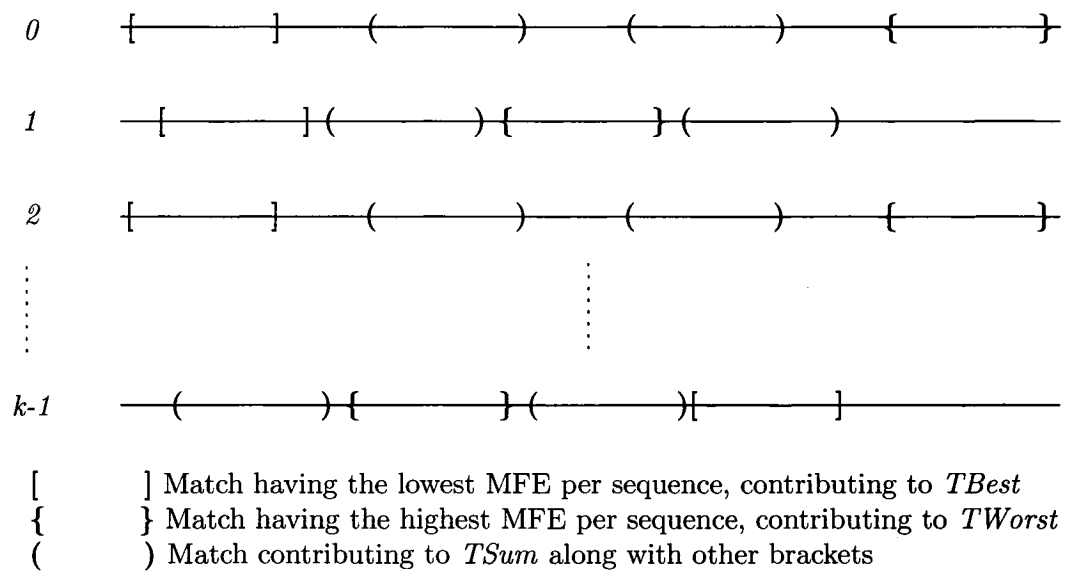

Figure 3.3: Schematic representation of scoring function calculation. Each bracket represents a match in the sequences.

ing statistics describe the relationship between scoring functions and the different performance measures. A detailed description of how ranking statistics are calculated and used for model (scoring function) comparison is discussed in Chapter 4.

3.3 Experiments

We ran the Seed algorithm on four sets of data to study the performance of the different scoring functions proposed. The first two experiments HSL3 01 and IRE 01 were designed to evaluate the suitability of this approach to identify automatically conserved stem-loop structures. Next, we relaxed the parameters and reran the first two experiments (HSL3 02, IRE 02) to allow stem-loop structures in general. Finally, we studied the performance of the functions on more complex structures, tRNA and 5S. See Chapter 2 for the parameter settings and the description of the data sets. We also compared the performance of top ranked motifs to RNAProfile [41] on

the same four data sets. This is a recently developed algorithm for finding conserved secondary structure motifs in a set of unaligned RNA sequences.

3.4 Results

Application of the scoring functions based on the methodologies specified previously showed interesting results. The experiments conducted show that the scoring functions can generally identify motifs with high positive predicted value and sensitivity to known motifs. The functions that we investigated are correlated to MCC in varying degrees, see Table 3.2. On average, the normalized version of the scoring functions show a higher correlation. Also, the ability of the functions to discriminate native folds is supported by the ranking statistics (Table 3.3). Furthermore, the various plots demonstrate that the algorithm successfully identified complete motifs and there is a strong correlation between the free energy and the different performance measures. Since *NTFirst* is fast to compute and has the largest $\hat{\rho}$ value, it has been used for the analysis of the results for all the six experiments. We first discuss here the results of the first four experiments and follow this with the results on the complex structures.

3.4.1 HSL3 data set

In the first experiment HSL3 01, Seed finds 65 motifs. All but four have a Matthews correlation coefficient score of 100%, i.e. both PPV and sensitivity are 100%. As expected, a negative correlation is observed for all the scoring functions except *TInfo*. Recall that the native structure is assumed to be the one with minimum free energy. The normalized variants of the scoring functions show a higher correlation, with *NTLeft* the highest (-0.959), see Table 3.2. The relationship between the information content (IC) and the performance index for this experiment is characteristic of the other experiments as well. Motifs with a high information content also have good PPV and sensitivity, see Figure 3.4. However, there are low information content motifs that have high PPV and sensitivity. For all twelve functions, the "best" scoring motif

also has the best PPV and sensitivity. The average number of matches per motif per sequence varies from 1 to 5, with an average of 1.2 matches per sequence.

In the second run on the same data set, HSL3 02, the parameters were relaxed. A total of 357 motifs were found, see Table 2.5. Nearly one third of the conserved motifs have a PPV of 100%. *NTWorst* has the best correlation to MCC (-0.978). All the correlation scores are -0.95 or better. For this experiment, the free energy scores are particularly effective to separate the high PPV/sensitivity motifs from the rest, see Figure 3.5. This is also confirmed by the ranking statistics; the value of $\hat{\rho}$ ranges from 0.864 to 0.926 (Table 3.3).

3.4.2 IRE data set

The third experiment (IRE 01) was carried on the IRE data set. A total of 32 structures were found, three of which have 100% PPV and more than 90% sensitivity. The number of matches per sequence varies from 1.1 to 3.9, with an average number of 1.9 matches per sequence. *TFirst* is the scoring function with the best correlation to MCC. Figure 3.6 illustrates the separation of native folds from the rest achieved using *NTFirst*. Except for *TInfo* and its normalized variant, the ranking statistics are high (> 0.859). For this experiment, strong correlation is seen for all the scoring functions, see Table 2.5.

We then varied the parameters for the IRE experiment (IRE 02) to allow stem-loop in general to be discovered. A total of 110 motifs were found, see Table 2.5. Of the scoring schemes based on free energy, *NTWorst* has the best correlation to MCC. This observation is similar to HSL3 02 experiment, where relaxed parameters were used. However, the magnitude of the correlation for the *TInfo* score is higher than that of *NTWorst*. The structural motifs with the highest MCC score rank first using *NTFirst*, see Figure 3.7. The ability of the energy functions to rank the motifs is good, with $\hat{\rho}$ ranging from 0.702 to 0.780.

3.4.3 tRNA data set

We also studied more complex structures. The results on the tRNA data set were promising. The algorithm discovered 5,518 motifs. *NTLeft* gave the highest correlation to MCC (−0.834). This scoring function also has the highest ranking statistics, see Table 3.3. The tRNA structures have been used as a benchmark for the development minimum free energy methods. Perhaps the free energy models are particularly proficient for this class of RNAs. Indeed, it is interesting to observe the high degree of correlation of *NTFirst* to MCC, particularly for the left tail of the distribution, see Figure 3.8. The motif with the highest *NTFirst* score has 16 base pairs, 100% PPV, 76.2% sensitivity. It matches 5 of the 7 input sequences.

3.4.4 5S data set

The last experiment was carried out on 7 5S ribosomal RNA sequences. A total of 364,505 consensus structures were found. The maximum sensitivity achieved is much lower than for the previous experiments with average sensitivity below 50%. Figure 3.9 shows the scatter plot of the correlation between *AvgSensitivity* and *TFirst*. *TSum* has the highest correlation to MCC. Better correlation scores are observed when focusing on motifs consisting of 15 or more base pairs, the correlation scores between *TSum* and *AvgPPV*, *AvgSensitivity* and *AvgMCC* are −0.783, −0.838 and −0.823, respectively. The PPV of the "best" motif according to *NTFirst* is high, on average 85%. However, there are many motifs with high PPV, 100%, for the intermediate *NTFirst* values.

3.5 Discussion of results

Application of the Seed algorithm against experimental data sequences produced a significant number of candidate consensus secondary structure motifs. The bulk of motifs obtained were, for most part, generic structures. For example, in the HSL3 02 experiment, 29 of 32 structures

were generic (structurally different). In case of IRE experiments the ratio was 232/357 and 102/110, respectively. However, for HSL3 01, tRNA and 5S, a higher degree of specificity was obtained, see Table 2.5. All the scoring functions were strongly correlated to MCC, however no function performed better than the rest. It is interesting to note that the objective function having the highest $\hat{\rho}$ values is *NTFirst*. This function is fastest to compute since the matcher stops after locating the first occurrence of the motif in each input sequence. The occurrence is also the one that comes first in the lexicographic order. The bias of this function is more difficult to rationalize that that of the other functions. The other objective functions are biased towards finding the leftmost, largest number of occurrences, best or worst free energy matches.

3.5.1 Free energy functions can identify motifs with high PPV values

With the free energy based functions, lowest free energy scoring motifs have a high PPV and sensitivity while highest scoring motifs have a low PPV/sensitivitiy. The functions display good correlation to PPV/sensitivity; the correlation to sensitivity is generally higher than to PPV. The ability of all the functions to rank the motifs is good, with $\hat{\rho}$ ranging from 0.84 to 0.87. For 5 out of 6 experiments, the top ranked motifs has the highest Matthews correlation coefficient.

For HSL3 02 and IRE 02 experiments, the parameters used were less restrictive. It is interesting to note that for both the experiments, *NTWorst* had the highest correlation to MCC. This could be due to the presence of more false positives as input constraints were relaxed.

For single stem structures, motifs with high information content were also found to have high PPV and sensitivity. However, the performance of IC decreases as the complexity of motifs sought increases. Overall, the free energy based ranking functions performed better, the average weighted ranking ($\hat{\rho}$) of *NTFirst* is 0.87, compared to 0.69 for *TInfo*.

3.5.2 Higher information content correspond to higher PPV motifs

In case of single stem motifs, HSL3 and IRE, there are few low information content motifs that have high PPV values. These motifs correspond to generic structures. As expected, information content base scoring function, *TInfo*, will rank such motifs lower than those containing sequence information. For complex structures, more instances of low information content and high PPV motifs can be seen, see Figure 3.8 and 3.9. This is probably due to the fact that the number of multi-stem motifs composed of two or more single stem motifs is possibly large. For example, in Figure 3.8, the first vertical line stretching across the PPV axis represents 858 structurally different motifs with the same information content value. All of these motifs are composed of 3 single stem motifs consisting of 3 base pairs. Out of these, 13 have a PPV of 100%. Intuitively, as more motifs are combined and instantiated, we tend to see an increase in their information content. However, since the formulation of complex (multi-stem) motifs is constrained to minimum support and valid composition, motifs of higher complexity are less in number. The data points in the top plots of Figure 3.8 and 3.9 serve to indicate this fact. We can see a reduction of higher IC motifs. Motifs with higher information content generally correspond to higher PPV values.

3.5.3 More matching sequences less sensitivity

For complex structures, tRNA and 5S, Seed is able to predict structures with high PPV, often 100%, but of lower sensitivity. This is due to the fact that a consensus structure motif is less representative of each matching sequence. As such, a consensus structure can never achieve 100%.

For tRNA, the top ranked motif according to *NTFirst* has 16 base pairs and matches 5 of the 7 input sequences. Of all the structures predicted by the Seed algorithm, this motif has the highest *AvgSensitivity* (76.2%). With 5S data set, the top ranked motif has 22 base pairs and an average sensitivity of 48.2%. This motif matches 5 of the given 7 input sequences and

has the highest *AvgSensitivity* in the search space. Thus, the energy based scoring functions successfully identify motifs with the highest sensitivity in the search space.

3.5.4 Comparison with RNAProfile

We compared the performance of top ranked motifs to RNAProfile on the same four data sets. Several experiments were ran, varying in the small increments the region parameters (minimum `l`, and maximum `L`), and the number of hairpins (`H`) sought was specified. The combinations of (parameters) values producing the best performance were used for comparison (HSL3 : `H=1, l=16, L=30`; IRE: `H=1, l= 20, L=40`; tRNA: `H=3, l=60, L=78`; 5S: `H=2, l=30, L=120`).

On the HSL3 data set, both algorithms identified motifs with 100% PPV and sensitivity. In case of IRE, the performance of both systems were comparable. Specifically, the PPV/sensitivity of the structures predicted by RNAProfile ranged from 71.4–100/62.5–100%, excluding one prediction which had no true positives. The motif predicted by Seed matched 10 out of 13 sequences with PPV/sensitivity ranging from 100–100/80–100%.

In case of complex structures, tRNA and 5S, Seed outperformed RNAProfile in terms of PPV. For the tRNA data set, the PPV/sensitivity range for RNAProfile and Seed were found to be 25–100/23.8–90.5% and 100–100/76.2–76.2%, respectively. RNAProfile failed to predict the overall Y shape of the 5S RNAs. The minimum and maximum PPV/sensitivity of motifs predicted by RNAProfile was found to be 0–79.3/0–60.5% whereas by Seed it is 77.3–86.4/44.7–50%.

3.6 Combining scoring functions

For all data sets, the energy scoring functions are able to detect (rank) meaningful motifs. However, in comparison, the objective function, *TInfo* is weakly correlated to the different performance measures. In the case of complex structures, higher information content motifs share the same performance measure with low information content motifs. The strategy of using

a linear combination of multiple objective functions can perhanps help alleviate this limitation. Higher information content can be used to pick conserved motifs. Combined with information content, the free energy functions can perhaps help detect not only conserved motifs, but also of high PPV/sensitivity in a data set. We explore this idea to build scoring functions using regression analysis in the next chapter.

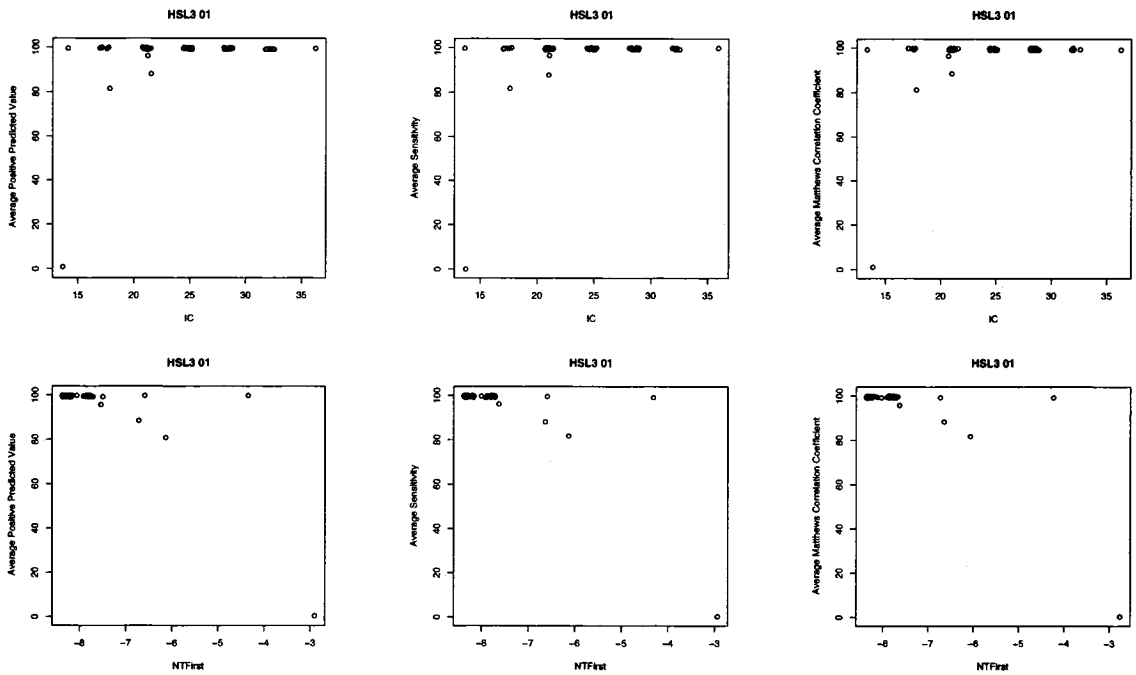

Figure 3.4: Performance diagrams for the HSL3 01 experiment. *IC* (top), *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the HSL3 dat aset. Each data point represents the predicted rank (according to *IC* and *NTFirst* respectively) and average performance index (PPV, sensitivity or MCC) for a consensus structure. For better visualization a random noise has been introduced so that the data points do not overlap.

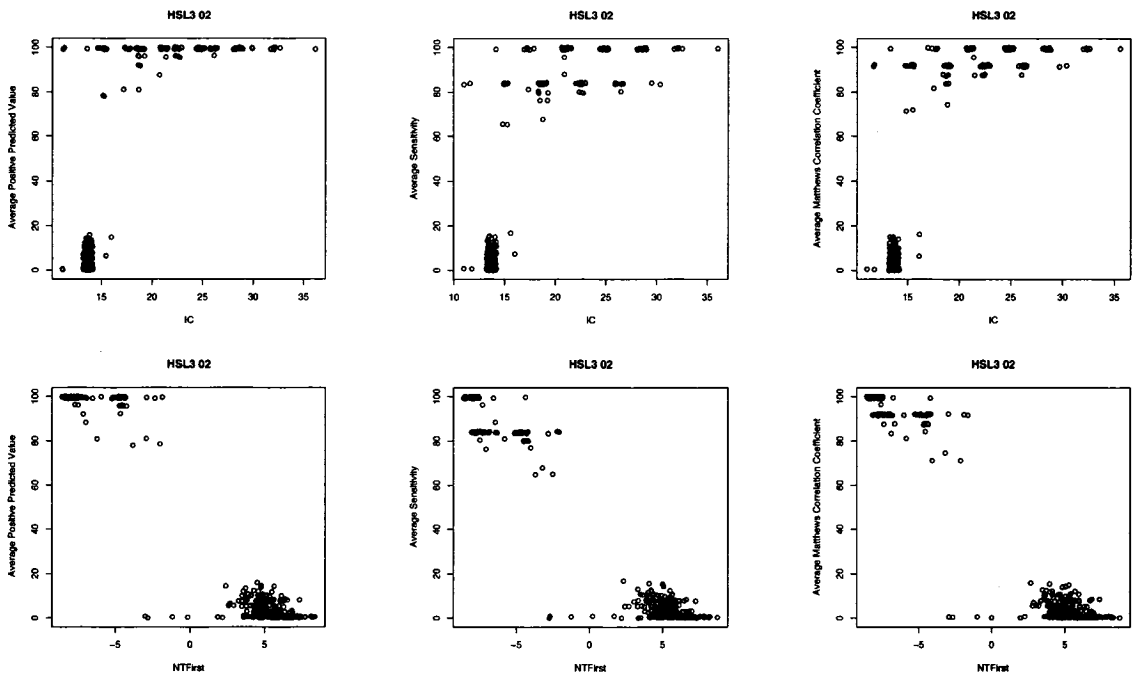

Figure 3.5: Performance diagrams for the HSL3 02 experiment. *IC* (top) *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the HSL3 data set.

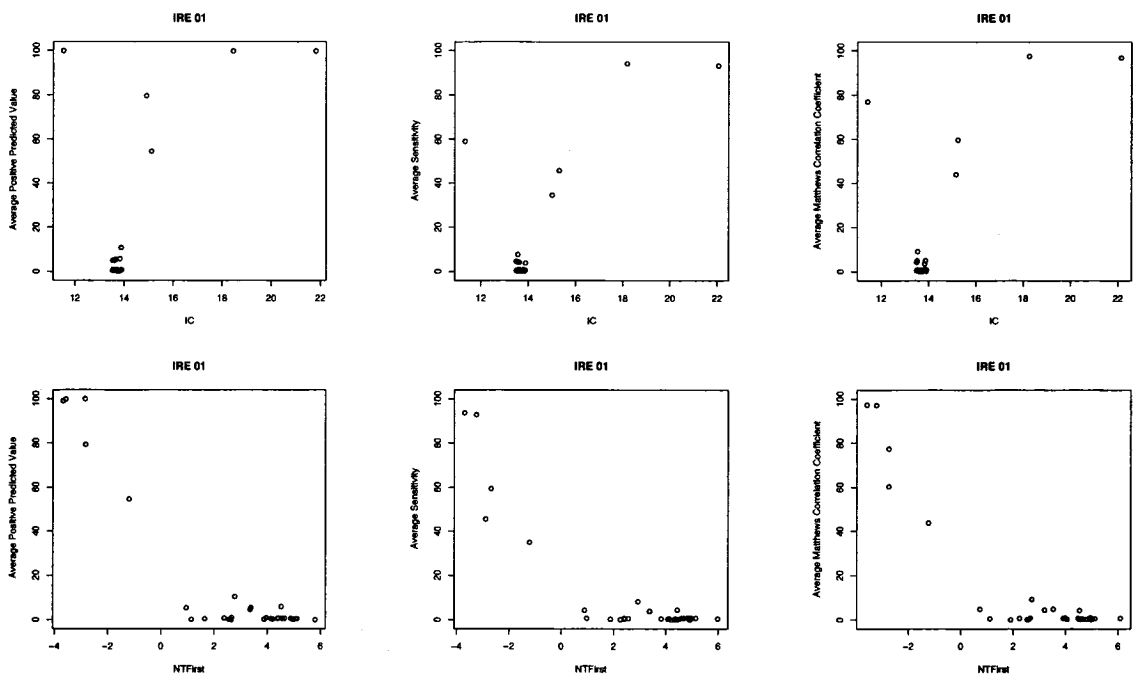

Figure 3.6: Performance diagrams for the IRE 01 experiment. *IC* (top) *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the IRE data set.

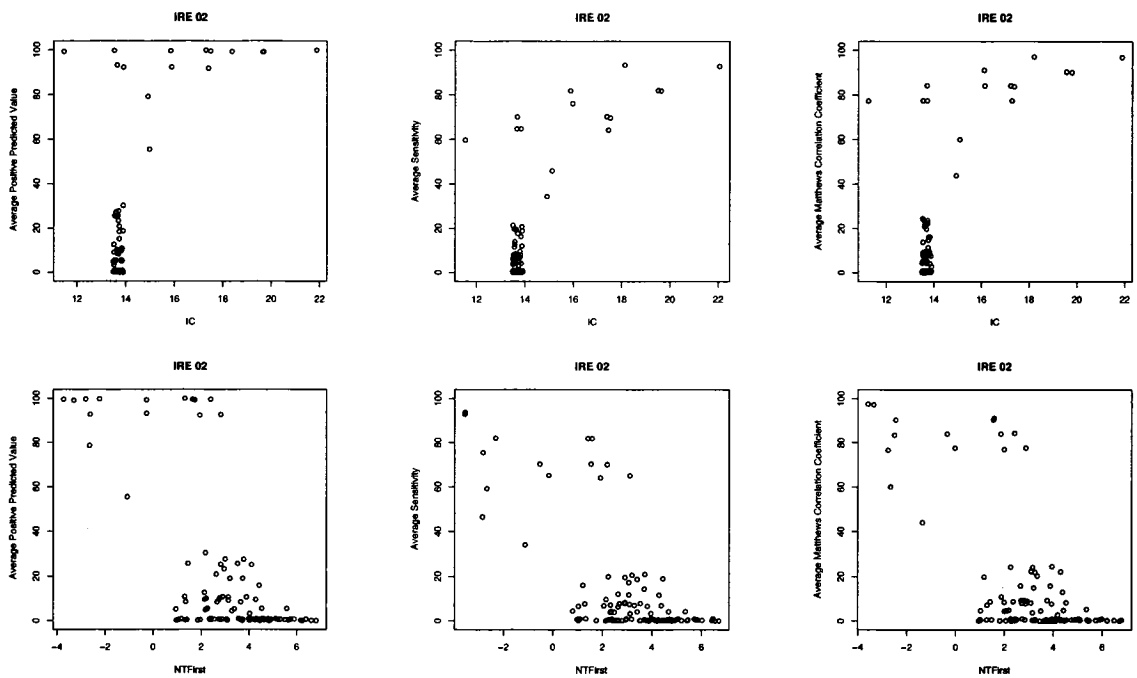

Figure 3.7: Performance diagrams for the IRE 02 experiment. *IC* (top) *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the IRE data set.

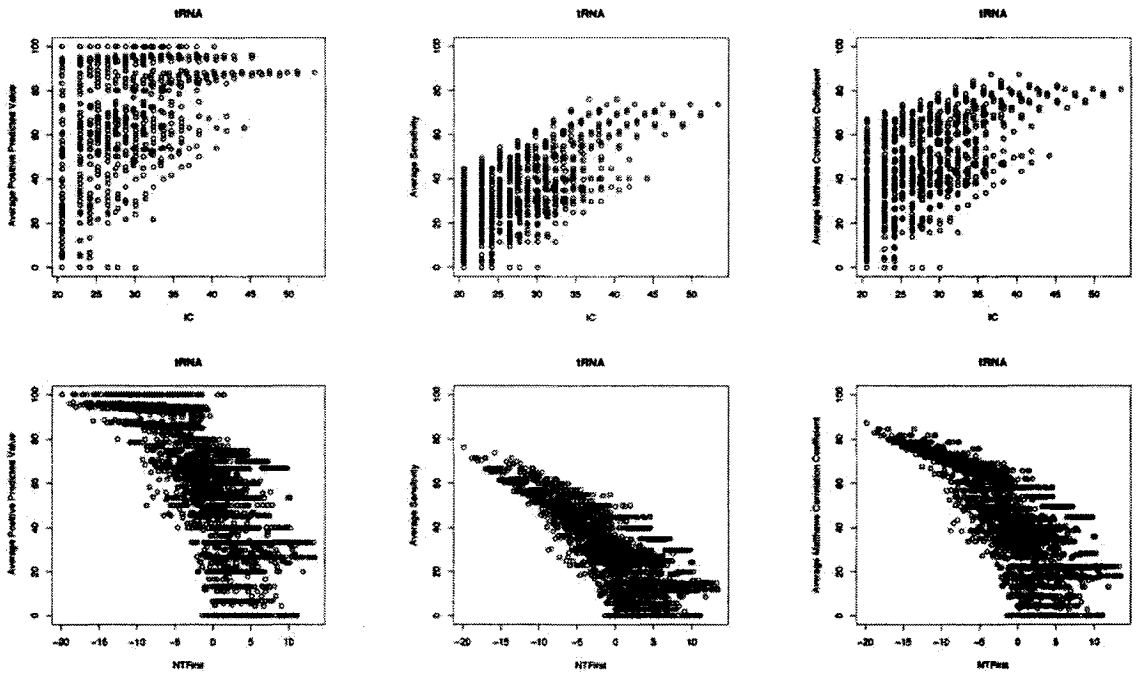

Figure 3.8: Performance diagrams for the tRNA experiment. *IC* (top) *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the tRNA data set.

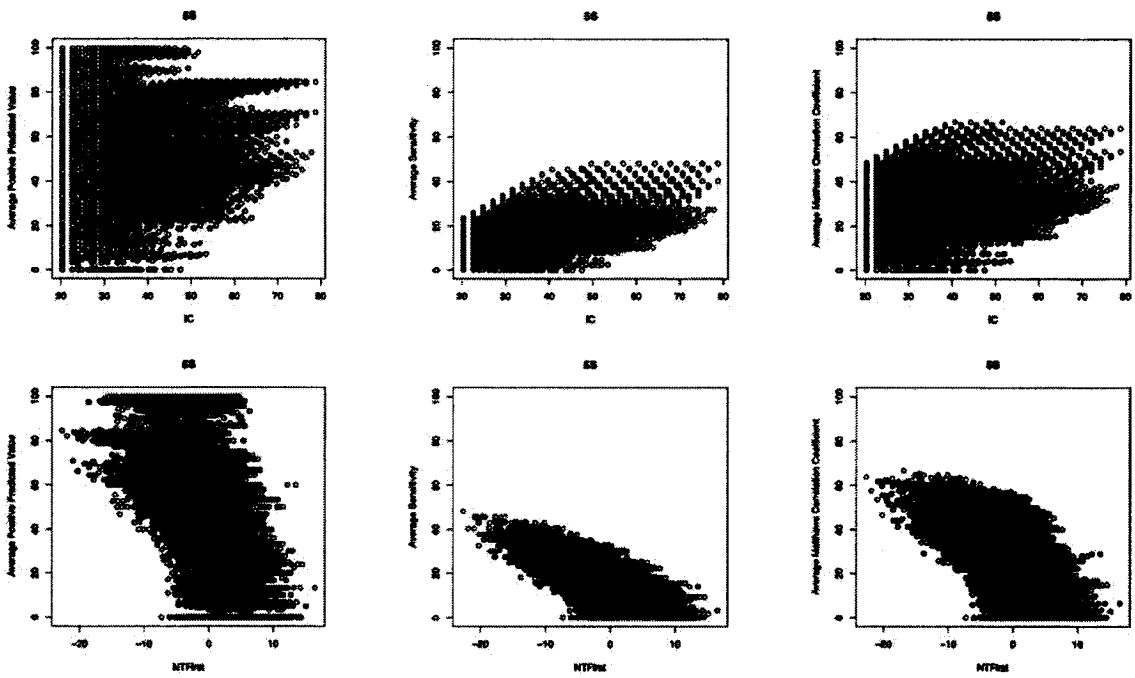

Figure 3.9: Performance diagrams for the 5S experiment. *IC* (top) *NTFirst* (bottom) scores against the PPV, sensitivity and Matthews correlation coefficient for the 5S data set.

Table 3.2: Correlation results. For each column, the numbers represent the correlation coefficient with the average Matthews correlation coefficient. Bold values correspond to the highest correlation with MCC.

Score	HSL3 01	HSL3 02	IRE 01	IRE 02	tRNA	5S	Average
TInfo	0.337	0.849	0.681	0.723	0.663	0.519	0.629
TAvg	-0.868	-0.964	-0.857	-0.626	-0.792	-0.680	-0.798
TBest	-0.385	-0.960	-0.869	-0.608	-0.829	-0.755	-0.734
TFirst	-0.465	-0.970	**-0.891**	-0.673	-0.827	-0.743	-0.762
TLeft	-0.529	-0.970	-0.869	-0.638	-0.831	-0.744	-0.764
TSum	-0.483	-0.950	-0.840	-0.656	-0.785	**-0.764**	-0.746
TWorst	-0.561	-0.973	-0.880	-0.661	-0.825	-0.735	-0.773
NTAvg	-0.450	-0.973	-0.836	-0.644	-0.797	-0.679	-0.730
NTBest	-0.879	-0.966	-0.857	-0.610	-0.830	-0.754	-0.816
NTFirst	-0.776	-0.974	-0.887	-0.682	-0.830	-0.743	-0.815
NTLeft	**-0.959**	-0.976	-0.860	-0.647	**-0.834**	-0.744	**-0.837**
NTSum	-0.865	-0.948	-0.841	-0.663	-0.784	-0.757	-0.810
NTWorst	-0.740	**-0.978**	-0.880	**-0.685**	-0.828	-0.734	-0.808

Table 3.3: Rank statistics. Ranking statistics $\hat{\rho}$ for the objective functions for all six experiments. MCC has been used as the response variable. Bold values indicate the highest value of ranking statistics.

Score	HSL3 01	HSL3 02	IRE 01	IRE 02	tRNA	5S	Average
TInfo	0.994	0.754	0.740	0.509	0.686	0.465	0.691
TAvg	**1.0**	0.918	0.939	0.702	0.844	0.668	0.845
TBest	0.959	0.923	**0.950**	0.712	0.873	0.748	0.861
TFirst	0.979	0.900	0.931	0.778	0.874	0.735	0.866
TLeft	0.959	0.918	0.887	0.729	0.877	0.736	0.851
TSum	0.962	0.864	0.898	0.743	0.857	0.766	0.848
TWorst	0.998	0.909	0.899	0.727	0.872	0.724	0.855
NTInfo	0.975	0.565	0.587	0.327	0.609	0.454	0.586
NTAvg	0.977	0.923	0.926	0.715	0.845	0.667	0.842
NTBest	**1.0**	**0.926**	0.939	0.715	0.873	**0.746**	0.867
NTFirst	0.999	0.903	0.924	**0.780**	0.874	0.734	**0.869**
NTLeft	**1.0**	0.922	0.884	0.751	**0.877**	0.735	0.862
NTSum	0.998	0.908	0.943	0.741	0.876	0.738	0.868
NTWorst	0.998	0.912	0.859	0.750	0.873	0.723	0.853

3.7 Summary

Thermodynamic study of RNA helps to provide a useful method to predict secondary structure for a given sequence. The number of possible structures for a given RNA sequence is large; exponential w.r.t its length [56]. Thus, thermodynamic study restricts the set of hypothesis by providing models with reasonable approximations. In this chapter we introduced several objective functions based on thermodynamic principles and information content. Their correlation to MCC and rank based evaluation were performed. For all experiments conducted, the diagrams, correlation coefficients and ranking statistics support the use of free energy for ranking consensus motifs. Top-ranked motifs generally correspond to high PPV/sensitivity motifs while bottom-ranked motifs correspond to low PPV/sensitivity motifs. This work also demonstrates that the motifs with positive free energy can be eliminated from the search space.

Chapter 4

Regression analysis

This chapter presents our work on combining the different scoring functions introduced in Chapter 3. We present here a general framework for analyzing RNA structures using statistical regression methods such that the motifs that are most similar to the true biological motif are ranked highest. By applying multiple scoring functions in combination, we expect to raise further the performance of scoring functions alone. This work has been accepted for publication at the Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada, 2006 [3].

4.1 Statistical models

We studied correlation coefficient to measure the usefulness of the objective scoring functions defined in the previous chapter. Strong correlation was observed between the normalized scores and PPV/sensitivity across the data sets used. The value of r ranged from -0.783 to -0.95. For some instances, it was even lower (better). However, no particular function outperformed the other; the highest correlation coefficient was different across each data set, see Table 3.2. Preliminary work had shown that the linear combination of free energy score and information content outperformed either of the two scores alone.

Since the free energy scores were particularly effective to separate high PPV/sensitivity motifs from the rest, we assume that variations in the scoring functions can be represented by a statistical model. We used the most classical field of statistical analysis — regression, to analyze a set of RNA secondary structure motifs. A statistical model can be derived using techniques from regression analysis to obtain a template or model that is able to discriminate native folds from the rest.

We built multiple regression models to show the relationship between the normalized scores and different performance measures (see Section 3.4.3). We used Matthews correlation coefficient, PPV and Distance (between two RNA structures) as the response for building the models. Several models were obtained from the normalized scores to find weights that produce the best classification performance to predict whether a given RNA is native.

Different length of motifs and number of base pairs occurring within will introduce variations in the thermodynamic energies. We normalized the variants by the length as well as the occurring number of base pairs observed in the motif. This allowed to put the scores on a common scale by removing the undue influence of the scores with large outlying numerical values. We selected a subset of the scoring functions that have a more biological interpretation, namely *TBest*, *TWorst* and *TSum*. While *TBest* and *TWorst* consider the best (lowest) and the worst (highest) free energy matches, *TSum* captures the influence of the remaining matches not participating in best or worst matches. In [12], the correlation between the normalized energy (energy per base pair) and %-GC was studied. It was observed that they are negatively correlated. This is expected due to the fact that G:C base pairs have lower free energy than the other base pairings. We include this distinguishing feature with the other defined functions.

Minitab [23] statistical software package was used to built the regression models. We describe below the normalized scores and the rational behind each to predict the response.

4.1.1 Energy contribution per match

To calculate the energy contributed per match for a given motif, we used the same energy functions defined in Chapter 3. We divide each score by the number of sequences taken under consideration, except for *TSum* where we divide it by the number of matches made. The resulting scores are noted with the letter N_{1x} followed by the name of the base score. The subscript x equals 1 indicates the normalized function used as an attribute alone and 2 indicates a linear combination of the normalized function and information content.

$$N_{11}\,TBest = \frac{TBest}{NoSeq} \qquad N_{12}\,TBest = \frac{TBest}{NoSeq} - TInfo$$

$$N_{11}\,TWorst = \frac{TWorst}{NoSeq} \qquad N_{12}\,TWorst = \frac{TWorst}{NoSeq} - TInfo$$

$$N_{11}\,TSum = \frac{TSum}{NoMatch} \qquad N_{12}\,TSum = \frac{TSum}{NoMatch} - TInfo$$

NoSeq represents the number of sequences in which the match occurred and *NoMatch* is the total number of matches made.

4.1.2 Energy contribution per nucleotide

Since the length of the motif is different across different data sets, it is expected to find variations in the thermodynamic energies. One way to remove this variation is to bring the energies onto a common level. This is possible by obtaining the energy contribution per nucleotide for a given motif. Normalized variants of the above defined functions are obtained by diving each score by the length of the motif. We normalize *TInfo* in the same fashion. With the linear combination of the normalized information content we get seven new attributes to be considered for model building.

$$N_{21}\,TBest = \frac{N_{11}\,TBest}{MLen} \qquad N_{22}\,TBest = \frac{N_{11}\,TBest}{MLen} - \frac{TInfo}{MLen}$$

$$N_{21}\,TWorst = \frac{N_{11}\,TWorst}{MLen} \qquad N_{22}\,TWorst = \frac{N_{11}\,TWorst}{MLen} - \frac{TInfo}{MLen}$$
$$N_{21}\,TSum = \frac{N_{11}\,TSum}{MLen} \qquad N_{22}\,TSum = \frac{N_{11}\,TSum}{MLen} - \frac{TInfo}{MLen}$$
$$N_{2}\,TInfo = \frac{TInfo}{MLen}$$

MLen is the length of the motif.

4.1.3 Energy contribution per base pair

As mentioned earlier, the number of base pairs in a stem contributes to the stability of a RNA structure. Hence we would like to differentiate between structures having similar energy but different number of base pairs. For this reason, we normalize the base scores by dividing them by the number of base pair found in the motif. Similarly, we also consider a linear combination of the normalized information content and the energies. The base scores $N_{31}\,TBest$, $N_{31}\,TWorst$, $N_{31}\,TSum$, $N_{32}\,TBest$, $N_{32}\,TWorst$, $N_{32}\,TSum$ are calculated as follows.

$$N_{31}\,TBest = \frac{N_{11}\,TBest}{NoBp} \qquad N_{32}\,TBest = \frac{N_{11}\,TBest}{NoBp} - \frac{TInfo}{NoBp}$$
$$N_{31}\,TWorst = \frac{N_{11}\,TWorst}{NoBp} \qquad N_{32}\,TWorst = \frac{N_{11}\,TWorst}{NoBp} - \frac{TInfo}{NoBp}$$
$$N_{31}\,TSum = \frac{N_{11}\,TSum}{NoBp} \qquad N_{32}\,TSum = \frac{N_{11}\,TSum}{NoBp} - \frac{TInfo}{NoBp}$$
$$N_{3}\,TInfo = \frac{TInfo}{NoBp}$$

where, *NoBp* is the number of base pair found in the motif.

Having obtained all the explanatory variables, we can now define the maximum model, that is, the model containing all explanatory variables that could possibly be present in the final model.

We define the following regression models for each set of normalized variants.

$$Y_{11} = \beta_0 + \beta_1 N_{11}\, TBest + \beta_2 N_{11}\, TWorst + \beta_3 N_{11}\, TSum + \beta_4\, TInfo + \beta_5 \%GC$$

$$Y_{12} = \beta_0 + \beta_1 N_{12}\, TBest + \beta_2 N_{12}\, TWorst + \beta_3 N_{12}\, TSum + \beta_4 \%GC$$

$$Y_{21} = \beta_0 + \beta_1 N_{21}\, TBest + \beta_2 N_{21}\, TWorst + \beta_3 N_{21}\, TSum + \beta_4 N_2\, TInfo + \beta_5 \%GC$$

$$Y_{22} = \beta_0 + \beta_1 N_{22}\, TBest + \beta_2 N_{22}\, TWorst + \beta_3 N_{22}\, TSum + \beta_4 \%GC$$

$$Y_{31} = \beta_0 + \beta_1 N_{31}\, TBest + \beta_2 N_{31}\, TWorst + \beta_3 N_{31}\, TSum + \beta_4 N_3\, TInfo + \beta_5 \%GC$$

$$Y_{32} = \beta_0 + \beta_1 N_{32}\, TBest + \beta_2 N_{32}\, TWorst + \beta_3 N_{32}\, TSum + \beta_4 \%GC$$

4.2 Building base models using selection criteria

When defining the maximum model, it is important to include all explanatory variables which might have an effect on the response variable. However, there are reasons as to why one has to be careful in selecting the explanatory variable. Including many explanatory variables may increase the risk of collinearity, that is, two or more variables being linearly dependent. One of the common purported remedy of collinearity is variable selection. Since the number of variables in our case is not large, we opted for a best subset regression. In this procedure, all possible regressions were computed for each subset of models containing one to the maximum number of variables. Subset models may actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors [20].

Once computed, we adopt a statistical selection criterion approach to determine which model is better than the rest. A selection criterion should be able to order all the possible models from best to worst. Although many different criteria have been suggested through time; some are better than others, but there is no single criterion which is overall preferred. We discuss here the different criteria used for the base model selection. The definitions are adapted from [28].

For a given data set, the above process is carried for all the maximum models obtained using

the different normalized scores. The resulting model that best fits the base data is defined as the base model of that particular data set.

4.2.1 The R^2 criterion

The coefficient of determination R^2 is a crude measure of how well a model accounts for the variation in the data. It is the proportion of the total amount of variation in the data that can be explained by the fitted model, that is,

$$R^2 = \frac{S_{yy} - RSS}{S_{yy}}$$

where,

$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the corrected sum of squares of the responses,

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares,

$\bar{y} = \frac{\sum y_i}{n}$, n is the number of data points,

y_i is the observed value of y and $\hat{y}_i$ is y calculated from the regression equation.

The closer the model fits the data, the larger the R^2 will be. This method, however, has a number of drawbacks. The most important being that, due to the way R^2 is defined, the largest model (the one with the most number of explanatory variable) will always have the largest R^2 whether the extra variables provide any important information about the response variable or not. A common way to avoid this problem is to use an adjusted version of R^2 instead of R^2 itself. The adjusted R^2 statistic, for a model with k explanatory variables and n data points is given by

$$R_a^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

where n is the number of data points. R_a^2 does not necessarily increase when the number

of explanatory variables increases. According to R_a^2 (and R^2) criteria, one should choose the model which has the largest R_a^2 (R^2, respectively).

4.2.2 The C_m criterion

A useful selection criteria uses the Mallow C_m statistic

$$C_m = \frac{RSS_m}{S^2} + 2(m+1) - n$$

where m is the number of explanatory variables in the reduced model, and $S^2 = RSS_k/(n-k-1)$ is the unbiased estimate of the error variance. Suppose that the reduced model is as good as (or better than) the maximum model. In that case, the error variance for the reduced model, $RSS_m/(n-m-1)$ is close to (or smaller than) the error variance for the maximum model $RSS_k/(n-k-1)$. The C_m statistic becomes

$$\begin{aligned} C_m &= (n-m-1)\frac{RSS_m/(n-m-1)}{RSS_k/(n-k-1)} + 2(m+1) - n \\ &\leq n - m - 1 + 2(m+1) - n \\ &= m + 1 \end{aligned}$$

Thus, if C_m is approximately equal to (or smaller than) $m+1$, the reduced model fits the data at least as good as the maximum model (in the sense that the error variation is smaller for the reduced model). The C_m criterion says to choose the reduced model with the smallest value of C_m.

4.3 Evaluation of base models

With the base model determined for the corresponding data set, we evaluate the model on the basis of its performance on other data sets. Along with the standard residue based approach to evaluate regression models, we also adopt a rank based evaluation [43]. We discuss here both the evaluation methods.

4.3.1 Residue based evaluation

The standard approach to evaluating regression model performance is through additive, residual based loss functions. These measures often represents the true cost of the prediction errors. One test procedure to evaluate the model is by calculating the standard error of estimate.

The standard error of the estimate is a measure of the accuracy of predictions made with a regression line. The regression line seeks to minimize the sum of the squared errors of prediction. The square root of the average squared error of prediction is called the standard error of estimate. The formula is given as

$$\sigma_{est} = \sqrt{\frac{\sum (y - y_i)^2}{n}},$$

where n is the number of pairs data points.

4.3.2 Rank based evaluation

While residue based evaluation takes into account the global performance, rank based evaluation provides insights about the local performance of a model. In addition, they provide a visualization method of comparing models and are robust to extreme outliers [43]. They also provide statistics that are similar to the commonly used correlation coefficients in data analysis (Spearman's ρ and Kendall's τ). We explain here the method of calculating the two statistics.

We assume a base model $\hat{y} = m(x)$ which attempts to predict a response $y \in R$ as a function of a feature vector x. Assume further we have a test set of size n ($\{x_i, y_i\}_{i=1}^n$) which is not used in the modeling, which we want to utilize to evaluate the performance of the model m in predicting y. Using a rank based measure, we evaluate the performance of the scoring model $m(x)$ in sorting the values of y from large to small or vice versa. Without loss of generality, we also assume that the test set is sorted in decreasing order of model scores, that is:

$$m(x_1) > m(x_2) > ... > m(x_n)$$

The responses of the test set are ranked in the decreasing order. Let s_i be the rank of observation i in this order:

$$s_i = \quad | \{j \leq n \mid y_i \leq y_j\} |$$

The number of ranking order switches is calculated as:

$$T = \sum_{i<j} \mathbf{1}\{s_i > s_j\} \tag{4.1}$$

and the weighted sum of order switches

$$R = \sum_{i<j} (j - i) \cdot \mathbf{1}\{s_i > s_j\} \tag{4.2}$$

The first measure simply counts how many of the pairs in the test data are ordered incorrectly by the model $m(x)$. The second weights these incorrect orderings by the difference in their model ranks thereby providing a magnitude of error being committed.

Finally, the two correlations, are obtained by the following equations.

$$\hat{\tau} = 1 - \frac{4T}{n(n-1)}, \; \hat{\rho} = 1 - \frac{12R}{n(n-1)(n+1)} \tag{4.3}$$

Values of $T = R = 0$ correspond to a perfect model performance by giving a value of 1 to $\hat{\tau}$ and $\hat{\rho}$. Similarly, -1 corresponds to making all possible errors.

In Seed, the construction step of motifs involves an instantiation subprocess, where all the base pairs of every stem motifs are replaced by the actual base pair that occurs in the seed sequence. Each instantiated motif is matched among the remaining sequences and the ones that posses a minimum support are reported. Clearly, the scoring function based on information content will rank partially or fully instantiated motifs higher than those containing only structural motifs. However, since the instantiation step occurs in a combinatorial manner, motifs occurring at the same location and having same measure of information content will be ranked same. This would

Table 4.1: Transformation of rank : **O_Rank** corresponds to the original rank and **T_Rank** is the transformed rank. The test set is sorted in decreasing order of model scores.

Index	**O_Rank** (s_i)	**T_Rank**	y	$\hat{y}$
7	1	1	3.0	3.5
6	1	2	3.0	3.5
5	6	4	0.0	2.4
4	6	5	0.0	2.4
3	3	3	2.0	2.2
1	6	6	0.0	0.2
2	6	7	0.0	0.2

result in sets of responses having the same rank. Having unique ranks makes the discussion and presentation of rank based method simpler.

The above discussion are based on the assumption that the model scores are unique. To assign unique ranks to each response, we re-rank them in decreasing order of their response values. This does not effect the above statistics as the ranking order switches remain the same. Table 4.1 shows an example explaining the rank transformation.

In Table 4.1, using the transformed rank (**T_Rank**), T receives a value of 2 for $s_i = \{4, 5\}$, and $s_j = \{3\}$. Similarly, R evaluates to 3. Using equation 4.3, the correlations equate to

$$\hat{\tau} = 0.809 \qquad \hat{\rho} = 0.892$$

4.3.3 Visualization of ranking performance

Visualization of ranking performance can often provide additional insights about model performance. We plot the percentage of correct ranked pairs as a function of the predicted rank. Starting from the largest model prediction $m(x_1)$, we calculate in decreasing order for each observation x_i the percentage of correctly ranked pairs (y_i, y_j) over all $j \neq i$.

The area above the curve is the sum of the incorrectly ranked pairs which is equal to $2 \cdot T/(n(n-$

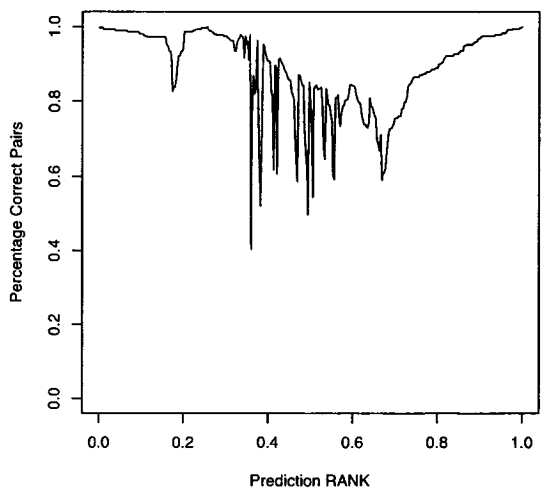

Figure 4.1: Percent of correctly ranked pairs in involving a particular prediction

1)) [43]. Therefore, the area under the curve equals $\tilde{\tau}$, where $\tilde{\tau}$ is the re-normalized version of $\hat{\tau}$ ($\tilde{\tau} = 1/2 + \hat{\tau}/2$). For a perfect model comprising of 100% correctly ranked pairs, we would see a constant performance for each prediction rank. However in general, a model is not perfect and makes predictions that are some times large or small. Even for a perfect prediction for a particular observation with $m(x_i) = y_i$ will have a number of inversely ranked pairs due to errors of the other predictions. Therefore, the upper limit of the performance for a given prediction is not 100% but determined by the model performance of predictions around it.

The performance of a particular region in the graph is therefore understood by two properties

- the distance of the upper envelope from the 100% line, and
- the distance of the actual performance from the upper envelope.

where the envelope (not shown in the figure) is the line joining the peaks of the curve. Figure 4.1 illustrates the ranking performance of Y_{5S}^{MCC} regression model (built from 5S data set using

MCC as response variable) on HSL3 data set. In this example, the model ranking of the top 30% cases on the left side of the graph appears to be very good whereas the ranking of the intermediate values is comparatively bad. Also, it shows large outliers between 70 to 50 percent correct pairs.

4.4 Experiments

We set out to learn regression models build of different scoring functions to rank favorable secondary structures from a set of motifs. For each of the four data set, several models were built on different explanatory variables discussed earlier. The model that best fits the data was identified as the base model for the corresponding data set. Furthermore, different response variables were used to study the relationship between the different functions and the performance measures. Residue and Rank based evaluation were carried to study the performance of the regression models. We compared the PPV and sensitivity of the top ranked motifs of all the data sets predicted by each model. In addition, the top 1, 2, 4 and 8 structurally different motifs were also compared.

4.5 Results

4.5.1 Generation of base models

As discussed earlier, the selection strategy for choosing the best model was due to best subset regression. In this, we identify the best-fitting models that can be constructed with the specified predictors variables.

The procedure examines all the possible subsets of the predictors, beginning with all the models containing one predictor, and then all models containing two predictors, and so on. Each subset differs from the preceding subset by the addition or deletion of only one variable. Below we

discuss the statistics of the top models for each data set consisting of all or some of explanatory variables.

In best subset regression, we look at the following statistics:

1. R_a^2 : We want the value of R_a^2 to be as close to 1 (100%) as possible without creating other problems of collinearity.
2. Mallows C_m : We want the value of C_m to be less than or equal to $m+1$ where m is the number of predictors.

We present below the results where first MCC is used as a response, recall that MCC represents a tradeoff between PPV and sensitivity. Following this, the results of PPV and Distance are illustrated. For HSL3 and IRE data sets, the experiments (HSL 02, IRE 02) that produced more number of motifs were used to built regression models as it gave more of a chance to capture of the variability of the different scores.

MCC as response

HSL3

The application of best subset regression on HSL3 data set yielded a variety of models. Table E.1 shows the possible regressions that were computed. For each model, the R^2, R_a^2 and Mallow's C_m values are reported.

Taking a closer look at the results of the different criteria, we observe that the model preferred by the R_a^2 criteria is the model containing $\%GC$, $N_2\,TInfo$, $N_{21}\,TBest$, $N_{21}\,TSum$ and $N_{21}\,TWorst$ ($R_a^2 = 97.8$). Also, the value of C_m is equal to 6.0 which is the lowest among the other models under the same predictors. Although the R^2 and R_a^2 values for other set of predictors are very similar, we choose the model with the highest R^2 value as the appropriate model for HSL3 data set.

Therefore, the least square model is given by

$$Y_{HSL3}^{MCC} = 0.09 + 203\%GC + 12.1N_2\,TInfo - 125N_{21}\,TBest - 163N_{21}\,TWorst + 230N_{21}\,TSum$$

IRE

Table E.2 shows the possible regressions that were computed for IRE data set. For each model, the R^2, R_a^2 and Mallow's C_m values are reported. The R_a^2 criterion suggests the maximum model containing $\%GC$, $NTInfo$, $N_{11}\,TBest$, $N_{11}\,TSum$, $N_{11}\,TWorst$ ($R_a^2 = 67.5\%$), while the C_m criterion suggests the model containing only $N_{22}\,TBest$ and $N_{22}\,TSum$ ($C_m = 1.9$). As always, R^2 criterion suggests the maximum model for all the predictors.

Although, the model preferred by R_a^2 is not the best according to the C_m criterion, it is still a good model. Conversely, the model suggested by C_m, has a lesser value or R_a^2 (65.4%), so, according to this criteria, it is a reasonably good model. Thus it seems that both of the models in question are good according to all the three criteria, so we can choose either one. If simplicity of the model is essential, the simple model

$$Y = \beta_0 + \beta_1 N_{32}\,TBest + \beta_2 N_{32}\,TSum$$

might be the preferred choice, but since the prediction is the main purpose of the analysis, the larger model

$$Y = \beta_0 + \beta_1 \%GC + \beta_2 NTInfo + \beta_3 N_{11}\,TBest + \beta_4 N_{11}\,TWorst + \beta_5 N_{11}\,TSum$$

may be better. The least squares models is given by,

$$Y_{IRE}^{MCC} = -110 + 308\%GC + 8.23NTInfo + 0.93N_{11}\,TBest - 0.71N_{11}\,TWorst - 3.42N_{11}\,TSum$$

tRNA

Table E.3 shows the regressions that were computed for tRNA data set. The value of R_a^2 ranges from the low value of 55.7 to the high value of 72.2 across the set of predictors. The model suggested by R_a^2 is the maximum model containing the following independent variables : $\%GC$, $TInfo$, $N_{11}\,TBest$, $N_{11}\,TSum$, $N_{11}\,TWorst$. Conversely, the C_m criterion prefers the model containing $\%GC$ and , $N_{32}\,TWorst$ ($C_m = 1.3$). The model suggested by C_m has a lower R_a^2 value ($R_a^2 = 61.2$). As a result, the former stands out to be the most appropriate model as the C_m value is within the limits ($C_m = 6.0$) and has the highest value of R_a^2. Also, the model having the second lowest value of C_m has a slightly lower R_a^2.

Thus, the least squares line for the final models is given by,

$$Y_{tRNA}^{MCC} = 25.5 - 74.9\%GC + 0.809\,TInfo - 1.28N_{11}\,TBest + 1.33N_{11}\,TWorst - 2.62N_{11}\,TSum$$

5S

The final data set used for model building is 5S data set. Similarly to tRNA data set, the model preferred by R_a^2 consists of the following predictors : $\%GC$, $TInfo$, $N_{11}\,TBest$, $N_{11}\,TSum$, $N_{11}\,TWorst$ ($R_a^2 = 59.9$). The models suggested by C_m ($C_m = 5.0$), has lesser values of R_a^2 (51.6, 55.2 and 54.1). For the remaining subset, the value of C_m is off limits, that is, they have very large values. Therefore the appropriate model for prediction appears to be the maximum model. The details of the possible regression models are given in Tables E.4.

The least squares line for the final models is given by,

$$Y_{5S}^{MCC} = 14.3 + 22.1\%GC + 0.212\,TInfo - 2.66N_{11}\,TBest + 2.74N_{11}\,TWorst - 1.45N_{11}\,TSum$$

PPV as response

Similar analyses were carried out for building base models having PPV and Distance as response. The details of R_a^2 and C_m can be observed from the Tables E.5–E.8. The following models were obtained

$$Y_{HSL3}^{PPV} = -2.87 - 265\%GC + 11.4N_2\,TInfo - 135N_{21}\,TBest - 194N_{21}\,TWorst + 288N_{21}\,TSum$$

$$Y_{IRE}^{PPV} = 21.0 + 134N_{22}\,TBest - 257N_{22}\,TSum$$

$$Y_{tRNA}^{PPV} = 44.0 - 54.5\%GC - 7.6N_3\,TInfo - 12.1N_{31}\,TBest - 27.2N_{31}\,TSum$$

$$Y_{5S}^{PPV} = 29.4 + 34.8\%GC + 3.49N_3\,TInfo - 46.0N_{21}\,TBest + 54.2N_{21}\,TWorst - 40.3N_{21}\,TSum$$

Distance as response

In case of Distance, the following models were obtained. Refer Tables E.9–E.12 for the statistics.

$$Y_{HSL3}^{D} = 83.7 - 434\%GC - 14.8N_2\,TInfo + 111N_{21}\,TBest - 69.6N_{21}\,TWorst - 146N_{21}\,TSum$$

$$Y_{IRE}^{D} = 79.1 - 192\%GC - 60.4N_2\,TInfo + 78.3N_{21}\,TBest - 48.8N_{21}\,TWorst - 88.5N_{21}\,TSum$$

$$Y_{tRNA}^{D} = 65.5 - 88.4\%GC - 40.3N_2\,TInfo + 45.6N_{21}\,TBest + 28.8N_{21}\,TWorst - 40.2N_{21}\,TSum$$

$$Y_{5S}^{D} = 136 - 224\%GC - 39.8N_2\,TInfo + 76.4N_{21}\,TBest - 19.5N_{21}\,TWorst - 17.9N_{21}\,TSum$$

4.5.2 Residual based evaluation

The standard error of the estimate is a measure of the accuracy of predictions made with a regression line. The smallest value of σ_{est} is zero, which occurs when all the points fall on the regression line. Thus, when σ_{est} is small, the fit is excellent, and the linear model is likely to be an effective analytical and prediction tool.

As expect, the best value of σ_{est} can be seen across the diagonal of the Tables 4.2–4.4 as it represents the performance of the regression model on the base data set. For example, when

response variable is MCC, the value of σ_{est} is 6.46, 8.32, 10.78 and 8.24 of the four models on the base data. Also, the smallest values of σ_{est} is seen when Distance is used as the response variable, see Table 4.4.

Table 4.2: Standard error of estimate : Response is AVGMCC

	HSL3	**IRE**	**tRNA**	**5S**
Y_{HSL3}	6.46	14.33	25.21	13.22
Y_{IRE}	33.08	8.32	48.12	31.21
Y_{tRNA}	35.87	25.87	10.78	34.72
Y_{5S}	39.61	24.10	24.37	8.24

Table 4.3: Standard error of estimate : Response is AVGPPV

	HSL3	**IRE**	**tRNA**	**5S**
Y_{HSL3}	7.50	16.88	40.62	32.73
Y_{IRE}	53.14	9.93	62.46	41.66
Y_{tRNA}	24.72	34.92	15.36	23.19
Y_{5S}	38.27	59.54	20.70	14.33

Table 4.4: Standard error of estimate : Response is DISTANCE

	HSL3	**IRE**	**tRNA**	**5S**
Y_{HSL3}	5.20	22.59	10.96	56.95
Y_{IRE}	21.44	3.73	12.24	46.72
Y_{tRNA}	23.74	8.69	4.66	58.05
Y_{5S}	60.10	59.81	55.69	6.00

The performance of the model can be judged by comparing the values of σ_{est} obtained across the different data sets. For example, the smallest value of σ_{est} is 3.73 for base model Y_{IRE} when response variable is Distance. For this model, the value of σ_{est} on tRNA data set is 12.24. We can see that the standard error of estimate is not very small (close to its base value, 3.73). On the other hand, it is not a large number. Because there is no predefined upper limit on σ_{est}, it is difficult to assess the model in this way. In case of Y_{tRNA} for the same response variable, we

see that the model shows similar performance on IRE data set. However, the error is large for tRNA and 5S data set. Similarly, for the other models the standard error of estimate is small for its base data set but large for the remaining data sets.

Although, σ_{est} is useful in comparing model, the values are not interpretable. It does appear to be a better performance metric in our case since we want the regression model to be able to separate the motifs which are likely to have high response that the others.

4.5.3 Rank based evaluation

Tables 4.5–4.7 list the values obtained for Spearman and Kendalls statistics. In common with other measures of correlation, Kendall's τ and Spearman's ρ will take values between -1 and +1, with a positive correlation indicating that the ranks of both variables increase together whilst a negative correlation indicates that as the rank of one variable increases the other one decreases.

In each table, the performance of every base model is shown on the other data sets used and on itself. Overall, the ranking statistics show significant results for all the response variables. The highest average correlation is obtained for Y_{tRNA} when Distance is used as the dependent variable.

When the response variable is *AVGMCC*, the average performance of the models is good with $\hat{\tau}$ ranging from 0.599 to 0.660 and $\hat{\rho}$ ranging from 0.739 to 0.801. Y_{tRNA} is the regression model with the highest average statistics. Except for the regression model Y_{IRE}, all the models show high correlation ($\hat{\tau}$) on their base data, see Table 4.5. Y_{HSL3} performs better for IRE data set, the weighted ranking ($\hat{\rho}$) of Y_{HSL3} is 0.719, compared to 0.706 for Y_{IRE}. It is interesting to observe that models built on complex structures, Y_{tRNA} and Y_{5S} show comparable correlations on HSL3 and IRE data set. However, the performance of Y_{HSL3} and Y_{IRE} decreases as the complexity of the motifs sought increases; with Y_{IRE} showing a slighter better performance than Y_{HSL3}.

Table 4.5: Kendall's and Spearman's correlation : AVGMCC

	HSL3		IRE		tRNA		5S		Average	
	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$
Y_{HSL3}	0.802	0.907	0.612	0.720	0.607	0.794	0.381	0.536	0.601	0.739
Y_{IRE}	0.761	0.886	0.614	0.706	0.608	0.799	0.411	0.572	0.599	0.741
Y_{tRNA}	0.793	0.905	0.577	0.676	0.718	0.887	0.551	0.737	0.660	0.801
Y_{5S}	0.779	0.889	0.512	0.596	0.685	0.863	0.576	0.761	0.638	0.777

The average performance of the models is lower when $AVGPPV$ is used as the response variable with $\hat{\tau}$ ranging from 0.521 to 0.647 and $\hat{\rho}$ ranging from 0.643 to 0.791. The weighted ranking ($\hat{\rho}$) of Y_{tRNA} is highest for HSL3, IRE and tRNA data sets, see Table 4.6. Y_{tRNA} is the regression model with the highest average statistics. The ranking statistics of Y_{HSL3} and Y_{IRE} is low for 5S, with $\hat{\tau} = 0.235$, $\hat{\rho} = 0.321$ for Y_{HSL3} and $\hat{\tau} = 0.246$, $\hat{\rho} = 0.339$ for Y_{IRE}.

We also performed several analyses to examine the relationship between Distance and scoring functions. Table 4.7 shows the Kendall's and Spearman's correlation. Y_{tRNA} gave the highest average correlation with $\hat{\tau} = 0.756$ and $\hat{\rho} = 0.904$. A similar trend is seen for this class of regression models; significant statistics are seen in the lower triangle of the table. The ability of the models built on complex structures (tRNA and 5S) to rank the motifs is good, with $\hat{\rho}$ ranging from 0.880 to 0.964. The performance of Y_{HSL3} is lowest for tRNA data set ($\hat{\tau} = 0.202$, $\hat{\rho} = 0.277$).

Ranking based evaluation measures the success of regression models in ranking the test set observation in the correct order. We explore the performance of the defined models by comparing the top ranked structurally different motifs.

4.5.4 Model performance

We present here the results of the top motifs predicted by the different regression models. Tables 4.8, 4.10 and 4.12 show the sensitivity, PPV, MCC and Distance of top motifs predicted

Table 4.6: Kendall's and Spearman's correlation : AVGPPV

	HSL3		IRE		tRNA		5S		Average	
	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$
Y_{HSL3}	0.792	0.895	0.608	0.709	0.461	0.647	0.235	0.321	0.524	0.643
Y_{IRE}	0.751	0.866	0.604	0.686	0.484	0.670	0.246	0.339	0.521	0.640
Y_{tRNA}	0.818	0.919	0.638	0.750	0.656	0.842	0.476	0.653	0.647	0.791
Y_{5S}	0.793	0.892	0.527	0.609	0.618	0.813	0.507	0.683	0.611	0.749

Table 4.7: Kendall's and Spearman's correlation : DISTANCE

	HSL3		IRE		tRNA		5S		Average	
	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$
Y_{HSL3}	0.824	0.948	0.750	0.916	0.202	0.277	0.498	0.686	0.569	0.707
Y_{IRE}	0.890	0.975	0.875	0.976	0.375	0.506	0.575	0.772	0.679	0.807
Y_{tRNA}	0.850	0.961	0.770	0.928	0.698	0.851	0.705	0.874	0.756	0.904
Y_{5S}	0.859	0.964	0.762	0.924	0.679	0.828	0.716	0.880	0.754	0.899

by the different models. Following each table, the structure and sequence of the corresponding motifs are shown (Table 4.9, 4.11 and 4.13).

In each table, both the observed and the predicted performance measure of motif is shown. For example, in Table 4.8, the top ranked motif picked by Y_{HSL3} on HSL3 data set is 261; represented under the column **motif**. The observed *AvgSensitivity*, *AvgPPV* and *DIST* are 100%, 100% and 0. Since *AvgMCC* is used as the response variable we only report its predicted value, which is 107.18%. Similarly, the performance measures for other models are shown. Finally, we normalized distance by information content and length (see Section 4.7). The results are shown in Tables 4.14 and 4.15.

We also compared the top 1, 2, 4 and 8 structurally different motifs to evaluate the performance of the models. Tables 4.16, 4.17 and 4.18 shows the maximum, minimum and average sensitivity and PPV measures of the top 1–8 structurally different motifs.

	Data set	Motif	Observed				Predicted
			AVGSEN	AVGPPV	AVGMCC	DIST	AVGMCC
Y_{HSL3}	HSL3	261	100	100	100	0	107.18
	IRE	85	92.7	100	96.28	4	45.48
	tRNA	569	73.7	88.2	80.62	12	57.39
	5S	38708	18.1	41.2	27.3	118	101.624
Y_{IRE}	HSL3	261	100	100	100	0	287.28
	IRE	85	92.7	100	96.28	4	119.67
	tRNA	569	73.7	88.2	80.62	12	432.06
	5S	39929	30.1	72.5	46.71	74	19.64
Y_{tRNA}	HSL3	261	100	100	100	0	55.87
	IRE	85	92.7	100	96.28	4	42.17
	tRNA	4984	71.2	96	82.67	24	40.20
	5S	39929	30.1	72.5	46.71	74	149.98
Y_{5S}	HSL3	356	83.3	100	91.26	4	40.20
	IRE	109	58.9	100	76.74	11	31.71
	tRNA	4968	76.2	100	87.29	20	54.71
	5S	39929	30.1	72.5	46.71	74	80.43

Table 4.8: Characteristics of top motifs picked by models having **AVGMCC** as response showing the observed and predicted performance measures. **DIST** is the edit distance of the predicted structure to the reference structure.

	Structure/Sequence
Y_{HSL3}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......))))).....(((...........)))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
	(((....(((.(((........)))........................)))..............)))(((.......)))........(((....))) NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGNNNNNNNNNCNNNNNNNNNNNNNNNNN
Y_{IRE}	(((((((....))))))) / GGCTCTNNNNAGAGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......))))).....(((...........)))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
	(((.....(((.((((........(((.............))).........))))....)))...........(((.......)))...............))) / NNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNNCNNNNNNNNNNNNNNNNNNN
Y_{tRNA}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((((..(((...........))).(((((.......)))))....................))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNNGGNNGNNNNNNNNNNNNNNNNNNNNNNNNNNN
	(((.....(((.((((........(((.............))).........))))....)))...........(((.......)))...............))) / NNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNNCNNNNNNNNNNNNNNNNNNN
Y_{5S}	(((((......))))) / NNNNNNNNNNNNNNNN
	(((((......))))) / NNNNNNNNNNNNNNNN
	(((((((.((((.........)))).(((((.......)))))....................))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNN
	(((.....(((.((((........(((.............)))).........))))....)))...........(((.......)))...............))) / NNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNNCNNNNNNNNNNNNNNNNNNN

Table 4.9: Characteristics of top motifs picked by models having **AVGMCC** as response showing the structure and sequence of the motifs as generated by Seed.

	Data set	Motif	Observed				Predicted
			AVGSEN	AVGPPV	AVGMCC	DIST	AVGPPV
Y_{HSL3}	*HSL3*	269	100	100	100	0	110.67
	IRE	1	0	0	0	14	49.73
	tRNA	269	68.7	87.5	77.53	14	61.39
	5S	38708	18.1	41.2	27.3	118	125.29
Y_{IRE}	*HSL3*	261	100	100	100	0	19.35
	IRE	85	92.7	100	96.28	4	2.8
	tRNA	569	73.7	88.2	90	12	35.13
	5S	124281	48.2	84.6	63.85	70	55.08
Y_{tRNA}	*HSL3*	325	83.3	100	91.26	4	134.77
	IRE	107	45.5	78.6	59.80	11	81
	tRNA	4984	71.2	96	82.67	24	109.9
	5S	39929	30.1	72.5	46.71	74	126.17
Y_{5S}	*HSL3*	356	83.3	100	91.26	4	117.36
	IRE	109	58.9	100	76.74	11	90.71
	tRNA	5080	66.3	95.7	79.65	26	87.37
	5S	39929	30.1	72.5	46.71	74	127.55

Table 4.10: Characteristics of top motifs picked by models having **AVGPPV** as response showing the observed and predicted performance measures. **DIST** is the edit distance of the predicted structure to the reference structure.

	Structure/Sequence
Y_{HSL3}	((((((....)))))) / GGCNCTNNNNAGNGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......))))).....(((...........)))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
	(((....(((.(((........))).......................))).............)))(((.......))).......(((....))) NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNN
Y_{IRE}	(((((((....))))))) / GGCTCTNNNNAGAGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......))))).....(((...........)))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
	((((((((.....((((((((........(((.............)))........)))))...))).........(((....))).................)))))))) / NNNNGNNNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNNNNNNCNNNN
Y_{tRNA}	(((((......)))))/GGCTCNNNNNNGAGCC
	(((.(((((......))))))))/NNNNNNCNNNNNNNNNNGNNNNN
	(((((((..(((...........))).(((((.......)))))...................))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNNNNNNNNNNNNNNNNNNNNN
	(((.....(((.((((........(((.............)))........))))....)))...........(((.......)))...............))) / NNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNCNNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNN
Y_{5S}	(((((......))))) / NNNNNNNNNNNNNNNN
	(((((......))))) / NNNNNNNNNNNNNNNN
	(((((((.((((.........)))).(((((.......)))))...................))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
	(((.....(((.((((........(((.............)))........))))....)))...........(((.......)))...............))) / NNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNCNNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNN

Table 4.11: Characteristics of top motifs picked by models having **AVGPPV** as response showing the structure and sequence of the motifs as generated by Seed.

	Data set	Motif	Observed				Predicted
			AVGSEN	AVGPPV	AVGMCC	DIST	DIST
Y_{HSL3}	*HSL3*	260	83.3	100	91.26	2	-17.48
	IRE	1	0	0	0	14	2.24
	tRNA	1277	5.7	13.3	8.7	43	-4.73
	5S	299344	0	0	0	83	-19.26
Y_{IRE}	*HSL3*	261	100	100	100	0	-70.8
	IRE	85	92.7	100	96.3	4	14.65
	tRNA	269	68.7	87.5	77.5	14	22.81
	5S	291949	0	0	0	84	18.27
Y_{tRNA}	*HSL3*	261	100	100	100	0	-63.67
	IRE	85	92.7	100	96.3	4	10.88
	tRNA	569	73.7	88.2	80.62	12	11.26
	5S	291476	3.2	13.3	6.52	89	10.33
Y_{5S}	*HSL3*	261	100	100	100	0	-28.29
	IRE	85	92.7	100	96.3	4	62.8
	tRNA	269	68.7	87.5	77.5	14	56.50
	5S	300665	0	0	0	82	49.29

Table 4.12: Characteristics of top motifs picked by models having **DISTANCE** as response showing the observed and predicted performance measures. Predicted **DIST** is the edit distance of the predicted structure to the reference structure.

	Structure/Sequence
Y_{HSL3}	(((((....))))) / NNNNNNNNNNNNNN
	((((((......)))).....))) / NNNNNNNNNNNNNNNNNNNNNN
	((((((......)))....(((...........))).))) / NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
	(((.(((......)))).....)))..(((....))) / NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Y_{IRE}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......))))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	((((..((((.........)))).(((((.......))))).....(((...........))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNN
	(((.(((.......)))....))).(((....))) / NCNNCNNNNNNNNNNNGNNNNNGNNNNNNNNNNNN
Y_{tRNA}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......))))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......))))).....(((...........))))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNN
	(((...(((.(((....)))))))....))) / NNNNNNNNNNTNNNNNNNNANNNNNNNNNN
Y_{5S}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......))))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	((((..((((.........)))).(((((.......))))).....(((...........))))))) / NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNN
	(((((((((......)))....))))).(((....))) / NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Table 4.13: Characteristics of top motifs picked by models having **DISTANCE** as response showing the structure and sequence of the motifs as generated by Seed.

	Data set	Motif	Observed AVGSEN	AVGPPV	AVGMCC	NDIST	Predicted NDIST
Y_{HSL3}	*HSL3*	261	100	100	100	0	-0.11
	IRE	1	0	0	0	0.61	0.411
	tRNA	269	68.7	87.5	77.5	0.21	0.42
	5S	38708	18.1	41.2	27.3	1.17	0.07
Y_{IRE}	*HSL3*	261	100	100	100	0	-5.89
	IRE	81	81.1	100	90.1	0.35	-0.14
	tRNA	269	68.7	87.5	77.5	0.27	0.51
	5S	300665	0	0	0	3.05	-0.53
Y_{tRNA}	*HSL3*	261	100	100	100	0	0.40
	IRE	85	92.7	100	96.3	0.18	1.67
	tRNA	569	73.7	88.2	80.62	0.22	-0.43
	5S	124281	48.2	84.6	63.85	0.89	-1.82
Y_{5S}	*HSL3*	269	100	100	100	0	0.5
	IRE	86	93.3	100	96.6	0.5	7.62
	tRNA	269	68.7	87.5	77.5	0.88	1.92
	5S	124281	48.2	84.6	63.85	3.18	0.75

Table 4.14: Characteristics of top motifs picked by models having **DISTANCE** (normalized) as response showing the observed and predicted performance measures. Predicted **NDIST** is the edit distance of the predicted structure to the reference structure.

	Structure/Sequence
Y_{HSL3}	((((((....)))))) / GGCTCTNNNNAGAGCC
	((((((......)))).....))) / NNNNNNNNNNNNNNNNNNNNN
	((((..((((.........)))).(((((.......)))))).....(((...........))))))) / NNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNCNNNNNN
	(((((((((......)))....)))))).(((....))) / NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Y_{IRE}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((..((((......)))).))) / NNNNNNCNNNNNNNNNNGNNNNN
	((((..((((.........)))).(((((.......)))))).....(((...........))))))) / NNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNCNNNNNN
	(((((((((......)))....)))))).(((....))) / NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Y_{tRNA}	((((((....)))))) / GGCTCTNNNNAGAGCC
	(((.(((((......)))))))) / NNNNNNCNNNNNNNNNNGNNNNN
	(((((..((((.........)))).(((((.......)))))).....(((...........))))))))) / NNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNCNNNNNN
	(((((((((.....(((((((((........(((.............)))).........)))))...))).............(((....))).................)))))))) / NNNNGNNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNNNCNNNN
Y_{5S}	((((((....)))))) / GGCNCTNNNNAGNGCC
	(((.(((((......)))))))) / NNNNNNNNNNNNNNNNNNNNN
	((((..((((.........)))).(((((.......)))))).....(((...........))))))) / NNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNCNNNNNN
	(((((((((.....(((((((((........(((.............)))).........)))))...))).............(((....))).................)))))))) / NNNNGNNNNNNNNGCGNNGNGNNNNNNNNNGNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNNGNNNNNNNNCNNNNNNNNNNNNNNNNNNCNNNN

Table 4.15: Characteristics of top motifs picked by models having **DISTANCE** (normalized) as response showing the structure and sequence of the motifs as generated by Seed.

	Model	1 Structure		2 Structures		4 Structures		8 Structures	
		MAXSEN	MAXPPV	MAXSEN	MAXPPV	MAXSEN	MAXPPV	MAXSEN	MAXPPV
MCC	Y_{HSL3}	**69.50**	**88.23**	**72.60**	**94.13**	72.60	**94.13**	**72.60**	**94.13**
	Y_{IRE}	**77.0**	**95.47**	**77.0**	**95.47**	**77.0**	**95.47**	**77.0**	**95.47**
	Y_{tRNA}	76.67	88.90	83.33	95.47	83.33	95.47	83.33	95.47
	Y_{5S}	74.0	**100**	92.07	**100**	**93.67**	100	**93.67**	100
PPV	Y_{HSL3}	34.57	54.90	**72.60**	**94.13**	72.60	**94.13**	**72.60**	**94.13**
	Y_{IRE}	70.33	88.90	71.17	90.90	71.17	90.90	71.17	93.73
	Y_{tRNA}	58.60	88.90	77.50	89.47	79.83	94.43	80.70	100
	Y_{5S}	70.83	100	72.40	100	90.47	**100**	90.47	**100**
NDIST	Y_{HSL3}	34.57	54.90	55.40	88.23	**67.90**	88.23	67.90	88.23
	Y_{IRE}	58.73	66.67	61.37	77.77	61.37	77.77	63.13	85.20
	Y_{tRNA}	**83.33**	**95.47**	**83.33**	**95.47**	**83.33**	**95.47**	**83.33**	**95.47**
	Y_{5S}	**93.67**	100	**93.67**	100	93.67	100	93.67	100

Table 4.16: Characteristics of top 1–8 structurally distinct motifs picked by models showing the maximum sensitivity and positive predict value. Bold values indicate the highest measure in its class.

	Model	1 Structure		2 Structures		4 Structures		8 Structures	
		MINSEN	MINPPV	MINSEN	MINPPV	MINSEN	MINPPV	MINSEN	MINPPV
MCC	Y_{HSL3}	**51.40**	**64.70**	**22.77**	**30.50**	9.80	**13.90**	**9.80**	**13.90**
	Y_{IRE}	**67.83**	**78.70**	**57.67**	**73.27**	**29.90**	**39.93**	**29.03**	**39.37**
	Y_{tRNA}	65.27	77.77	56.37	77.20	5.27	10.53	5.27	10.53
	Y_{5S}	69.83	**100**	67.97	**93.33**	**19.60**	19.60	**17.63**	18.73
PPV	Y_{HSL3}	22.77	30.50	**22.77**	**30.50**	9.80	**13.90**	**9.80**	**13.90**
	Y_{IRE}	58.20	64.03	52.63	63.47	17.63	18.73	15.70	17.77
	Y_{tRNA}	33.03	44.43	33.03	43.87	5.27	10.0	5.27	10.0
	Y_{5S}	66.0	92.87	49.33	59.53	11.10	**23.33**	11.10	**23.87**
NDIST	Y_{HSL3}	22.77	30.50	15.70	18.73	**15.70**	17.77	0.0	0.0
	Y_{IRE}	50.97	52.07	43.47	52.07	0.0	0.0	0.0	0.0
	Y_{tRNA}	**74.90**	**92.43**	**63.37**	**87.87**	**12.27**	**21.20**	**11.40**	**20.63**
	Y_{5S}	**79.60**	86.27	**68.73**	85.40	17.63	18.73	15.70	17.77

Table 4.17: Characteristics of top 1–8 structurally distinct motifs picked by models showing the minimum sensitivity and positive predict value. Bold values indicate the highest measure in its class.

	Model	1 Structure		2 Structures		4 Structures		8 Structures	
		AVGSEN	AVGPPV	AVGSEN	AVGPPV	AVGSEN	AVGPPV	AVGSEN	AVGPPV
MCC	Y_{HSL3}	**60.45**	**76.47**	**47.68**	**62.32**	41.20	**54.02**	**41.20**	**54.02**
	Y_{IRE}	**72.42**	**87.08**	**67.33**	**84.37**	**53.45**	**67.7**	**53.02**	**67.42**
	Y_{tRNA}	70.97	83.33	69.85	86.33	44.30	53	44.30	53
	Y_{5S}	71.92	**100**	80.02	**96.67**	**56.63**	59.80	**55.65**	59.37
PPV	Y_{HSL3}	28.67	42.70	**47.68**	**62.32**	41.20	**54.02**	**41.20**	**54.02**
	Y_{IRE}	64.27	76.47	61.90	77.18	44.4	54.82	43.43	55.75
	Y_{tRNA}	45.82	66.67	55.27	66.67	42.55	52.22	42.98	55
	Y_{5S}	68.42	96.43	60.87	79.77	50.78	**61.67**	50.78	**61.53**
NDIST	Y_{HSL3}	28.67	42.70	35.55	53.48	**41.80**	53	33.95	44.12
	Y_{IRE}	54.85	59.37	52.42	64.92	30.68	38.88	31.57	42.60
	Y_{tRNA}	**79.12**	**93.95**	**73.35**	**91.67**	**47.8**	**58.33**	**47.37**	**58.05**
	Y_{5S}	**86.63**	93.13	**81.20**	92.70	55.65	59.37	54.68	58.88

Table 4.18: Characteristics of top 1–8 structurally distinct motifs picked by models showing the average sensitivity and positive predict value. Bold values indicate the highest measure in its class.

4.6 Discussion of results

In general, regression based models were effective to associate the variation of scores obtained from different functions and different performance measures. Results indicate that the performance of models based on PPV and MCC were better than Distance. We found that for our case study, rank based evaluation was more appropriate evaluation metric than residue based evaluation. We discuss here the implication of the results in detail.

4.6.1 Using thermodynamic based regression models to identify native folds

For the first data set, HSL3, we can see that the sensitivity/PPV of the predicted structures by different models is often 100%. The Y_{HSL3}^{MCC} model (regression model built on HSL3 data set) performed well on IRE, tRNA and itself. For IRE data set, the top ranked motif has a sensitivity of 92.7% and a PPV of 100%. There is a slight drop in the performance on tRNA data set resulting in a sensitivity of 73.7% and PPV of 88.2% . The model shows poor performance on 5S data set with sensitivity of 18.1% and PPV of 41.2%. This is due to the fact that the number of examples used in making the model is sufficiently not enough to identify the variation of scores in the 5S data set. Moreover, the normalized scores for this particular data set are not effective to capture the variation.

The performance of Y_{HSL3}^{PPV} shows similar results as Y_{HSL3}^{MCC} except for IRE data set where the top motif has MCC of 0%. This is because the predicted and true structure are at different locations in the sequence. If brought to the same offset, the sensitivity/PPV increases to 44.4/57.1%. The visual representation of this error can be seen in Figure D.5. This results in a slight drop in the value of percentage correct pairs (0.991) for the top ranked motifs. The results for Y_{HSL}^{DIST} on all data sets shown in table 4.12 are less promising. It performs poorly in recognizing motifs outside the training data.

Y_{IRE}^{MCC} performs well on all the data sets. The top motif picked from 5S has an PPV and

sensitivity of 30.1% and 72.5% respectively (Table 4.8). For the remaining test cases, this model has the same performance as Y_{HSL3}^{MCC}. We see an increase in the performance on identifying motifs on 5S data (see Table 4.10) when PPV is used as the response variable. For the remaining data sets, there is no further increase in performance measure. Unlike Y_{HSL3}^{DIST}, Y_{IRE}^{DIST} shows good performance on HSL3, IRE and tRNA data set. The performance of 5S is similarly low.

To study the performance of models built on higher complexity data set, we built models on 7 tRNA and 7 5S sequences. We can see that Y_{tRNA}^{MCC} performs well on all the data sets (Table 4.8). Average PPV of the top tRNA motif increased to 96% as compared to 88.2% ranked by Y_{tRNA}^{MCC}. However, the sensitivity drops to 71.2%. Excluding the results on the base data set, Y_{tRNA}^{MCC} has the highest average sensitivity (74.26%) over the other models. The performance of Y_{tRNA}^{PPV} and Y_{tRNA}^{MCC} are similar for tRNA and 5S data. However, the performance on HSL3 and IRE data set is low. Models built on Distance performed equally well except for 5S where it showed a low performance.

Finally, models built on 5S data set appear to perform well on all the 4 data sets. It is interesting to see that all the models on 5S have the highest average PPV over other models. This suggests that models built on complex structures and large number of examples (motif) are capable to detect motifs of higher PPV. Y_{5S}^{DIST} shows a poor performance on itself. This is most likely attributed to the way distance is calculated between the given structure and the reference structure. To investigate this, we normalized the distance by information content and length (for HSL3 only) of the motifs. Results show a global improvement in the performance. Tables 4.14–4.15 shows the measures of the predictions made by all the models.

Using normalized distance, we see that the average performance of Y_{HSL3}^{NDIST} is 28.93% as opposed to 1.9% by Y_{HSL3}^{DIST} in terms of sensitivity. Also PPV is increased by 38.5%. Performance of Y_{IRE}^{DIST} remains the same for all the data sets. The prediction of 5S structure by Y_{tRNA}^{NDIST} has an PPV of 84.6% where it was only 13.3% by Y_{tRNA}^{DIST}. Also, Y_{5S}^{NDIST} performs well on all the 4 data sets. Indeed, it is interesting to see an increase in maximum sensitivity and PPV as more motifs

are included, more specifically the maximum sensitivity and PPV of top 8 structurally distinct motif range from 63.13–93.67% and 85.20–100%. Also, a decrease in minimum sensitivity and PPV is seen with growing number of motifs, indicating significant ranking.

4.6.2 MCC, PPV or Distance

We wanted to test which response variable was more suitable to build regression models such that motifs of high sensitivity/PPV are ranked the highest. This was possible, since we built models on MCC, PPV and Distance. In order to compare the performance, we evaluate the models based on measures of the top 1, 2, 4 and 8 structurally distinct motifs. The comparison will only be objective if the models are not evaluated on the base data set (data set used to build the model) as each base model is expected to perform well on the training data.

The instantiation step in Seed allows all the base pairs of every stem to be replaced by the actual base pair that occurs in the seed sequence in a combinatorial manner. As a result, motifs having the same structure with different sequence are produced. These structures exhibit similar energy scores and hence they are often ranked in tandem. Therefore, in our comparison we consider structurally different motifs in the order they are ranked. Tables 4.16–4.18 show the performance of the different models. Here, the scores represent the maximum, minimum and average measure of a model performance on the different data sets. The performance on the training data is excluded.

Results suggest that the performance of MCC and Distance (normalized) models is better than PPV models. We expected PPV based regression models to pick motifs of higher PPV. However, they only appear to perform well with 5S model when 4 or more structures are considered. Elsewhere, similar performance is shown by MCC based models, see Table 4.18. NDIST based tRNA model show a better performance for the top 1–8 predicted structures. In comparison, MCC models build on HSL3 and IRE data set shows promising result. They have high measures of performance on all the data sets (except for the average sensitivity when 4

structures are considered). For the top two structurally different motif, NDIST based 5S model performance are high in sensitivity measure but low in PPV. Higher PPV motifs are predicted by MCC based 5S model.

In general, MCC based models outperform NDIST based models in terms of PPV. The sensitivity measure of NDIST based models appears to be high for the first two predicted structures. As more structures are included, the performance (AVGSEN, AVGPPV) of most of the models tends to decrease indicating that the favorable motifs are ranked higher. The performance of the models appear to be the similar as more structures are considered.

4.6.3 Identifying complex structures from models built on simple structures and vice-versa

Data sets of varying complexity were used for this work. This allowed us to measure the success of identifying complex structures from models built on simple structures and vice-versa. Rank based evaluation of the regression equations serve to indicate that the ranking ability of models built on simple structures is low for complex structures (tRNA and 5S). Considering the top ranked motifs, HSL3 based models fail to predict a 5S structure beyond a AVGMCC of 27.3%. However, models built on complex structures have high ranking correlations across all data set. This indicates that more number of motifs indeed helps the scoring functions achieve better ranking ability. Regression models based on tRNA and 5S predict single stem structures, HSL3 and IRE, with minimum AVGMCC of 91.26% and 59.80%.

With the result from Tables 4.8–4.15, all the models show an increase in performance with increasing data set size. Since outliers were allowed during model building, it was conceivable that models resulting from more examples would have more of a chance of making mistakes, as well as more of a chance to capture the variability of different scores. Empirically, it appears that having more examples does the second better than the first.

4.7 Summary

In this chapter we introduced thermodynamic based regression models built on data sets of varying complexities to identify biological relevant motifs. Regression analysis permits to assign separate weight for each score, allowing one to emphasize or compensate the variance that differs across the different scores. It was found that the models were able to identify high PPV motifs. Models based on MCC showed better performance than PPV and Distance based models. The presence of outliers had considerable effect on the residual based measures and made them unreliable and inappropriate for model comparison. On the other hand, the ranking based evaluation were much more stable and robust to outliers, allowing a more confident comparison.

Our major conclusion are 1) models build on less complex structures appear to be unstable, that is they are not effective to associate the variations of scores obtained from different functions and different performance measures and 2) that ranking performance of regression models increases when more number of examples are used.

Chapter 5

Minimum description length principle

The first approach used to rank RNA consensus motifs was based on the nearest neighbor thermodynamic stability. Although nearest neighbor energy parameters do work well for shorter sequences, they are known to pose difficulties in correctly predicting large RNA structures [11]. We therefore, investigate an alternative paradigm, minimum description length (MDL) encoding, for evaluating the significance of consensus motifs. We evaluate the motifs using the description length as a selection criteria. A manuscript presenting this work has been published in the International Conference on Bioinformatics and Computational Biology conference [2].

5.1 Introduction

The minimum description length principle [6, 50] was formulated in the context of computational complexity and coding theory. Its use for statistical model selection has developed over the last decade or so, largely as a result of the work by Rissanen (1978). The purpose of statistical modeling is to discover regularities in the observed data.

The secondary structure prediction is seen here as a scientific discovery process. Paraphrasing

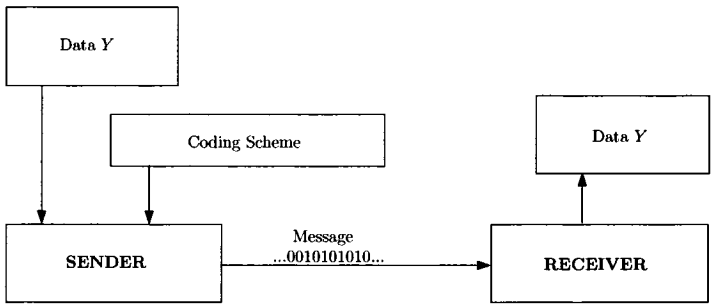

Figure 5.1: MDL schematic diagram

Grünwald, an important component of any discovery process consists of finding regularities in data. Regularities can be used to compress the data. Thus, when considering competing hypotheses, the hypothesis that achieves the highest compression can also be considered the most significant [15]. Accordingly, the minimum description length principle postulates that the best model (motif in our case) is the one that minimizes the length in bits, of both the description of the model and the data encoded by the model.

In this chapter, we will consider the basic concept of MDL and, in particular how to calculate the description length of the motifs. The discussion in Section 5.2 has been adapted from [7, 19].

5.2 The MDL principle

Let the data to be summarized be $Y = (y_1, y_2, ..., y_n)$, where $y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is a finite and countable set. Suppose we wish to summarize Y by a statistical model from a class $\mathcal{M} = \{P_\theta(Y) : \theta \in \Theta(m), m \in M\}$ where m stands for a model, M a set of models, θ a parameter vector taking values in some parameter space Θ.

5.2.1 Representing the data

Figure 5.1 illustrates the task of a sender who must represent the data $Y = (y_1, y_2, \cdots, y_n)$, as a sequence of symbols 0 and 1. The sender must choose a coding scheme to represent Y and the receiver must be able to decode the message, recreating the original data Y. The coding scheme used to represent Y assigns to each possible y_i a particular codeword. In order for the decoder to be able to decode a codeword as soon as the message comes, these code words have to be unique and unambiguous, that is, no codeword is a prefix of another codeword. Such a code is termed as prefix code.

For example, if '0' is used to represent y_1, then '00' may not be used to represent any other value, y_2, for then the receiver would not be able to tell whether the received message '00' represented y_2 or y_1 followed by a second copy of y_1. Thus, the message consists of codeword for y_1 followed by the codeword of y_2 and so on.

5.2.2 Efficient codes and statistical models

An intuitive way of discovering whether or not the prefix property holds is by drawing a code tree. Figure 5.2 gives the possible code words as a tree. A coding scheme specifies some of these possible code words to be selected: the selected code words will denote values of Y. The following code words are selected in Figure 5.2: '1', '00', '011' and '010'. For prefix codes, if a code word such as '00' is selected then no code word below it in the tree may be selected, for '00' is a prefix of all codewords below it in the tree.

For each fixed level of the tree, if $l(X)$ denotes the length (in symbols) of code word X, then $\sum 2^{-l(X)} = 1$. Combining this with the prefix restriction, we obtain the *Kraft Inequality* of coding theory as stated in Theorem 1.

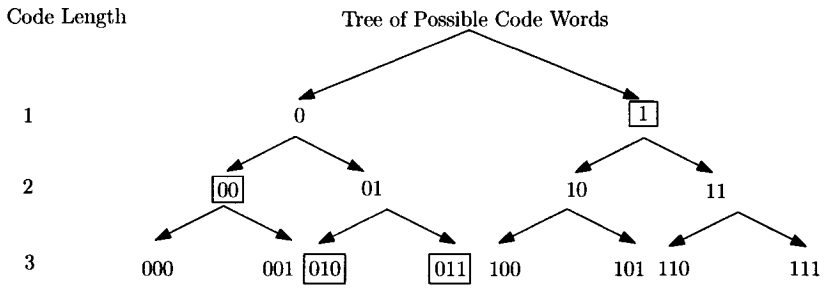

Figure 5.2: Selecting code words for prefix code

Theorem 1 *The sufficient and necessary condition for a code to be a prefix code is as follows,*

$$\sum_{X\,selected} 2^{-l(X)} \leq 1 \tag{5.1}$$

where l(X) denotes the length (in symbols) of codeword X.

For an efficient code the above equation will equal to 1. Now, let us assume that the elements of $\mathcal{Y}$ are generated according to some probability distribution P. Then the expected (average) code length of a code for encoding the information source X is

$$L(X) = \sum_{x \in X} P(x) \cdot l(x) \tag{5.2}$$

Moreover, we define the entropy of the distribution of X as

$$H(X) = -\sum_{x \in X} P(x) \cdot \log_2 P(x) \tag{5.3}$$

Theorem 2 *The expected code length of a prefix code for encoding the information source X is greater than or equal to the entropy of X, namely*

$$L(X) \geq H(X)$$

We can define a prefix code in which symbol $y \in Y$ is assigned a codeword with code length $l(y) = -\log_2 P(y)$ according to Theorem 1, since

$$\sum_{y \in Y} 2^{\log_2 P(y)} = 1 \tag{5.4}$$

The expected code length is bounded below by $H(X)$. Such a code in on average the most efficient prefix code, according to Theorem 2. We refer to such codes as non-redundant code, that is, they remove the redundancy without any loss of information by assigning short codewords to common symbols and long codewords to rare symbols.

By minimizing the code length we are choosing a model that closely corresponds to the observed distribution of data, thereby making the code optimal in information-theoretic sense to the observed frequency.

5.2.3 Model selection based on code length

In determining which model of $\mathcal{M}$ best represents Y, we choose the one that yields the smallest code length. In other words, we choose $\theta \in \Theta$ to minimize $-\sum_{y \in Y} P_\theta(y)$, where θ is an encoding scheme. We denote the minimizing value of θ by $\hat{\theta} = \hat{\theta}(Y)$, which is formally equal to the maximum likelihood estimator for θ.

In case if $\mathcal{M}$ contains only one model specified by θ_0, then the code produced makes it the shortest and only code for the data Y, and its length is $-\sum_{y \in Y} P_{\theta_0}(y)$. However, in general, the code produced by the coding scheme corresponding to $P_{\hat{\theta}}(X)$ will be the shortest code for the data Y, but it still may not be the legitimate prefix code. The receiver in Figure 5.1 does not know which encoding scheme (θ) was used to decode the message. Therefore, the sender should first encode the value of θ and transmit it before sending data Y. If the user encodes $\hat{\theta}$ by a coding scheme corresponding to some probability function $g(\hat{\theta})$, then $-\log_2 g(\hat{\theta})$ is the extra code length that needs to be transmitted. Complicated models with large parameter spaces would tend to have large values of $-\log_2 g(\hat{\theta})$, and thus longer messages. Coding theory thus measures both the model's ability to describe the data and its complexity on a common scale: code length.

Therefore, the total description length consists of three parts

1. Some coded representation of g, of the model class $\mathcal{M}$ used to encode Y;
2. A coded representation of the optimal value of θ ($\hat{\theta}$), with a code length $-\log_2 g(\hat{\theta})$, coded by the coding scheme corresponding to $g(\hat{\theta})$;
3. A coded representation of the data Y, given the model and its parameters.

Since the code length for part 1 is usually the same for all model classes, it is generally omitted. Finally, the model with the minimum description length is selected,

$$L_{\min} = \min_{m} \left(-\log_2 P_\theta(Y) - \log g(\theta) \right) \tag{5.5}$$

In other words, the model that minimizes

$$L(M) + L(D|M) \tag{5.6}$$

is selected, where $L(M)$ is the length of the model and $L(D|M)$ is the length of the data (D) given the model. From here on, we will use M to signify a model for clarity.

5.3 Method

We propose a simple encoding that follows [51] and is as close as possible to the encoding used by Seed. Let $T = \{T_1, T_2, ..., T_k\}$ be the set of k input sequences, such that $T \in \sum_{RNA}$. $\sum_{RNA} = \{A, C, G, U\}$ is the 4-letter nucleotide alphabet. We also assume that the sequences have been sorted by length, so that T_1 is the shortest input sequence. Let M denote a consensus RNA secondary motif. For a given M, let T_+ denote the sequences of T matching M, and T_- the rest, that is, $T_- = T \backslash T_+$, it is also assumed that T_1 is not included in $T+$, see below.

Let $L(X)$ be a function that returns the length of its inputs X, in bits. The length (in bits) of the motif M given T is defined as follows.

$$L(T_1) + L(M) + L(T_+) + L(T_-)$$

where $L(T_1)+L(M)$ defines the length of the model (M) and $L(T_+)+L(T_-)$ defines the length of the data (D).

5.3.1 Encoding the model

We use information theory in its fundamental form to measure the significance of different motifs [6, 50]. The theory takes into account the probability of a nucleotide in a motif (or sequence) when calculating the description length of the motif (or sequence). Shannon showed that the length in bits to transmit a symbol b via a channel in some optimal coding is $-\log_2 P_x(b)$, where $P_x(b)$ is the probability with which symbol b occurs. Given the probability distribution P_x over an alphabet $\sum_x = \{b_1, b_2, ..., b_n\}$, we can calculate the description length of any string $b_{k_1}b_{k_2}...b_{k_l}$ over the alphabet $\sum_x$ by $-\sum_{i=1}^{l}\log_2 P_x(b_{k_i})$.

The length in bits to transmit the sequence T_1 is given as

$$dlen(T_1) = -\sum_{i=1}^{4} n_{a_i} P(a_i)$$

where the probability distribution P is calculated by the occurring frequencies of nucleotides in the data set, $a_i \in \sum_{RNA}$ and n_{a_i} is the number of occurrences of a_i.

The following example illustrates the encoding of the second part of the model M (a motif in our case).

```
                    1         2         3         4         5
           012345678901234567890123456789012345678901234567890123
Seq      = GCCGGCUUGGUAAUGGUAAACGCCUCCCCUGUCACGGGAGAUCGCAAAUCAAAA
Motif    = NCCNNNNNNNNNNNGGNNNNNNNCNNNNNNNNNNNNNNNG
Struc    = ((((.........))))......(((((.......)))))
Bitmap   = 0110                   10000
```

In the above example, the motif M was inferred using T_1. Accordingly, it is described with respect to T_1.

The given motif consists of two stems. It matches the sequence indicated by 'Seq' at position 0. As a sender, we first send the sequence T_1 followed by the information of the motif. A bitmap representation is used to encode the stem information of the motif. A bit value of '1' indicates the receiver to extract the information of the base from the sequence T_1. Assuming only Watson-Crick base pairs, the base at 3'-end can be re-constructed. Bit value of '0' indicates that the base not conserved across the sequences. Consequently, the base information is extracted from the remaining sequences (T_+) sent later. To reconstruct the motif at the other end, the location and length of the stems are also transmitted. Hence, we achieve compression by sending the positions of the stems followed by the bitmap representation instead of sending the complete motif.

We encode the above motif as 0, 3, 16, 0, 1, 1, 0, 23, 27, 17, 1, 0, 0, 0, 0, \$. where \$ is a delimiter to signal the end of the motif. The first three numerals indicate the start, end (position) and length of the first stem occurring in the motif, followed by the bitmap representation of that stem. This is followed for the remaining stems.

Let $\sum_1$ denote the alphabet $\{s, e, l, 1, 0, \$\}$, where s, e and l are the parameters that define the stem. Let P_1 denote the probability distribution over the alphabet $\sum_1$. $P_1(\$)$ can be approximated by the reciprocal of the average length of motifs. $P_1(s)$, $P_1(e)$ and $P_1(l)$ can be calculated as $n(P_1(\$))$ where n denotes the number of stems. Finally, $P_1(0) = (1 - (n + 1)P_1(\$))P(0)$ and $P_1(1) = (1 - (n + 1)P_1(\$))P(1)$.

Given P_1, we can calculate the description length of a motif. For the above example the description length is

$$dlen(M) = -(log_2P_1(\$) + 2log_2P_1(s) + 2log_2P_1(l) + 2log_2P_1(e) + 6log_2P_1(0) + 3log_2P_1(1))$$

5.3.2 Encoding positive and negative sequences

Sequences matching the motif (T_+) are encoded with the help of motif M while the rest (T_-) are encoded similarly to T_1. Let us illustrate with an example how to encode the matching sequences with the help of the motif.

```
                    1         2         3         4         5
          012345678901234567890123456789012345678901234567890123456789012345678
Seq     = CGUACACCCCGAAACCAAGGGUGCCCGCCGGCACUGUCACUGCCGACGAAAAAAAAAAAA
Motif   =      NCCNNNNNNNNNNNGGNNNNNNNCNNNNNNNNNNNNNNNG
Struc   =      ((((.........))))......(((((.......)))))
Bitmap  =      0110                   10000
```

For the receiver to recreate the sequence, we need to send the 5'-end nucleotide information of the stem whose bitmap has a value '0', the offset k of the motif and the remaining part of the sequence. The 3'-end of the stem can be created from the 5'-end information as they are complementary. Seed allows a range operator that permits additional base pairs to be considered by the pattern matcher. We use a separate delimiter i to indicate insertions of base pairs.

The message to be transmitted now includes the 5'-stem information of the stems followed by the offset k, and the remaining nucleotide information. An insertion is indicated by the delimiter i containing the position followed by the nucleotide inserted. We encode the above sequence as

$A, C, G, G, C, A, k, C, G, U, A, C, C, G, A, A, A, C, C, A, A, G, C, C, C, G, C, C, U, G, U, C, A, C, A,$
$C, G, A, A, A, A, A, A, A, A, A, A, A, \$$

where \$ indicates the end of the sequence and k is the offset location of the motif. Since there are no insertions in the above example, we do not include the delimiter i.

Let $\sum_2$ denote the alphabet $\{a_1, a_2, a_3, a_4, i, k, \$\}$, where a_1, a_2, a_3, and a_4, are the four nucleotides. Also, let P_2 denote the probability distribution over the alphabet $\sum_2$. $P_2(\$)$ and $P_2(k)$ can be approximated by the reciprocal of average length of the positive sequences. $P_2(i) = nP_2(\$)$, where n denotes the number of insertions. Finally, $P_2(a_i) = (1-(2P_2(\$)+nP_2(i)))P(a_i)$.

For the above sequence, the description length is

$$dlen(T_i, M) = -(log_2P_2(\$)+log_2P_2(k)+21log_2P_2(A)+8log_2P_2(G)+15log_2P_2(C)+3log_2P_2(U))$$

5.3.3 Significance of motif

Suppose there are n sequences out of which k sequences match the motif, the significance of motif M_j, denoted by $w(M_j)$ is defined as

$$w(M_j) = \sum_{i=1}^{n} dlen(T_i) - (dlen(M_j) + \sum_{i=1}^{k} dlen(T_{+_i}, M) + \sum_{i=1}^{n-k} dlen(T_{-_i})) \tag{5.7}$$

Intuitively, the more sequences in T_+ matching M_j and the less number of bits we use to encode M_j and to encode the those sequences based on M_j, the larger weight M_j has.

5.4 Experiments

We test the performance of MDL based scoring method on the four data sets introduced in Chapter 2. More specifically, RNA motifs from the experiments HSL 02, IRE 02, tRNA and 5S were used (see Table 2.5). We also perform rank based evaluation with different performance measures discussed in Chapter 4. Plots are drawn to highlight the relationship between the different performance measures and description length of the motifs. Finally, the performance measure of the top 1, 2, 4 and 8 structurally distinct ranked motifs are discussed.

Table 5.1: Kendall's and Spearman's correlation : MDL

	HSL3		IRE		tRNA		5S	
	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$	$\hat{\tau}$	$\hat{\rho}$
AVGSEN	0.686	0.813	0.561	0.640	0.585	0.770	0.277	0.395
AVGPPV	0.708	0.819	0.562	0.638	0.435	0.603	0.119	0.166
AVGMCC	0.686	0.813	0.561	0.640	0.540	0.727	0.199	0.288

5.5 Results

For each experiment, frequency based probability distributions for each of the four nucleotides were derived from the corresponding data set. Figures 5.3–5.6 show the different performance measures (AVGPPV, AVGSEN and AVGMCC) as a function of the description length. Separate plots are used for each measure as it would be easier to distinguish the performance. For 5S experiment, the number of motifs was large (364,505). We took a sample 10,000 motifs that were picked randomly for 5S data set. The results for 5S show the average measure of 10 samples of 10,000 motifs.

Table 5.3 shows the performance measures of the top 1, 2, 4 and 8 structurally distinct motifs predicted. The ranking statistics obtained for the data sets are shown in Table 5.1. We observe that the Spearman's rank correlation is high for HSL3, IRE and tRNA data sets. However, for 5S, the ranking statistics are low; the weighted ranking ($\hat{\rho}$) with AVGSEN is 0.395. It is interesting to note that the ranking statistics for AVGPPV decreases with increasing complexity of data set. We discuss these results in the next section.

5.6 Discussion of results

For all the data sets, MDL scoring method has been able to correctly identify (rank) RNA consensus secondary structures. The performance measure plots indicate that motifs with low

description length correspond to high sensitivity/PPV.

For single stem structure, HSL3, the top motif ranked by the scoring method has sensitivity and PPV of 100%, see Table 5.7. This motif matches all the 28 input sequences. Also for the second data set, IRE, we see that the high sensitivity/PPV motifs have lower description length; the top motif measuring at 92.7/100%. Figures 5.3 and 5.4 shows that the motif producing the maximum compression (minimum length) also have high MCC. As expected, the search space also contains small and/or generic motifs with low compression but high MCC; represented by spikes toward the right hand side of the diagram. In case of HSL3, there are many motifs with varying AVGSEN (79.5–100%), for the intermediate description length values. This could explain the small difference in the value of Kendall's $\hat{\tau}$ on AVGSEN when compared to AVGPPV, where the variation of positive predicted values is less. In general, the high values of ranking statistics support the performance of the model on HSL3 and IRE data sets. For both the data sets, the maximum sensitivity and PPV of the top 8 structurally distinct motifs are 100 and 100% respectively, see Table 5.3. Also, we see a decrease in the minimum sensitivity as more structures are considered, indicating that the high sensitivity/PPV motifs are ranked highest for single stem structures.

We now focus on the performance of the MDL scoring method for complex structures. For tRNA, the motif with the lowest description length has an average sensitivity/PPV of 73.7/88.2%. It matches 5 of the 7 input sequence. We see a strong correlation between the different performance measures and the description length (see Figure 5.5). This also fits in nicely with the significance results of $\hat{\rho}$ (0.727 for AVGMCC). Figure 5.5 indicates that a subset of motifs with high PPV, ($>$ 88.2%) receive intermediate description length values. The lower value of $\hat{\tau}$ (0.435) quantifies this irregularity.

On 5s data, the average sensitivity is lower than that for the other data sets. However, the top motif has a sensitivity of 40.8%, which is 7.8 percentage points less than the maximum observed sensitivity for the whole search space. Although, the ranking statistics for 5S indicate

poor ranking performance, the results of the top ranks motifs is optimal, in the sense that this the motif with the highest MCC in our search space. The maximum sensitivity and PPV of the top 8 structurally distinct motifs are 50 and 86.4% respectively (Table 5.3) .

For all the four experiments, the figures and ranking statistics support the use of MDL criterion for ranking consensus motifs. Top-ranked motifs generally correspond to high PPV/sensitivity motifs while bottom-ranked motifs correspond to low PPV/sensitivity motifs.

5.7 Summary

The MDL principle is a general method for inductive inference, which equates learning with finding a short description (encoding) of the data. In this chapter, we proposed a simple encoding for consensus RNA secondary structure motifs. The data is encoded by first encoding a model and then encoding the data with the help of the model. The description length obtained were used to rank the motifs. In this way, we hope to choose a model that optimizes the trade-off between model complexity and goodness-of-fit. Furthermore, this ranking scheme does not require *a priori* information such as thermodynamic parameters.

Data set	motif	AVGSEN	AVGPPV	AVGMCC
HSL3	269	100	100	100
GGCNCTNNNNAGNGCC / ((((((....))))))				
IRE	85	92.7	100	96.28
NNNNNNCNNNNNNNNNNGNNNNN / (((.(((((.....))))))))				
tRNA	569	73.7	88.2	80.62
NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN (((((..((((.........)))).(((((.......)))))).....(((..........)))))))				
5S	124279	40.4	70.9	53.52
NNNNGNNNNNNNNGCGNNGNGNNNNNNNNNNGNNNNNNNNNNNNNNNNCNNNNNNNNNNCNCNNNNNCGCNNNNNNNNNNNNNNNNNNNNNNCNNNNNNNNNNNNNGNNNNNNNCNNNN (((((((((.....(((((((((........(((.............)))).........)))))...))).....................(((.............)))..))))))))				

Table 5.2: Characteristics of top motif picked by MDL based scoring scheme showing the performance measures of the top ranked motifs. Each prediction is followed by the respective structure and sequence.

	Data set	1 Structure		2 Structures		4 Structures		8 Structures	
		SEN	**PPV**	**SEN**	**PPV**	**SEN**	**PPV**	**SEN**	**PPV**
MIN	*HSL3*	100	100	83.3	100	0	0	0	0
	IRE	80	100	70	100	60	100	0	0
	tRNA	58.8	63.6	58.8	58.8	52.9	56.2	52.9	56.2
	5S	36.8	63.6	36.8	63.6	34.2	61.9	23.7	45
MAX	*HSL3*	100	100	100	100	100	100	100	100
	IRE	100	100	100	100	100	100	100	100
	tRNA	81	100	81	100	81	100	81	100
	5S	42.1	72.7	50	86.4	50	86.4	50	86.4

Table 5.3: Characteristics of top 1–8 structurally distinct motifs picked by models showing the minimum and maximum sensitivity (**SEN**) and positive predict value (**PPV**).

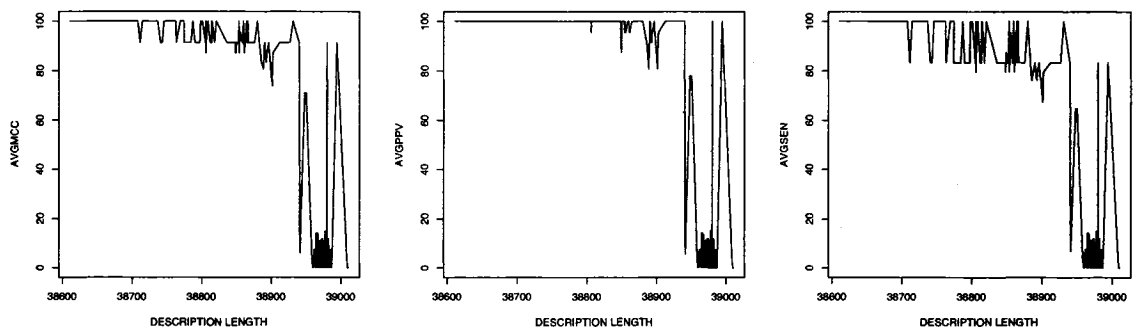

Figure 5.3: Performance measure as a function of description length on HSL3 data set

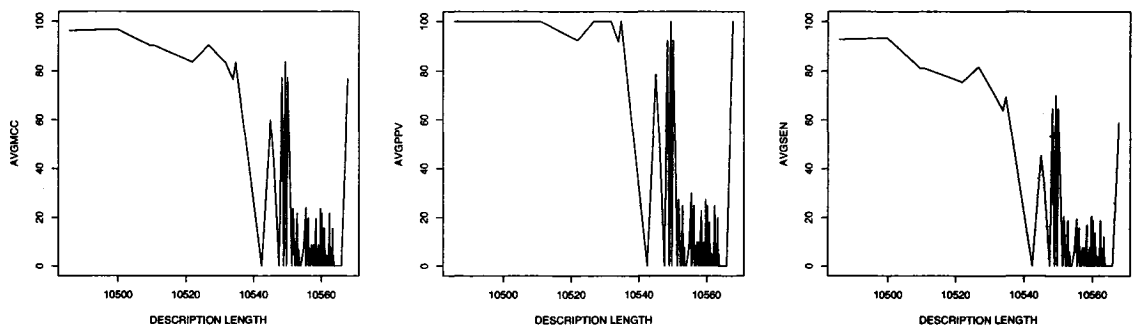

Figure 5.4: Performance measure as a function of description length on IRE data set

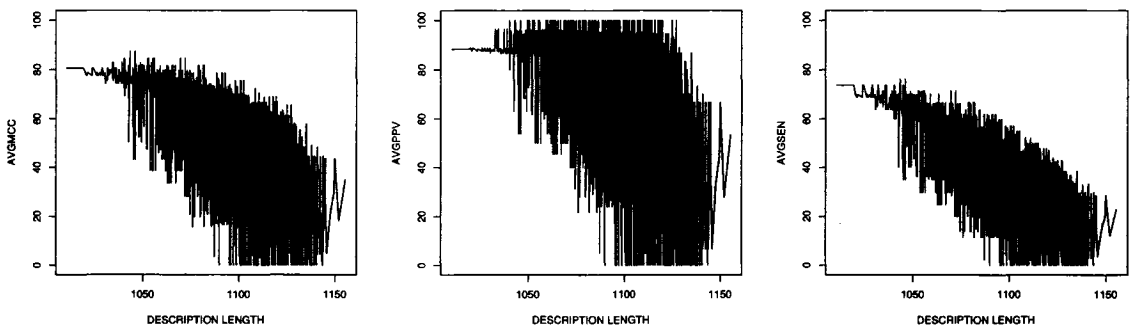

Figure 5.5: Performance measure as a function of description length on tRNA data set

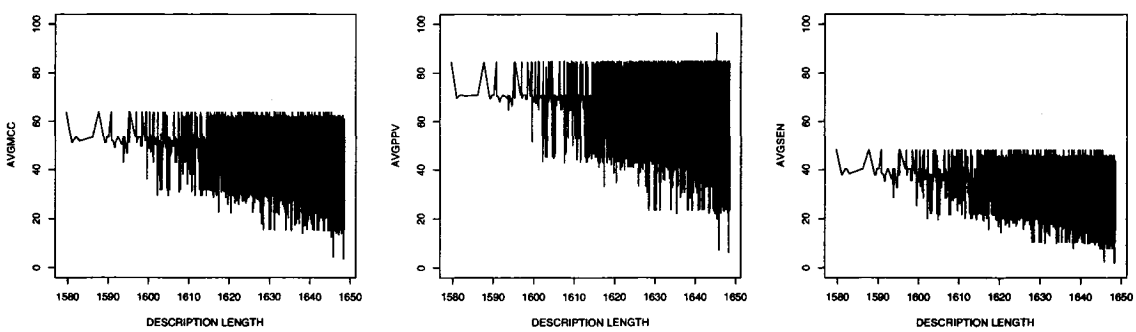

Figure 5.6: Performance measure as a function of description length on 5S data set

Chapter 6

Discussion and conclusion

6.1 Discussion

In this thesis, we developed three scoring schemes for the software system Seed to distinguish the biologically relevant RNA secondary structure motifs from the rest. We tested our methods on four different data sets having varying range of complexity. Two data sets we constructed consisted of selected members of UTRdb database where the coding regions are flanked by two untranslated regions (5' UTR and 3' UTR). Others were assembled using a subset of sequences from [31].

Our first approach was based on the sum of the thermodynamics free energy, based on the nearest neighbor model. We introduced simple objective functions consisting of a combination of the free energy of matching sequences. Each functions captured different features of a RNA structure. We found the energy functions were strongly correlated to MCC, suggesting that motifs with low folding free energies correspond to high sensitivity/PPV motifs. For all the data sets, the functions served as a good discriminator of false positives. In the second method we combined the scoring functions through regression analysis. Several models were formulated using different dependent variables on all data sets. Each model built on a particular data set was evaluated based on its performance of generalizing instances outside of the training

data. We found that the regression models can generally identify motifs with high measures of specificity and positive predicted values to known motifs. We also found that the performance of models based on the PPV and MCC were better than Distance. The poor prediction for Distance based model could be due to limited set of edit operations between the tree representations of RNA secondary structures [16]. We then introduced an alternative paradigm, MDL encoding, for evaluating the significance of consensus motifs. The performance of this method was evaluated on same four data sets stated above. The top ranked motifs correspond to high measures of sensitivity and PPV. As expected, the motifs matching more number of sequences and having more base pairs instantiated were ranked higher.

6.2 Comparison of methods

We compare the proposed scoring methods based on the following three criterion: ranking statistics, sensitivity/PPV measures of top ranked motif averaged for all data sets and MCC of top 1–8 structurally distinct motifs. The results of the following experiments were used for comparison: HSL 02, IRE 02, tRNA and 5S (see Table 3.3).

6.2.1 Ranking performance

Table 6.1 shows the ranking statistics ($\hat{\rho}$) for the scoring methods. Here, MCC has been used as the response variable. Overall, Y_{tRNA}^{DIST} regression model has the highest average statistics (0.904). The lowest ranking statistics is seen for *NTInfo*. Among the objective functions, *NTFirst* has the best Spearman's correlation. In general, the average statistics for the thermodynamics based objective function is higher than the regression models. The overall lower performance of MDL based scoring scheme is attributed due to the poor performance on 5S data set; the Spearman's correlation on 5S is 0.288.

Y_{IRE}^{DIST} shows the highest correlation on HSL3 02 and IRE 02. The high correlation on IRE

data set achieved by Y_{IRE}^{DIST} regression model is expected as it has been formulated on the same data set. Similar is the case on tRNA and 5S data sets; base models show higher ranking statistics. On the tRNA data set, Y_{tRNA}^{MCC} performs the best among the models with ($\hat{\rho}$) equal to 0.887. This model has the second highest correlation on IRE (0.928) and 5S (0.874) data set respectively, see Table 6.1. Although Y_{5S}^{DIST} has the highest ranking statistics on 5S data set, it fails to correctly predict high sensitivity/PPV motifs. The poor predicting performance and high ranking statistics is due to the way edit distance is calculated between the tree representation of predicted and reference structure. We therefore also compare the models based on the sensitivity/PPV of the top ranked structure.

6.2.2 Sensitivity/PPV of the top ranked motifs

The characteristics of the top consensus structures predicted by different models on all the data sets is shown in Table 6.2. Except for HSL3, the sensitivity of all the predicted structure is less in comparison to its PPV values. This is to be expected as a consensus structure will be less representative of each individual structure. As such it can never have a coverage of 100%.

The normalized variants of the objective functions show better performance over their base scores in terms of both sensitivity and PPV. The range of averaged sensitivity and PPV for the normalized scores are 76.3–79.4/93.2–96.2% in comparison to 60.1–78.7/91.9–95.3% for the base scores. The top ranked motifs predicted by *TSum* for HSL3 and IRE data set has a sensitivity of 83% and 58.9%, respectively. When normalized by the number of sequences matched the sensitivity increases to 100% and 93.3%.

Of the regression models, MCC based models predict structures with higher sensitivity and PPV. The range of averaged sensitivity and PPV are 62.1–74.1%. Among the Distance based regression models, Y_{HSL3}^{DIST} fails to rank favorable structures for all the data sets except HSL3. This attributes to lower averaged sensitivity/PPV of 22.3/28.3% for all the data set.

In comparison to regression models, MDL based scoring scheme predicted optimal structures

with high sensitivity and positive predicted value. For tRNA data set, the sensitivity of top ranked motif is equal to the maximum sensitivity achieved by the regression models and base scoring functions. However, on 5S data set the average sensitivity is lower than Y_{IRE}^{PPV} by 7.8 percentage points. Overall, the performance of MDL based scoring method is good with average sensitivity/PPV of 76.7/89.8%

Except *NTInfo*, the normalized variants perform well for all the data sets and are the best scoring functions in this comparison. The results indicate that while the regression models gave comparable performance on single stem structures, the (normalized) scoring functions are superior for tRNA and 5S data sets. More work is needed to see if there are detectable properties in the data that could indicate the selective use of objective functions.

6.2.3 MCC of the top 1–8 motifs

Table 6.3 shows the Matthew correlation coefficient of the top 1–8 structurally distinct motifs. For each scoring function, the observed maximum MCC values are averaged for all the data sets. Overall, the performance of the models are good with the maximum MCC ranging from 64.89 to 88.93%.

Similar performance is seen by the different energy based models models. It is interesting to observe that as more structures are considered, the performance of information content based models, *TInfo* and *NTInfo* is better. More specifically, when 2 structures are considered, the maximum MCC is 88.93% as compared to 87.56% by energy based scoring functions. However, when more structures are included, the performance of the objective functions are similar. The normalized variants of the energy based models show a constant performance of 88.25% indicating that the favorable motifs are ranked highest.

Of the regression models, Distance based models (Y_{HL3}^{DIST} and Y_{IRE}^{DIST}) show poor MCC values. The maximum MCC achieved considering 2 structures are 49.94% and 74.26% respectively. However, models built on complex structures, Y_{tRNA}^{DIST} and Y_{5S}^{DIST} perform better with maximum

MCC measure of 86.27% and 82.29% for 2 structures. In comparison, PPV based models show an increase in MCC when more structures are considered. The averaged maximum MCC for 8 structures ranges from 85.34% to 87.35%. In general, MCC based regression model show a slight increase in MCC over energy based objective functions with 4 or more structures. This class of models exhibit the highest MCC for 8 structures.

MDL based scoring method show good performance on all the data sets. The average maximum MCC is 86.33% for the top ranked structure, see Table 6.3. With 8 distinct structures, the model also has the highest value of maximum MCC (88.93%).

Table 6.1: Ranking statistics $\hat{\rho}$ for the scoring methods for four of the six experiments. MCC has been used as the response variable. Bold values indicate the highest value of ranking statistics.

Score	HSL3 02	IRE 02	tRNA	5S	Average
TInfo	0.754	0.509	0.686	0.465	0.604
TAvg	0.918	0.702	0.844	0.668	0.783
TBest	0.923	0.712	0.873	0.748	0.814
TFirst	0.900	0.778	0.874	0.735	0.822
TLeft	0.918	0.729	0.877	0.736	0.815
TSum	0.863	0.743	0.857	0.766	0.807
TWorst	0.909	0.727	0.872	0.724	0.808
NTInfo	0.565	0.327	0.609	0.454	0.489
NTAvg	0.923	0.715	0.845	0.667	0.788
NTBest	0.926	0.715	0.873	0.746	0.815
NTFirst	0.903	0.780	0.874	0.734	0.823
NTLeft	0.922	0.751	0.877	0.735	0.821
NTSum	0.908	0.741	0.876	0.738	0.816
NTWorst	0.912	0.750	0.873	0.723	0.815
Y_{HSL3}^{MCC}	0.907	0.720	0.794	0.536	0.739
Y_{IRE}^{MCC}	0.886	0.706	0.799	0.572	0.741
Y_{tRNA}^{MCC}	0.905	0.676	**0.887**	0.737	0.801
Y_{5S}^{MCC}	0.889	0.596	0.863	0.761	0.777
Y_{HSL3}^{PPV}	0.895	0.709	0.647	0.321	0.643
Y_{IRE}^{PPV}	0.866	0.686	0.670	0.339	0.640
Y_{tRNA}^{PPV}	0.919	0.750	0.842	0.653	0.791
Y_{5S}^{PPV}	0.892	0.609	0.813	0.683	0.749
Y_{HSL3}^{DIST}	0.948	0.916	0.277	0.686	0.707
Y_{IRE}^{DIST}	**0.975**	**0.976**	0.506	0.772	0.807
Y_{tRNA}^{DIST}	0.961	0.928	0.851	0.874	**0.904**
Y_{5S}^{DIST}	0.964	0.924	0.828	**0.880**	0.899
MDL	0.813	0.640	0.727	0.288	0.617

Table 6.2: Characteristics of the top motifs picked by the models showing the sensitivity (**SEN**) and positive predicted value (**PPV**).

Score	HSL3 02		IRE 02		tRNA		5S		Average	
	SEN	PPV	SEN	PPV	SEN	PPV	SEN	PPV	SEN	PPV
TInfo	100	100	92.7	100	73.7	88.2	48.2	84.6	78.7	93.2
TAvg	100	100	93.3	100	71.3	96.7	48.2	84.6	78.2	95.3
TBest	100	100	58.9	100	71.3	96.7	48.2	84.6	69.6	95.3
TFirst	100	100	93.3	100	71.3	96.7	48.2	84.6	78.2	95.3
TLeft	100	100	93.3	100	71.3	96.7	48.2	84.6	78.2	95.3
TSum	83	100	58.9	100	71.3	86.7	27.3	80.8	60.1	91.9
TWorst	100	100	93.3	100	63.7	96.7	48.2	84.6	76.3	95.3
NTInfo	100	100	92.7	100	73.7	88.2	48.2	84.6	78.7	93.2
NTAvg	100	100	92.7	100	76.2	100	48.2	84.6	79.3	96.2
NTBest	100	100	93.3	100	76.2	100	48.2	84.6	79.4	96.2
NTFirst	100	100	93.3	100	76.2	100	48.2	84.6	79.4	96.2
NTLeft	100	100	93.3	100	76.2	100	48.2	84.6	79.4	96.2
NTSum	100	100	93.3	100	76.2	100	48.2	84.6	79.4	96.2
NTWorst	100	100	93.3	100	76.2	100	48.2	84.6	79.4	96.2
Y_{HSL3}^{MCC}	100	100	92.7	100	73.7	88.2	18.1	41.2	71.1	82.4
Y_{IRE}^{MCC}	100	100	92.7	100	73.7	88.2	30.1	72.5	74.1	90.2
Y_{tRNA}^{MCC}	100	100	92.7	100	71.2	96	30.1	72.5	73.5	92.1
Y_{5S}^{MCC}	83.3	100	58.9	100	76.2	100	30.1	72.5	62.1	93.1
Y_{HSL3}^{PPV}	100	100	0	0	68.7	87.5	18.1	41.2	46.7	57.2
Y_{IRE}^{PPV}	100	100	92.7	100	73.7	88.2	48.2	84.6	78.7	93.2
Y_{tRNA}^{PPV}	83.3	100	45.5	78.6	71.2	96	30.1	72.5	57.5	86.8
Y_{5S}^{PPV}	83.3	100	58.9	100	66.3	95.7	30.1	72.5	59.7	92.1
Y_{HSL3}^{DIST}	83.3	100	0	0	5.7	13.3	0	0	22.3	28.3
Y_{IRE}^{DIST}	100	100	92.7	100	68.7	87.5	0	0	65.4	71.9
Y_{tRNA}^{DIST}	100	100	92.7	100	73.7	88.2	3.2	13.3	67.4	75.4
Y_{5S}^{DIST}	100	100	92.7	100	68.7	87.5	0	0	65.4	71.9
MDL	100	100	92.7	100	73.7	88.2	40.4	70.9	76.7	89.8

Table 6.3: Characteristics of the top 1–8 motifs picked by the models showing the averaged maximum MCC for all the data sets.

Score	1 Structure	2 Structures	4 Structures	8 Structures
TInfo	86.33	88.93	88.93	88.93
TAvg	87.56	87.56	88.25	88.25
TBest	82.32	87.56	87.56	88.25
TFirst	87.56	87.56	87.56	88.25
TLeft	87.56	87.56	87.56	88.25
TSum	78.33	87.35	87.35	88.54
TWorst	87.56	87.56	87.56	88.25
NTInfo	86.33	88.93	88.93	88.93
NTAvg	88.25	88.25	88.25	88.25
NTBest	88.25	88.25	88.25	88.25
NTFirst	88.25	88.25	88.25	88.25
NTLeft	88.25	88.25	88.25	88.25
NTSum	88.25	88.25	88.25	88.25
NTWorst	88.25	88.25	88.25	88.25
Y_{HSL3}^{MCC}	83.05	86.27	86.27	86.27
Y_{IRE}^{MCC}	88.93	88.93	88.93	88.93
Y_{tRNA}^{MCC}	82.31	88.25	88.93	88.93
Y_{5S}^{MCC}	86.27	83.61	88.93	88.93
Y_{HSL3}^{PPV}	57.37	86.27	86.27	86.27
Y_{IRE}^{PPV}	78.45	84.65	84.65	85.34
Y_{tRNA}^{PPV}	74.89	82.91	85.47	87.35
Y_{5S}^{PPV}	74.18	75.49	85.38	85.38
Y_{HSL3}^{DIST}	28.27	49.92	57.90	64.39
Y_{IRE}^{DIST}	70.21	74.26	75.88	78.6
Y_{tRNA}^{DIST}	80.72	80.72	83.96	86.27
Y_{5S}^{DIST}	71.82	80.72	80.72	82.29
MDL	86.33	88.93	88.93	88.93

6.3 Conclusion and future work

6.3.1 Conclusion

In this thesis we have introduced several scoring function formulations designed to pick high sensitivity/PPV motifs from an ensemble of RNA secondary structures. In Chapter 1, background information on RNA secondary structure, thermodynamics of RNA and the different prediction methods available today was given. Following this, we presented a recently developed algorithm, Seed, and illustrated the different phases of the algorithm (motif generation) with the help of an example. Three major approaches were used to predict biological relevant motifs produced by the algorithm. The first approach was based on the sum of thermodynamics free energy; several free energy functions were introduced in Chapter 3. Then in Chapter 4, we presented a general framework for analyzing RNA structures using statistical regression models. Finally, the third approach was based on the minimum description length principle, presented in Chapter 6. We evaluated the performance of the scoring methods on biological data illustrated in Chapter 2.

Results on the test data sets showed that our methods performed well. For single stem structures, HSL3 and IRE, the predictions made by most of the models have a high measure of PPV/sensitivity, often 100%. For complex structures, the methods could identify motifs with high PPV but lower sensitivity. With the free energy based functions, lowest free energy scoring motifs have a high PPV/sensitivity while highest scoring motifs have a low PPV sensitivity. The functions display good correlation to PPV and sensitivity; the correlation to sensitivity is generally higher than to PPV. All the results were supported by the ranking statistics. With regression models, we were able to detect greater performance by models that were build using more complex structures. The performance of models build on less complex structures generally showed inconsistent results on predicting complex structures. We are able to conclude that the variations in the scoring functions based on thermodynamic energy and information content can be represented by a statistical model. Following this, we were also able to implement a

MDL based scoring method. This criterion optimizes the trade-off between model (motif) complexity and how well does the model (motif) match the sequences (goodness-of-fit). As far as it is known to the author, this work represents the first MDL based methodology for evaluating RNA secondary structure motifs. The MDL method had its unique merits over thermodynamic based method as it successfully avoids using any biological information pertaining to a motif. Since this method does not use any biological information, this work could be extended to compare pseudoknotted structures for which no thermodynamics parameters are available.

With correct combination of parameters, Seed produces promising motifs since it implements an exhaustive search of the motif space. The advantage of using scoring functions is two-fold. Firstly, they give us intuitive means to compare large population of motifs and identify putative structures. Specifically, for biologists it will limit the number of structures that need to be examined. Secondly, they can be implemented in the algorithm to eliminate invalid motifs thereby reducing the search space and accelerating the run time.

All scoring methods proposed show similar performance on different data sets used in this thesis. The results in this study illustrate that nearest neighbor energies can potentially be used as novel scoring function for secondary structure programs, and can be used alone or in conjunction. The general approach of using regression models built on multiple functions to obtain different estimates of an RNA secondary structure can be useful models for motif discovery. Finally, the MDL approach appears to be promising in this context. Perhaps a combination of the above methods could result in a superior scoring scheme.

6.3.2 Future work

In future work, we will test the scoring schemes on other complex data sets to further evaluate their performance. We will also investigate and test more effective scoring methods for the Seed algorithm.

With the visualization of ranking performance of regression models, we were able to see that

certain examples were out of place. Analyzing the common characteristics of these examples could lead to conclusions about the shortcomings of the models and possible ways to improve its ranking performance. It is likely that some moderate refinements to regression models would improve its performance without altering its efficiency, for example, taking into account insertions/deletions (having a penalizing factor) and addition of other variants of thermodynamic scores.

As indicated by our regression results, all the (MCC based) models perform well and are comparable to each other. However, since the models are built on different data sets, their results differ. As shown in Table 4.8, the performance of a single model in predicting the highest sensitivity/PPV motif in the search space may vary. For example, among the MCC based regression models, Y_{5S}^{MCC} predicts a motif of higher sensitivity/PPV than Y_{tRNA}^{MCC} on tRNA data set. An interesting future step would be to build a consensus scoring model by combining regression models. By apply multiple regression models in combination, we can expect the positive and negative errors to cancel each other thereby reducing large errors in predicting the response variables.

It is known that for a given secondary structure adopted by a sequence having low energy does not indicate how close it is to the global minimum energy structure. The results from MFOLD agree the above observed fact on tRNA data set (see Table 6.4). This limitation in the nearest neighbor energy rules could be improved by using a consensus structure as structural constraint to an *ab initio* method. To investigate this claim, the base pairs of the top motif (4968) predicted by Y_{5S}^{MCC} scoring model were used as input constraints for MFOLD, which was ran on all the sequences of the tRNA data set matching best motif. With the combined approach, we observed an increase in both PPV and sensitivity from 69.8/70.5%, for MFOLD alone, up to 98.26/100%. Table 6.4 shows the tRNA structures obtained using MFOLD algorithm with constraints obtained from Seed.

With MDL, future work includes developing a more theoretical/general framework expressed in

Table 6.4: The tRNA structures obtained running MFOLD with constraints obtained from Seed. The motif 4,968 which has the highest predicted MCC according to $5S_{MCC}$, was used to constrain the MFOLD algorithm. This motif matches 5 of the 7 sequences. The number in parenthesis indicate the number of sub-optimal structures reported with similar performance measures.

Id	Method	% PPV	%sensitivity
RD0260	MFOLD (2)	28.6-29.2	28.6-33.3
	MFOLD (1)	66.7	57.1
	MFOLD (1)	57.1	57.1
	Seed + MFOLD	**100.0**	**100.0**
RD1140	MFOLD (1)	100.0	100.0
	Seed + MFOLD	**100.0**	**100.0**
RD2640	MFOLD (1)	63.6	66.7
	MFOLD (2)	18.2	19.0
	Seed + MFOLD	**100.0**	**100.0**
RE2140	MFOLD (1)	87.0	95.2
	MFOLD (2)	69.6	76.2
	Seed + MFOLD	**91.3**	**100.0**
RE6781	MFOLD (4)	28.0-31.8	33.3
	Seed + MFOLD	**100.0**	**100.0**

terms of code length functions rather than a specific encoding. A major advantage of using MDL principle as scoring function lies in its generality to compare standard secondary structural motifs, for which no experimentally derived parameters are available. More specifically, the scoring model can be used to compare numerous lowest free energy conformation and help identify the native conformation.

Bibliography

[1] M. Anwar, T. Nguyen, and M. Turcotte. Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics*, 7(244), September 1 2005.

[2] M. Anwar and M. Turcotte. An approach to selecting putative RNA motifs using MDL principle. *The 2006 International Conference on Bioinformatics and Computational Biology*, 2006.

[3] M. Anwar and M. Turcotte. Evaluation of RNA secondary structure motifs using regression analysis. *Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2006.

[4] D. P. Bartel. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297, 2004.

[5] P. N. Borer, B. Dengler, I. Jr. Tinoco, and O. C. Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.*, 86:843–853, 1974.

[6] A. Brazma, I. Jonassen, E. Ukkonen, and J.Vilo. Discovering patterns and subfamilies in biosequences. pages 34–43, 1996.

[7] Peter G. Bryant and Olga I. Cordero-Brana. Model selection using the Minimum Description Length principle. *The American Statistician*, 54(4):257–268, 2000.

[8] 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.

[9] Jamie Cannone, Sankar Subramanian, Murray Schnare, James Collett, Lisa D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi Madabusi, Kirsten Muller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin 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(1):2, 2002.

[10] Jamie Cannone, Sankar Subramanian, Murray Schnare, James Collett, Lisa D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi Madabusi, Kirsten Muller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs: Correction. *BMC Bioinformatics*, 3(1):15, 2002.

[11] K. J. Doshi, J. J. Cannone, C. W. Cobaugh, and R. R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.

[12] E. Freyhult, P. P. Gardner, and V. Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6(4), 2005.

[13] J Gorodkin, S. L. Stricklin, and G. D. Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucl. Acids Res.*, 29(10):2135–2144, 2001.

[14] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucl. Acids Res.*, 31(1):439–441, 2003.

[15] P. D. Grünwald, In Jae Myung, and M. A. Pitt. *Advances in Minimum Description Length Theory and Applications.* MIT Press, 2003.

[16] V. Guignon, C. Chauve, and S. Hamel. An edit distance between RNA stem-loops. In *SPIRE*, pages 335–347, 2005.

[17] R. R. Gutell. Comparative RNA web site. http://www.rna.icmb.utexas.edu, July 2004.

[18] Robin R Gutell, Jung C Lee, and Jamie J Cannone. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, 12(3):301–310, 2002.

[19] Mark H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[20] R. R. Hocking. A biometrics invited paper: The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

[21] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[22] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). *Chemical Monthly*, 125:167–188, 1994.

[23] Minitab Inc. Minitab release 14 for Windows.

[24] Klein R. J. and Eddy S. R. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4:44, 2003.

[25] Ashwini Jambhekar, Kimberly McDermott, Katherine Sorber, Kelly A. Shepard, Ronald D. Vale, Peter A. Takizawa, and Joseph L. DeRisi. Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud localization in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 102(50):18005–18010, 2005.

[26] E. H. Kislauskis and R. H. Singer. Determinants of mRNA localization. *Curr Opin Cell Biol*, 4:975–978, Dec 1992.

[27] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31:3423–3428, 2003.

[28] P. V. Larsen. Selecting regression models. 2003.

[29] T. Macke, D. Ecker, R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif – a new RNA secondary structure definition and discovery algorithm. *Nucleic Acids Research*, 29:4724–4735, 2001.

[30] B. Masoumi and M. Turcotte. Simultaneous alignment and structure prediction of RNAs: Are three input sequences better than two? In S. V. Sunderam et al., editor, *2005 International Conference on Computational Science (ICCS 2005)*, Lecture Notes in Computer Science 3515, pages 936–943, Atlanta, USA, May 22-25 2005.

[31] B. Masoumi and M. Turcotte. Simultaneous alignment and structure prediction of three RNA sequences. *International Journal of Bioinformatics Research and Applications*, 1(2):230–245, 2005.

[32] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, , and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. pages 7287–7292, 2004.

[33] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

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

[35] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Bioploymers*, 29:1105–1119, 1990.

[36] F. Mignoe, C. Gissi, S. Liuni, and G. Pesole. Untranslated regions of mRNAs. *Genome Biology*, 3(3):0004.1–0004.10, 2003.

[37] T. Nguyen. Efficient determination of concensus secondary structures in RNA, 2004.

[38] T. Nguyen and M. Turcotte. Exploring the space of RNA secondary structure motifs using suffix arrays. *6th International Symposium on Computational Biology and Genome Informatics (CBGI 2005)*, pages 1291–1298, 2005.

[39] K. C. Pang, S. Stephen, P. G. Engstrom, K. Tajul-Arifin, W. Chen, C. Wahlestedt, B. Lenhard, Y. Hayashizaki, and J. S. Mattick. RNAdb - a comprehensive mammalian noncoding RNA database. *Nucleic Acids Research*, 125–130, 2005.

[40] Ken C. Pang, Stuart Stephen, Par G. Engstrom, Khairina Tajul-Arifin, Weisan Chen, Claes Wahlestedt, Boris Lenhard, Yoshihide Hayashizaki, and John S. Mattick. RNAdb–a comprehensive mammalian noncoding RNA database. *Nucl. Acids Res.*, 33(suppl_1):D125–130, 2005.

[41] G. Pavesi, M. Stefani, G. Mauri, and G. Pesole. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Research*, 32(10):3258–69, 2004.

[42] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, and C. Saccone. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucl. Acids Res.*, 30(1):335–340, 2002.

[43] S. Rosset, C. Perlich, and B. Zadrozny. Ranking-based evaluation of regression models. In *The Fifth IEEE International Conference on Data Mining (ICDM '05)*, page In press, Houston, Texas, 2005.

[44] M. J. Serra and D. H. Turner. Predicting thermodynamic properties of RNA. 34:242–261, 1995.

[45] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, 26:148–153, 1998.

[46] M. Sprinzl and K. S. Vassilenko. Compilation of tRNA sequences and sequences of tRNA genes. http://www.uni-bayreuth.de/departments/biochemie/trna, August 2003.

[47] Mathias Sprinzl and Konstantin S. Vassilenko. Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, 33(suppl_1):D139–140, 2005.

[48] Maciej Szymanski, Volker A. Erdmann, and Jan Barciszewski. Noncoding regulatory RNAs database. *Nucl. Acids Res.*, 31(1):429–431, 2003.

[49] J. van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

[50] J. T. Wang, B. A. Shapiro, and D. Shasha, editors. *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications.* Oxford University Press, 1999.

[51] J. T. L. Wang, Q. H. Ma, D. Shasha, and Cathy H. Wu. Application of neural networks to biological data mining: A case sutdy in protein sequence classification. 2000.

[52] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Bioploymers*, 49:145–165, 1999.

[53] T. Xia, D. H. Mathews, and D. H. Turner. Thermodynamics of RNA secondary structure formation. pages 21–47, 1999.

[54] Hong Xue, Ka-Lok Tong, Christian Marck, Henri Grosjean, and J. Tze-Fei Wong. Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene*, 310:59–66, 2003.

[55] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31:3406–15, 2003.

[56] M. Zuker and D. Sankoff. RNA secondary structure and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

[57] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.

Appendix A

IUPAC nucleic acid codes

The IUPAC code is a set of symbols encoding each subset of the four nucleotides. Here is the complete list of these symbols together with the subsets they encode

Code	Corresponding Class	Name
A	-	Adenine
C	-	Cytosine
G	-	Guanine
T	-	Thymine
U	-	Uracil
R	[GA]	Purine
Y	[TC]	Pyrimidine
K	[GT]	Keto
M	[AC]	Amino
S	[GC]	-
W	[AT]	-
B	[GTC]	-
D	[GAT]	-
H	[ACT]	-
V	[GCA]	-
N	[ACGT]	Any

Table A.1: IUPAC code

Appendix B

Description of Seed parameters

`stem_min_len <n>` **(default 3):** During the first step of the search, Seed enumerates all the valid stems. For a given pair of positions i and j, Seed computes the longest extension l such that $S[i, i+l-1]$ and $S[j, j-l+1]$ forms a canonical stem (a stem made of canonical base pairs only, possibly including G:U base pairs), if the current number of mismatches is smaller than the user defined threshold then Seed repeats this process from $i+l+1$ and $j-l-1$. This option controls the minum value of l for each iteration of the extension process.

`stem_max_gu <n>` **(default 100):** Controls the maximum number of G:U base pairs.

`min_num_stem <n>` **(default 1):** Motifs having less than this number of stems are not reported.

`max_num_stem <n>` **(default 2):** Stops the algorithm when all the motifs having this many stems have been produced.

`stem_max_separation <n>` **(default 150):** In the stem enumeration step, limits the maximum distance between the start and end of a stem.

`skip_keep_longest_stems` **(default false):** Generates all the stems from stem min len up to the maximum possible size.

`loop_min_len <n>` **(default 4):** Defines the minimum size of loops.

`nogu` **(default false):** Disallow G:U base pairs.

`range <n>` **(default 1):** The number of additional base pairs that are allowed when matching a range element (when calculating the support of a motif).

`max_mismatch <n>` **(default 2):** The maximum number of mismatches that are allowed when enumerating stems and matching motifs.

`max_fixed_pos <n>` **(default 100):** The maximum number of base pairs that will be fixed in fix all stem.

`min_base_pair <n>` **(default 5):** The minum number of base pairs for the sum of all base pairs.

`min_support <n>` **(default 0.70):** The fraction of the input sequences that a motif matches to be retained.

`time_limit <n>` **(default 0):** Limits the execution time to the specified number of minutes.

Appendix C

Seed output

The following example shows the Seed output for IRE experiment. The output shows the motif '569' matches the first sequence '0' indicated by **match id** at an **offset** of 1 and 2.

```
<motifs>
......
<motif id="569">
  <seq>NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNCNNNNNNN</seq>
  <sec>(((((..((((.........)))).(((((.......)))))).....(((.........)))))))</sec>
  <match id="0">
     <offset>1</offset>
     <energy>-13.4</energy>
     <seq>CGACCGGGGCTGGCTTGGTAATGGTACTCCCCTGTCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTCG</seq>
     <sec>(((((...((((.........)))).(((((.......)))))).....(((.........))).)))))</sec>
  </match>
  <match id="0">
     <offset>2</offset>
     <energy>-12.7</energy>
     <seq>GACCGGGGCTGGCTTGGTAATGGTACTCCCCTGTCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTC</seq>
     <sec>(((((..((((.........)))).(((((.......)))))).....(((.........)))))))</sec>
  </match>
  <match id="1">
     <offset>0</offset>
     <energy>-15.9</energy>
     <seq>GCCCGGGTGGTGTAGTGGCCCATCATACGACCCTGTCACGGTCGTGACGCGGGTTCAAATCCCGCCTCGGGC</seq>
     <sec>(((((..((((.........)))))..(((((.......)))))).....(((.........)))))))</sec>
  </match>
......
</motifs>
```

Appendix D

Visualization of ranking performance

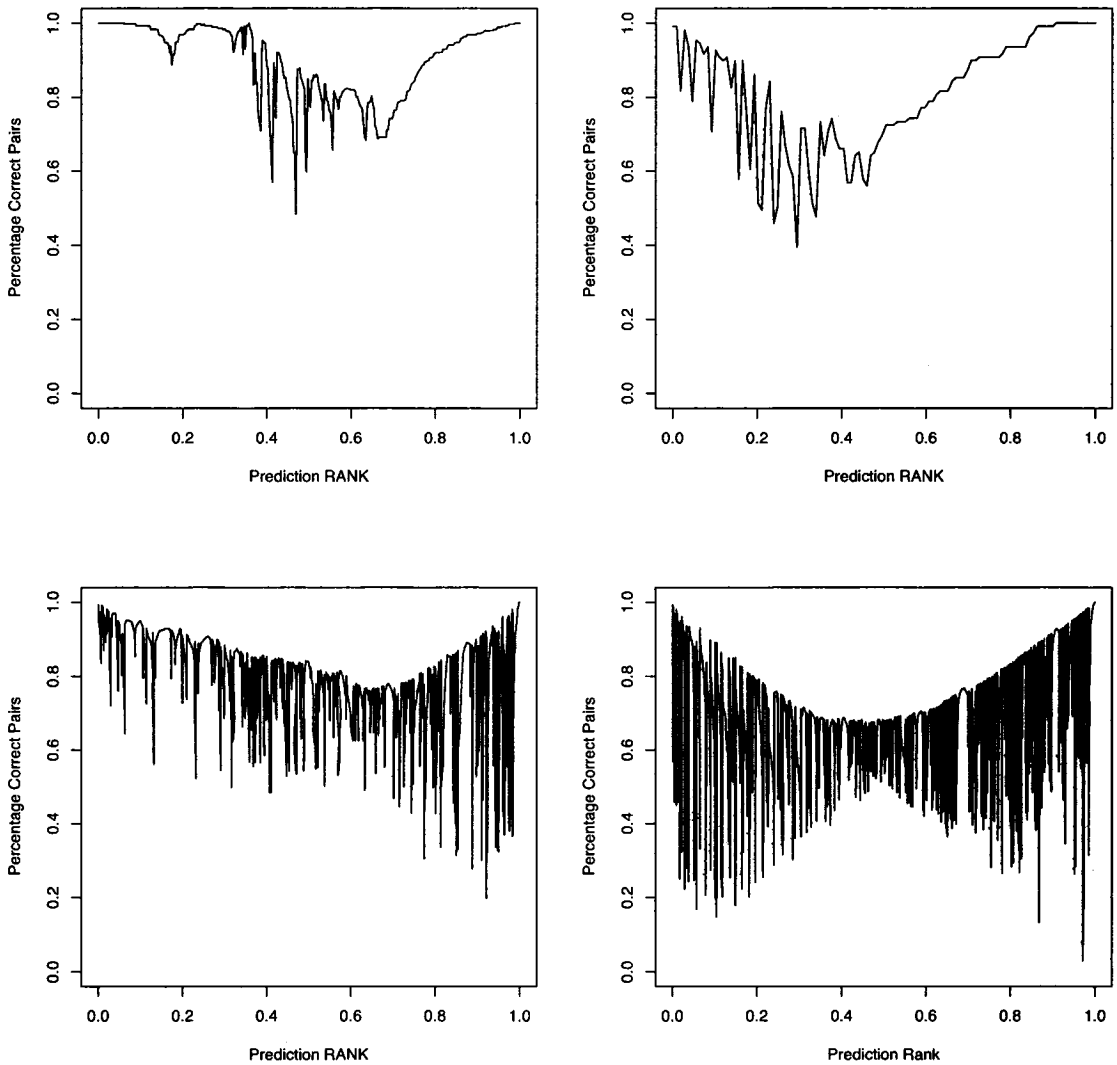

Figure D.1: Visual ranking performance of HSL3 base model. Here AVGMCC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

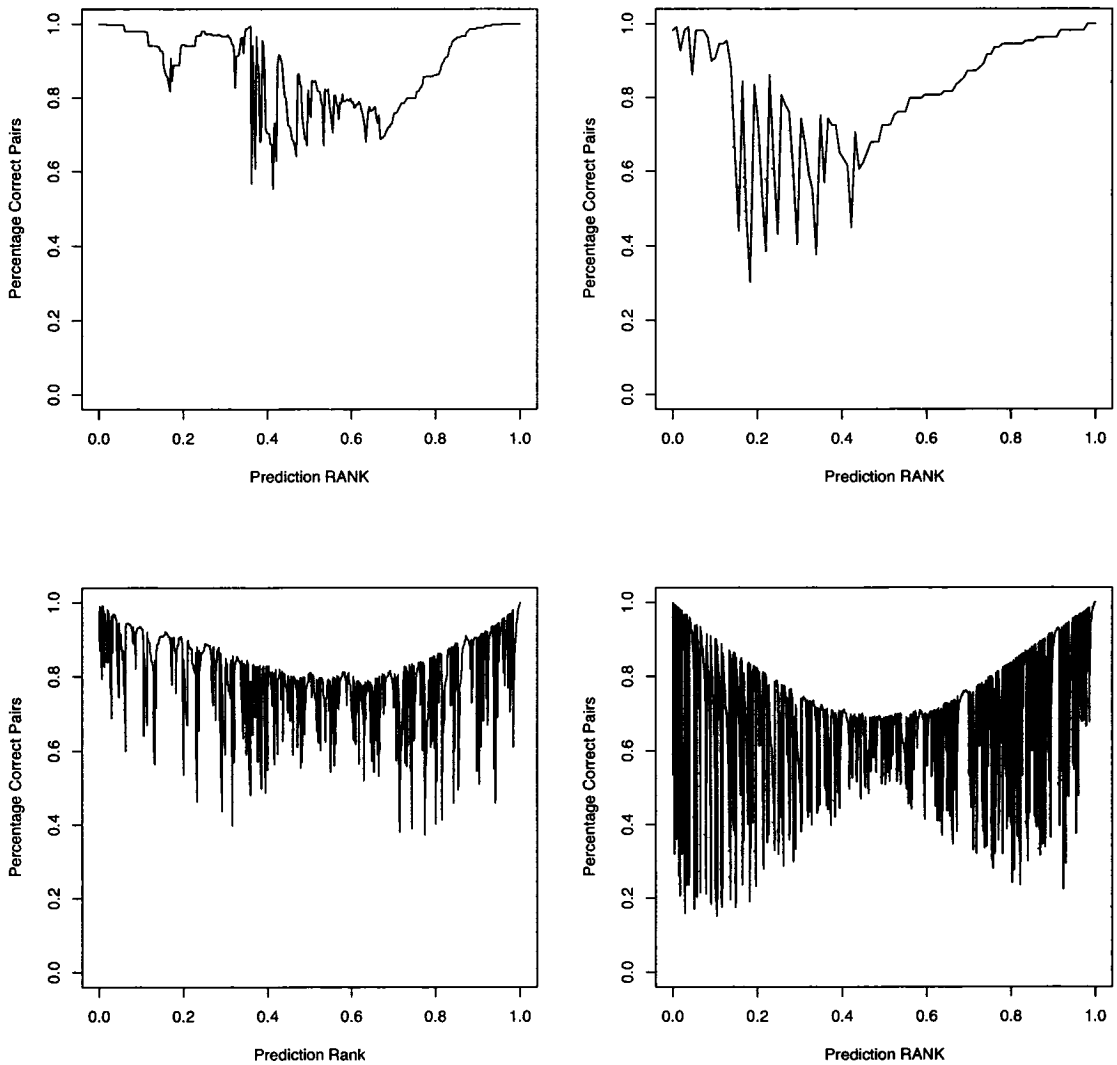

Figure D.2: Visual ranking performance of IRE base model. Here AVGMCC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

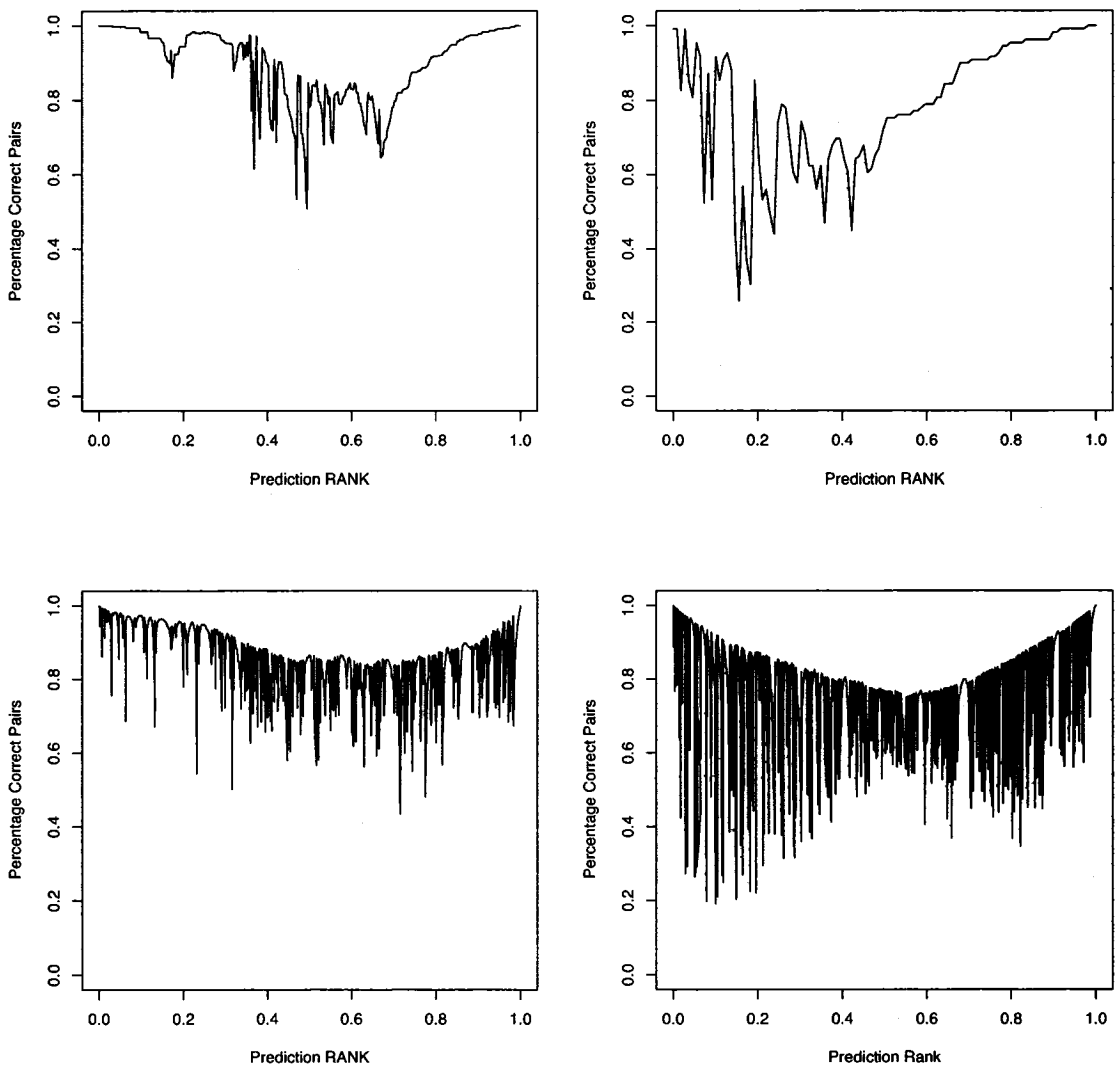

Figure D.3: Visual ranking performance of tRNA base model. Here AVGMCC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

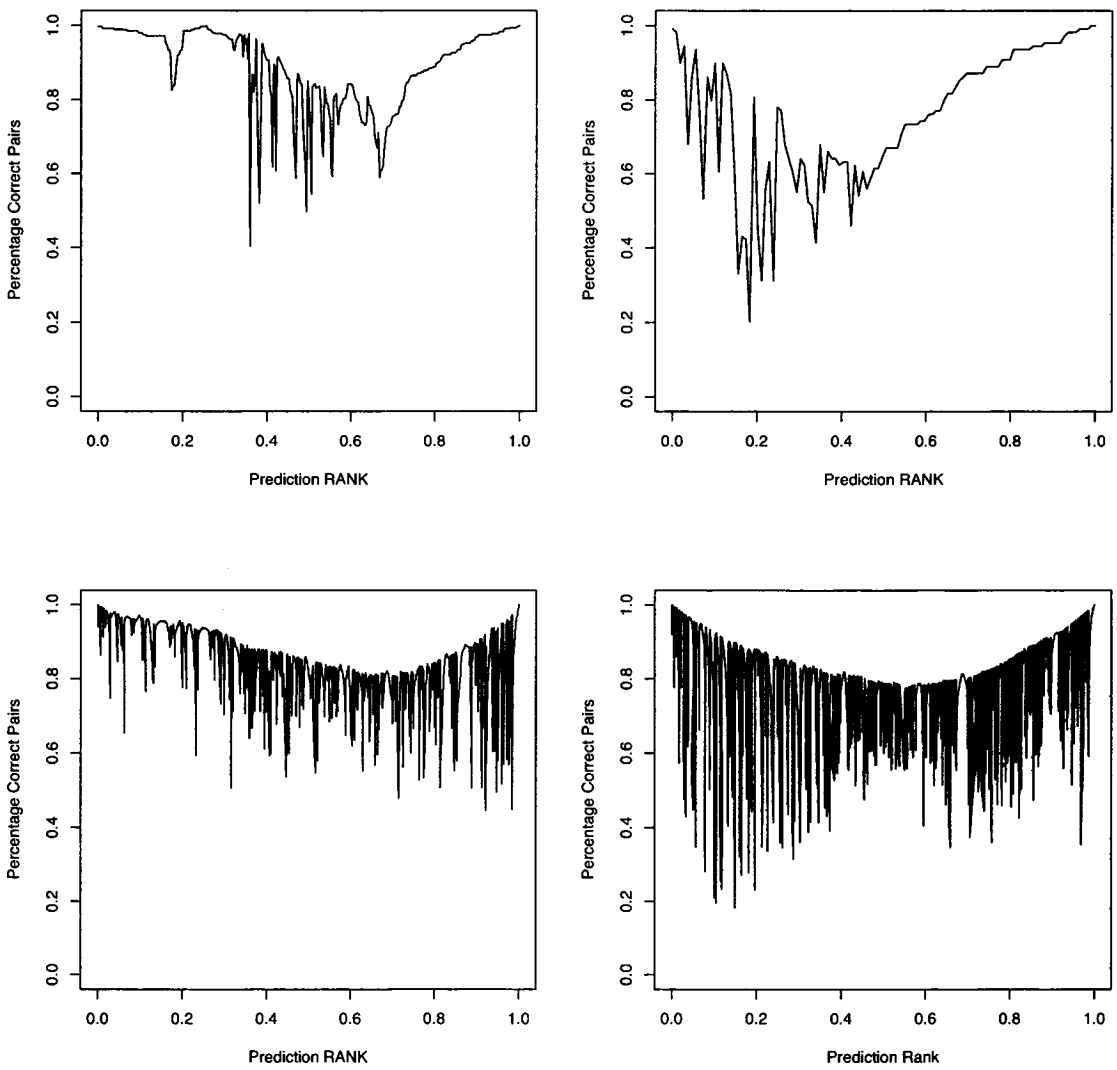

Figure D.4: Visual ranking performance of 5S base model. Here AVGMCC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

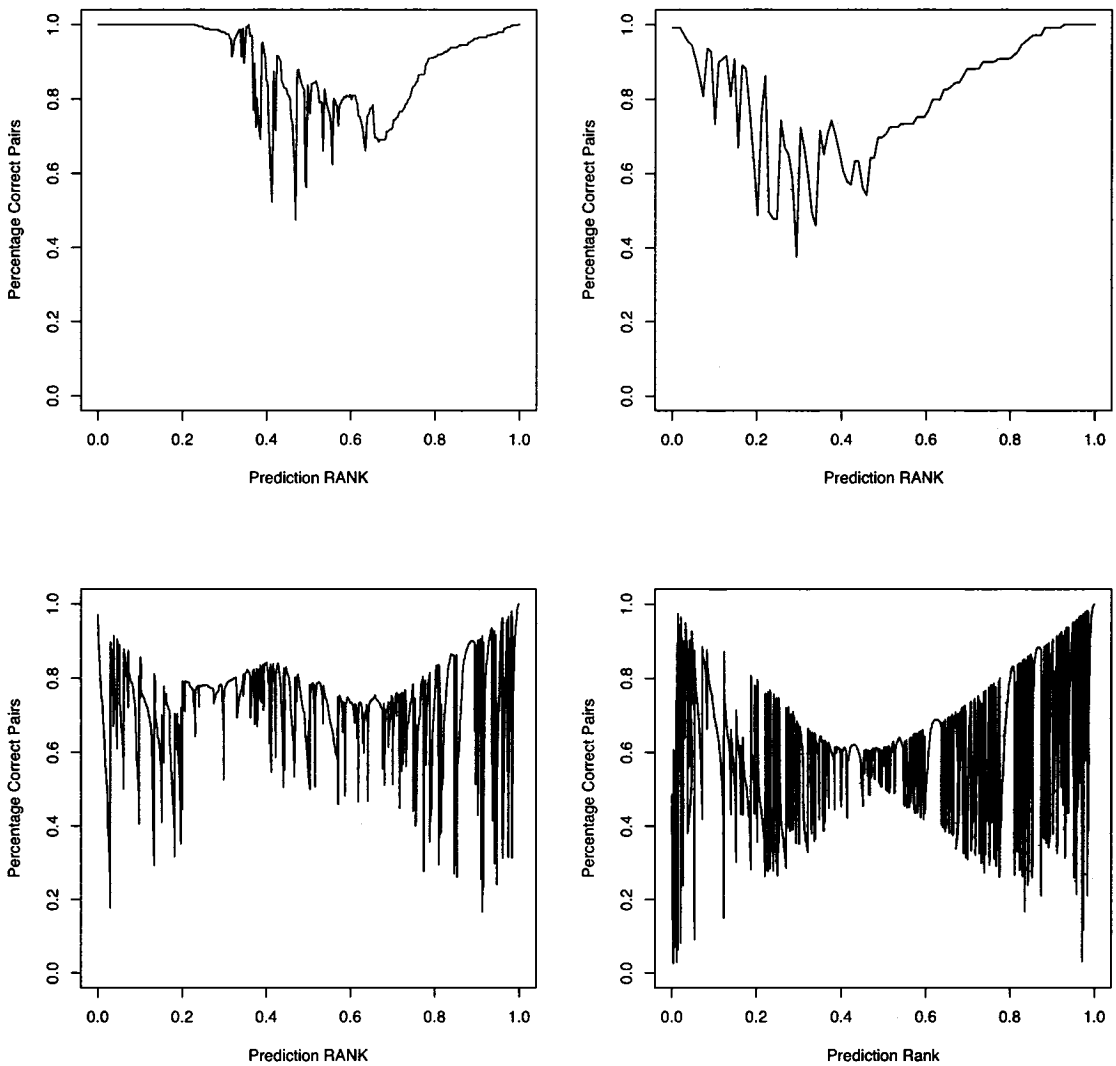

Figure D.5: Visual ranking performance of HSL3 base model. Here AVGACC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

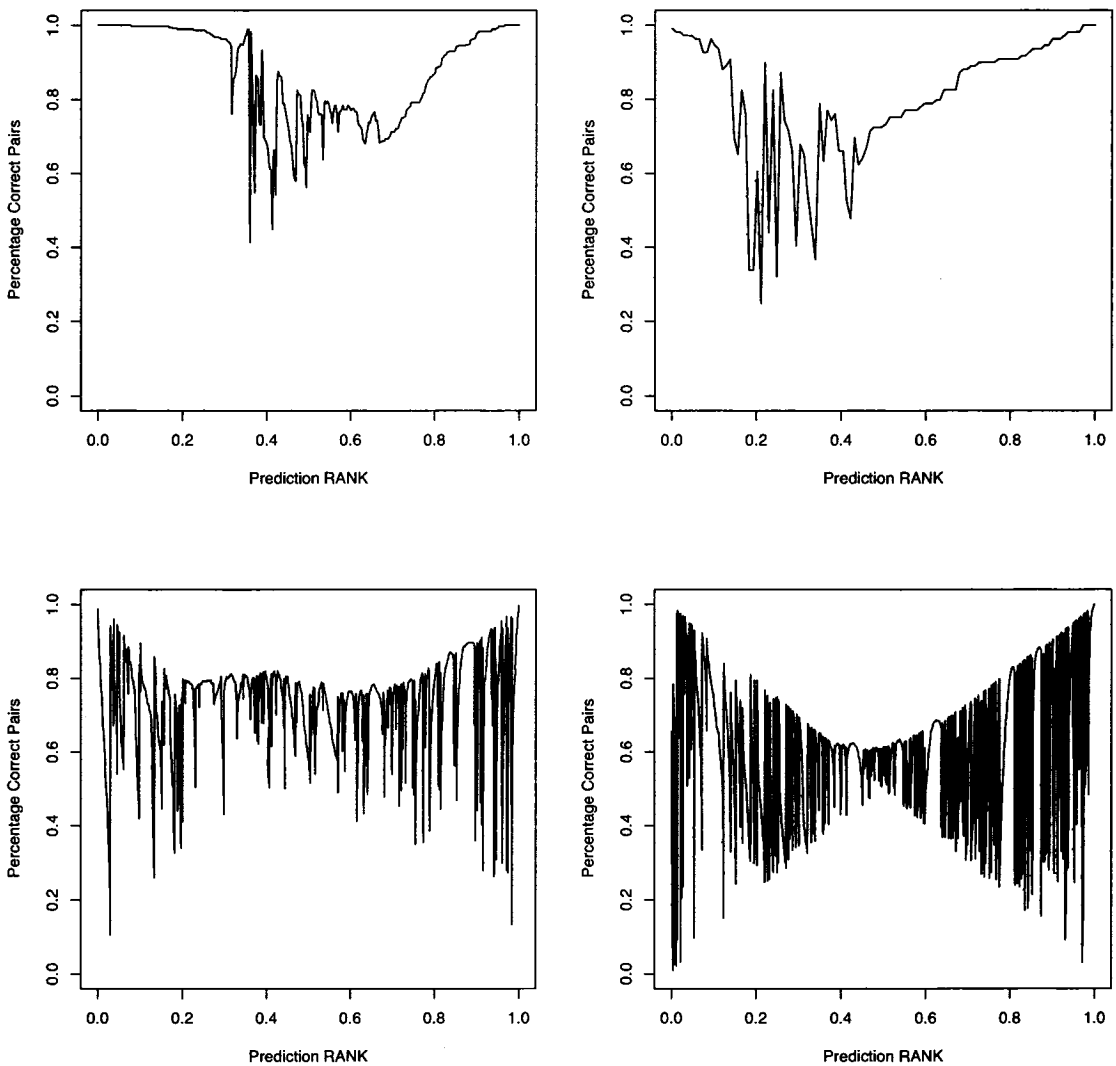

Figure D.6: Visual ranking performance of IRE base model. Here AVGACC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

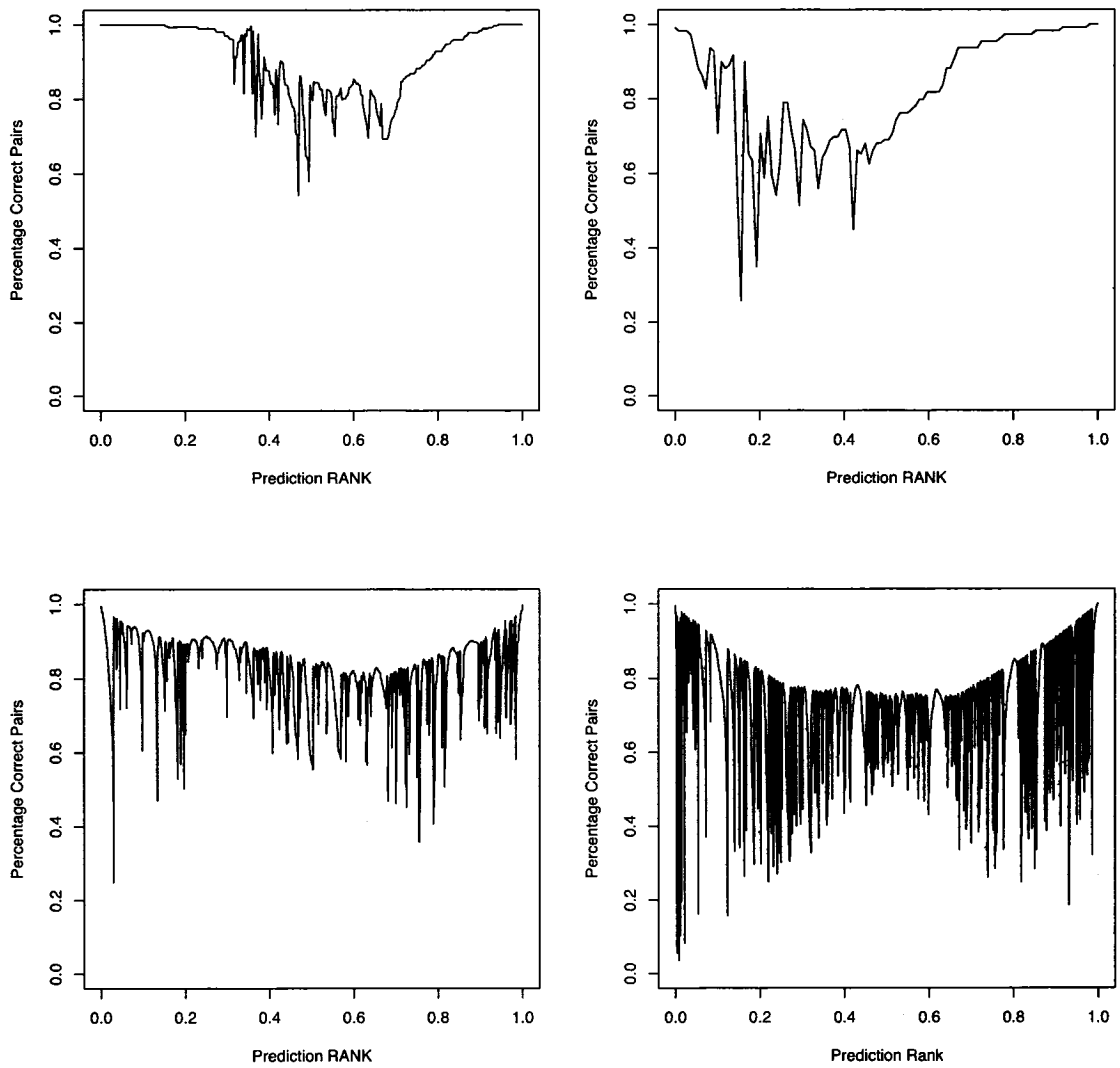

Figure D.7: Visual ranking performance of tRNA base model. Here AVGACC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

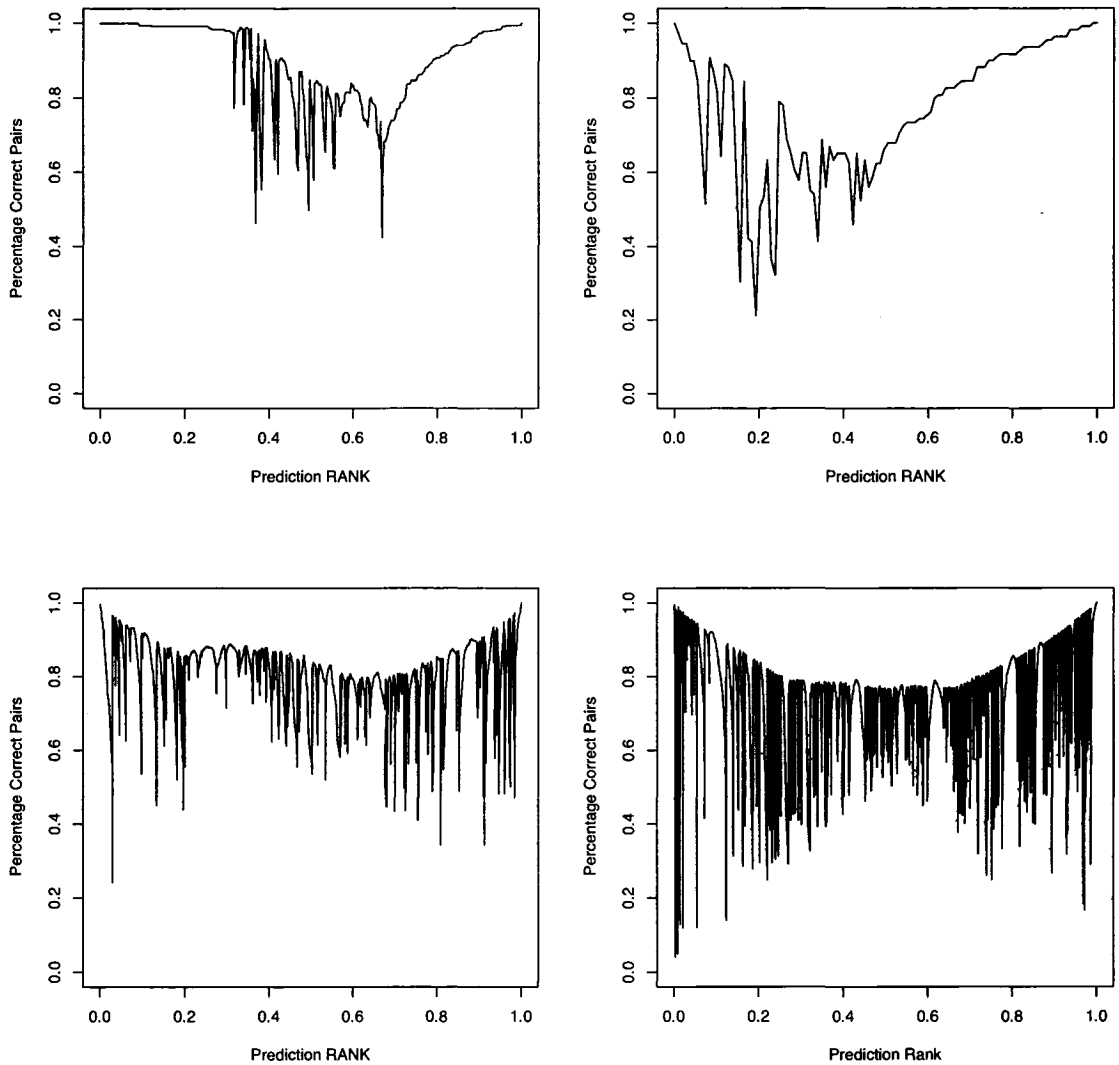

Figure D.8: Visual ranking performance of 5S base model. Here AVGACC is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

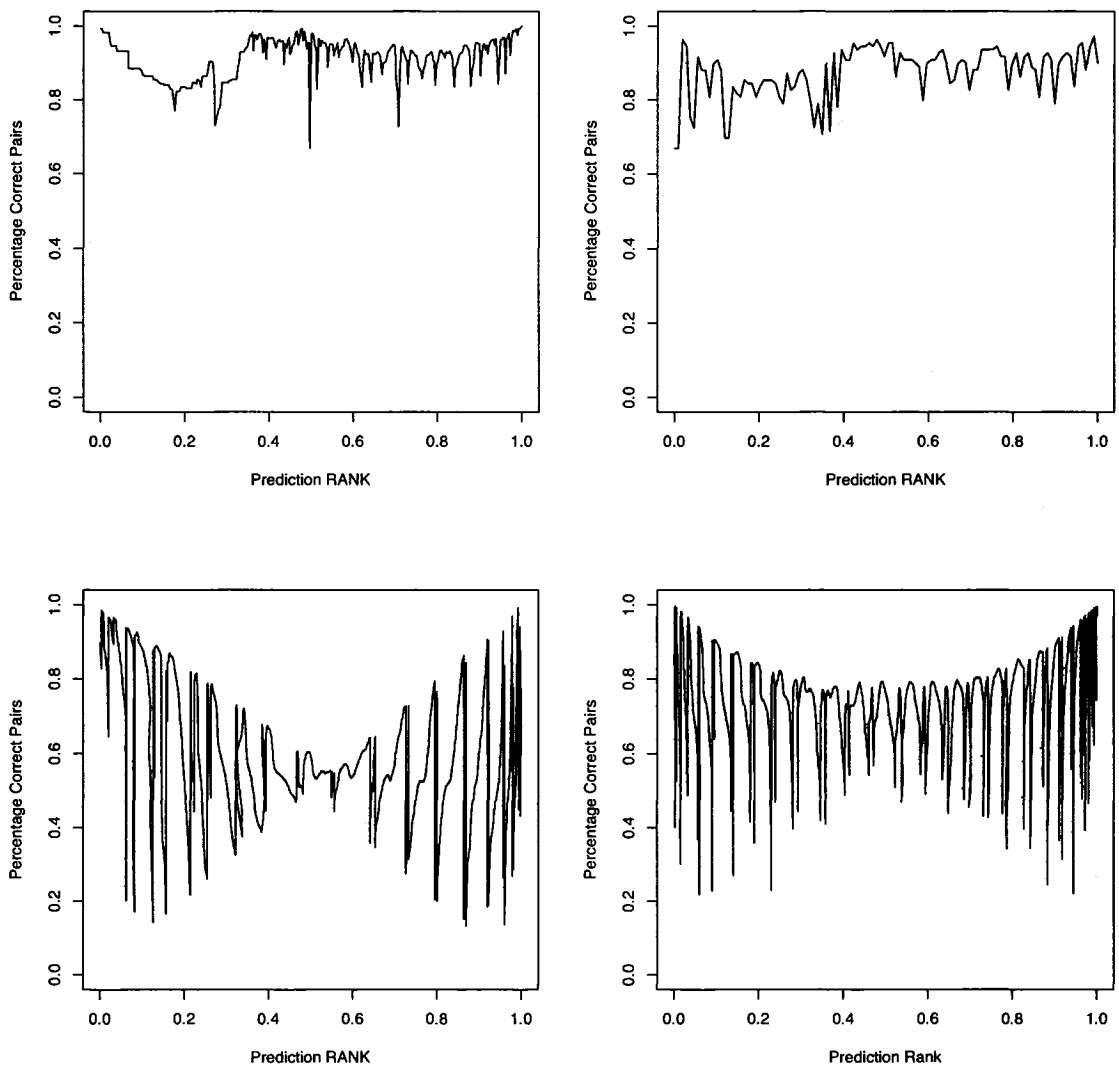

Figure D.9: Visual ranking performance of HSL3 base model. Here DISTANCE is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

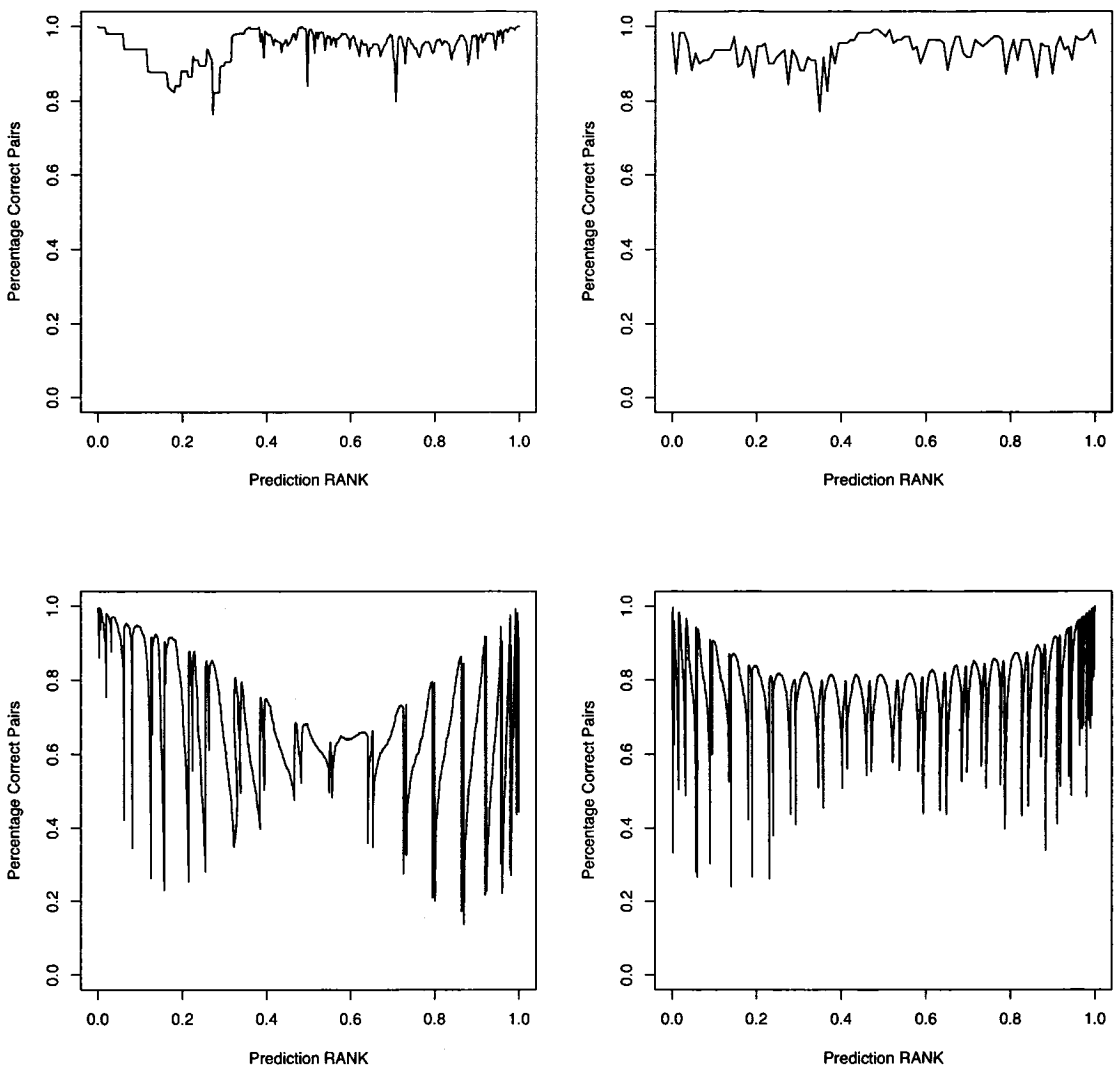

Figure D.10: Visual ranking performance of IRE base model. Here DISTANCE is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

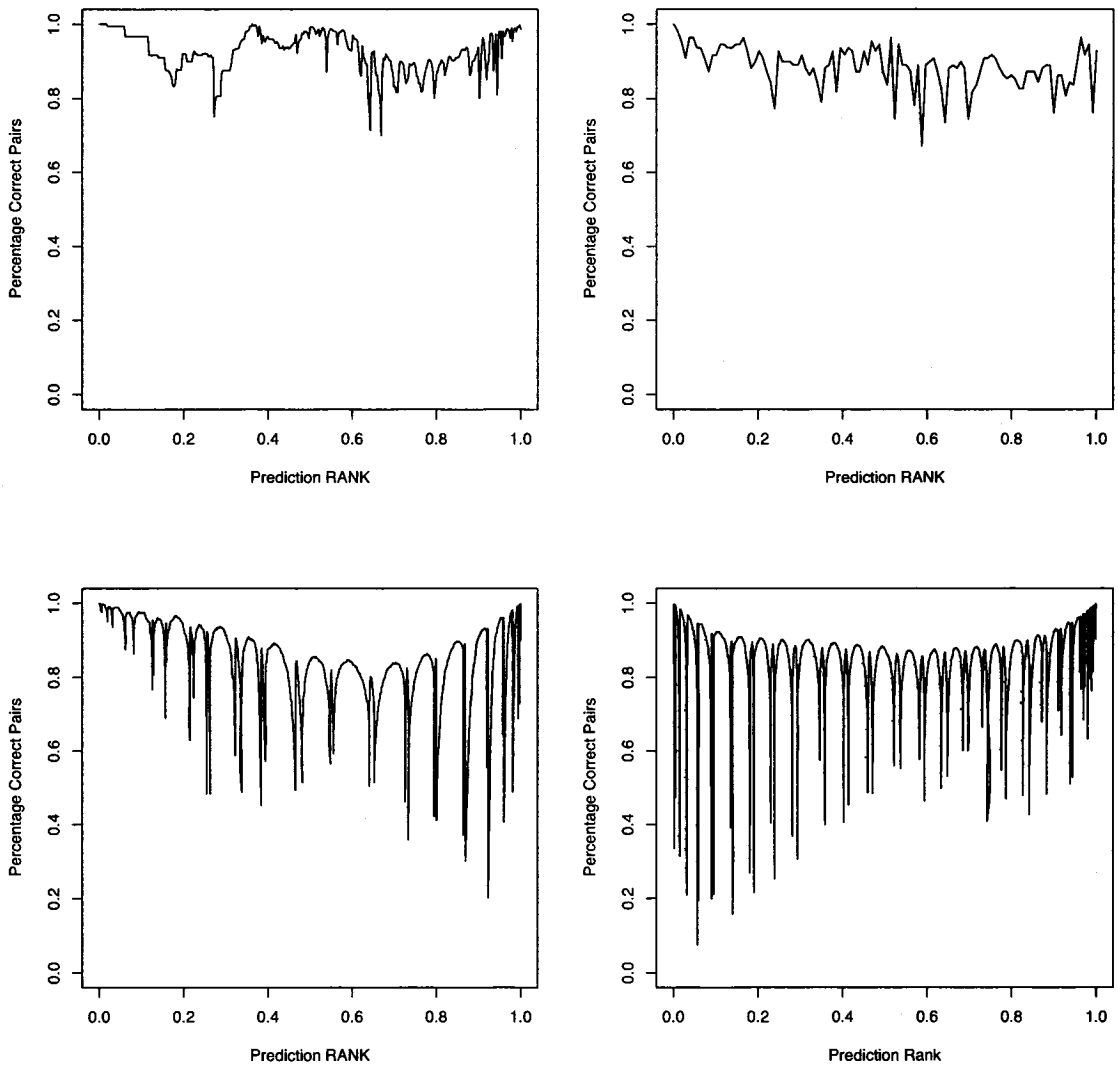

Figure D.11: Visual ranking performance of tRNA base model. Here DISTANCE is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

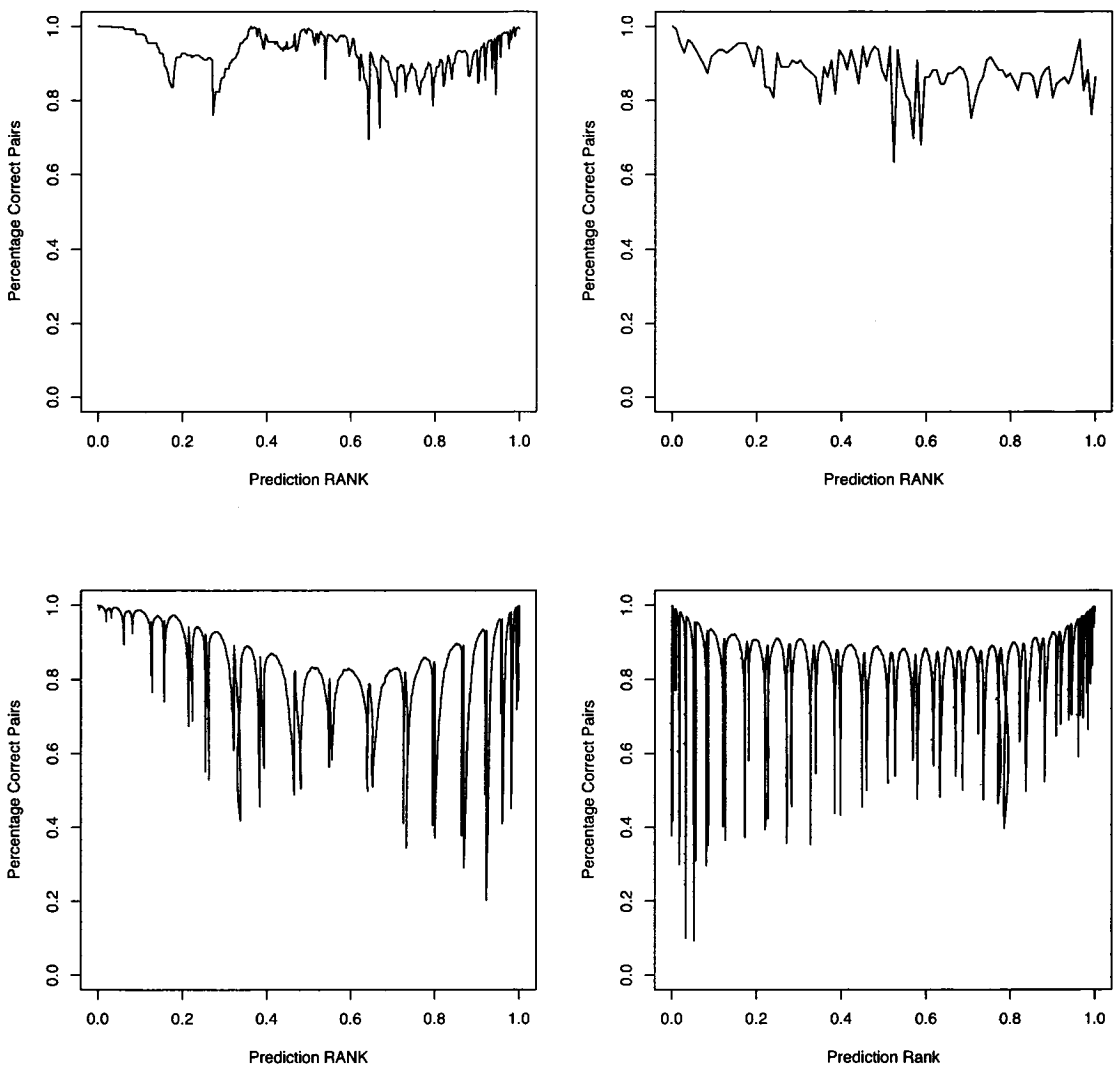

Figure D.12: Visual ranking performance of 5S base model. Here DISTANCE is used as the response variable. Starting from top left in clockwise direction: Performance of model a) on HSL3, b) on IRE, c) on 5S and d) on tRNA.

Appendix E

Best subset regression

This chapter illustrates the best subset regression for all data sets. For each table, **Pred set** represents the set of predictor variables used to build the models. For example, set 1 refers to building models using $N_{11}\,TBest$, $N_{11}\,TWorst$, $N_{11}\,TSum$, $TInfo$ and $\%GC$.

Pred set	Variables
1	$N_{11}\,TBest$, $N_{11}\,TWorst$, $N_{11}\,TSum$, $\%GC$, $TInfo$
2	$N_{12}\,TBest$, $N_{12}\,TWorst$, $N_{12}\,TSum$, $\%GC$
3	$N_{21}\,TBest$, $N_{21}\,TWorst$, $N_{21}\,TSum$, $\%GC$, $N_2\,TInfo$
4	$N_{22}\,TBest$, $N_{22}\,TWorst$, $N_{22}\,TSum$, $\%GC$
5	$N_{31}\,TBest$, $N_{31}\,TWorst$, $N_{31}\,TSum$, $\%GC$, $N_3\,TInfo$
6	$N_{32}\,TBest$, $N_{32}\,TWorst$, $N_{32}\,TSum$, $\%GC$

In each table, column **Pred** indicates the variables that have been selected to formulate the models using the selection criteria . A binary value of '1' indicates that the variable has been used in the order mentioned above and '0' otherwise. For example, a binary representation of **Pred** for HSL3 model build on AVGMCC using the first set of predictors is 11110 indicating that *TInfo* has not been used to build the model.

E.1 AVGMCC : Response

Table E.1: Best subset regression for HSL3

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	4	96.6	96.6	5.7	11110
2	4	96.5	96.4	5.0	1111
3	5	97.9	97.8	6.0	11111
4	4	97.6	97.6	5.0	1111
5	5	96.7	96.7	6.0	11111
6	4	96.7	96.6	5.0	1111

Table E.2: Best subset regression for IRE

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	68.0	67.5	6.0	11111
2	2	65.9	65.3	2.1	1011
3	3	65.5	64.6	3.1	01010
4	2	66.0	65.4	1.9	0101
5	3	62.3	61.2	3.3	10101
6	4	62.1	60.6	5.0	1111

Table E.3: Best subset regression for tRNA

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	72.2	72.2	6.0	11111
2	4	67.6	67.6	5.0	1111
3	4	70.4	70.4	4.6	01111
4	3	55.7	55.7	3.3	1011
5	4	68.2	68.2	4.5	10111
6	2	67.6	67.6	102.5	1010

Table E.4: Best subset regression for 5S

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	59.9	59.9	6.0	11111
2	4	51.6	51.6	5.0	1111
3	5	55.3	55.3	6.0	11111
4	4	31.4	31.4	5.0	1111
5	5	55.8	55.8	6.0	11111
6	4	43.8	43.8	5.0	1111

E.2 AVGACC : Response

Table E.5: Best subset regression for HSL3

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	4	96.1	96.1	4.6	11110
2	4	96.0	95.9	5.0	1111
3	5	97.4	97.3	6.0	11111
4	4	97.2	97.2	5.0	1111
5	5	96.5	96.5	6.0	11111
6	4	96.5	96.4	5.0	1111

Table E.6: Best subset regression for IRE

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	3	64.6	63.6	2.3	10011
2	2	63.5	62.8	2.0	1001
3	3	65.2	64.2	2.5	01011
4	2	65.7	65.0	1.3	0101
5	3	62.7	61.6	2.9	10011
6	2	61.3	60.6	3.0	1001

Table E.7: Best subset regression for tRNA

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	60.3	60.2	6.0	11111
2	4	51.8	51.8	5.0	1111
3	5	61.5	61.5	6.0	11111
4	2	40.6	40.6	1.9	0011
5	4	61.7	61.7	4.4	11011
6	3	53.6	53.5	4.5	1101

Table E.8: Best subset regression for 5S

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	43.8	43.8	6.0	11111
2	4	32.2	32.2	5.0	1111
3	5	43.3	43.3	6.0	11111
4	4	16.8	16.8	5.0	1111
5	5	45.6	45.6	6.0	11111
6	4	34.7	34.7	5.0	1111

E.3 DISTANCE : Response

Table E.9: Best subset regression for HSL3

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	80.4	80.2	6.0	11111
2	4	80.3	80.1	5.0	1111
3	5	94.3	94.2	6.0	11111
4	4	89.2	89.1	5.0	1111
5	4	80.8	80.6	4.6	11110
6	4	80.4	80.2	5.0	1111

Table E.10: Best subset regression for IRE

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	2	77.3	76.9	1.6	10010
2	2	76.7	76.3	1.4	1001
3	5	91.0	90.6	6.0	11111
4	3	77.5	76.9	3.8	1110
5	2	77.1	76.7	2.2	10010
6	2	76.4	76.0	1.7	1001

Table E.11: Best subset regression for tRNA

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	70.6	70.6	6.0	11111
2	4	70.6	70.6	5.0	1111
3	5	74.1	74.1	6.0	11111
4	4	73.9	73.9	5.0	1111
5	4	64.8	64.7	4.5	10111
6	3	64.2	64.2	3.2	1011

Table E.12: Best subset regression for 5S

Pred Set	Vars	R^2	R_a^2	Mallows C_m	Pred
1	5	71.6	71.6	6.0	11111
2	4	71.5	71.5	5.0	1111
3	5	73.8	73.8	5.0	11111
4	4	73.8	73.8	6.0	1111
5	5	69.6	69.6	6.0	11111
6	4	69.2	69.2	5.0	1111