An approach to selecting putative RNA motifs using MDL principle

Mohammad Anwar
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
Email: manwar@site.uottawa.ca

Marcel Turcotte
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
Email: turcotte@site.uottawa.ca

***Abstract*— The history of molecular biology is punctuated by a series of discoveries demonstrating the surprising breadth of biological roles of *ribonucleic acid* (RNA). An ensemble of evolutionary related RNA sequences believed to contain signals at sequence and structure level 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. For several decades, free energy minimization has been the most popular method for structure prediction. However, limitations of the free energy models as well as time complexity have prompted us to look for alternative approaches. We therefore, investigate another paradigm, minimum description length (MDL) encoding, for evaluating the significance of consensus motifs. Here, we evaluate motifs generated by Seed using the description length as a selection criteria. MDL scoring method was tested on four data sets of varying complexity. We found that the scoring method produces competing structures in comparison to the ones predicted with lowest free energy. The top rank motifs have high measures of positive predicted value to known motifs.**

Keywords: RNA, secondary structure, motif, minimum description length.

I. Introduction

RNA molecules are involved in a vast number of cellular functions, some of which include catalysis, splicing of premature messenger RNAs (mRNA), storage of genetic information and performing regulatory functions. To a great extent, the function of the RNA molecule is determined by its structure. RNA secondary structure prediction methods provide structural information that can serve as input constraint for solving the tertiary structure.

There are several automated methods for predicting secondary structures, either from a set of homologous RNA sequences or from a single sequence. The methods based on a set of sequences have the advantage of using evolutionary information, but have the disadvantage of requiring expensive computations. Also, these methods require an accurate alignment of the sequences, which is often not readily available. 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 [1], [2], [3]. Secondary structure prediction has also been approached by combining thermodynamics and comparative information. This class of algorithm attempts to fold and align the sequences simultaneously using a dynamic programming approach. The algorithms are both time and memory expensive, and restricted implementations are available in FOLDALIGN [4] and Dynalign [5].

As a result, the identification of RNA structures requires extensive human examination. We recently developed a novel method, Seed, to exhaustively search the space of RNA sequence and structure motifs using suffix arrays [6], [7]. The approach consists of two phases. First, the search space is generated from the seed sequence using suffix arrays. Secondly, suffix arrays are used to match secondary structure elements. The main steps of the Seed algorithm are as follows.

1) Select a seed sequence;
2) Construct the most specific motif;
3) General-to-specific search of the motif space;
4) Report the motifs.

The algorithm is built with user defined parameters to relax or restrict the size of the search space and accordingly the execution time. Seed identifies all the conserved RNA secondary structure motifs in a set of unaligned sequences. The search space is defined as the set of all the secondary structure motifs inducible from a seed sequence. Since Seed is exhaustive and independent of any scoring scheme, it provides an ideal environment to evaluate and study scoring methods to identify native folds.

For several decades now, free energy minimization has been the *de facto* method for studying RNA secondary structure. It suggests that the structure with the lowest free energy should be the most stable, and hence, the active fold. Also assumed is that the free energies of individual structural motifs are additive. Melting experiments are performed to determine the free energy parameters for small structures. Since the free energy can be decomposed into a sum of independent contributions, it can be solved exactly and efficiently when formulated as a dynamic programming problem [8].

Nearest neighbor model is the most common model used for predicting the stability of RNA. Figure 1 illustrates 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

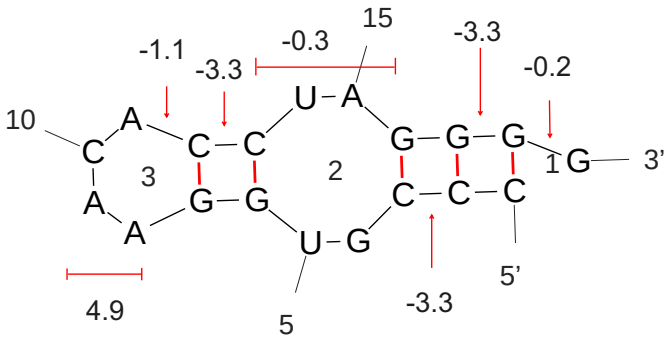

Fig. 1. Prediction of the conformational free energy for a conformation of **5'-CCCGUGGAACACCUAGGGG-3'**. The total free energy is the sum of each increment. The total free energy amounts to -5.5 kcal/mol.

region are calculated as the sum of the adjacent stacked pairs. For example, the consecutive CG base pairs contribute -3.3 kcal/mol each. Predicting 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 differ for hairpin, bulge and internal loops. Unpaired nucleotides in (internal) loops can provide favorable energy increments as either stacked nucleotides or as mismatched pairs.

Although the predictions have been on an average about 70% successful for sequences less than 100 nucleotides long [9], the performance of thermodynamics based methods is limited by thermodynamic models and parameters. There are several reasons why free energy minimization methods can fail.

- The lowest free energy conformation may not coincide with the native conformation. This can be due to experimental errors in determining the free energy parameters, errors due to the extrapolation of the parameters, or simply because there are numerous lowest free energy conformations, and it can be difficult to distinguish the native conformation from the others;
- Certain classes of RNA have more than one active structure. This is the case for several RNA regulatory elements termed riboswitches [10], [11], [12];
- No tertiary interactions are included in the model, no pseudoknotted structures and no interactions with cellular environment are considered.

The MDL statistical inference approach has proven to be highly valuable for numerous model selection problems. An MDL framework is outlined for tackling the problem of selecting the native fold from the rest. We present in this paper, an alternative approach to evaluate the significance of motifs such that the motifs ranking the highest are also biologically relevant. The paper is organized as follows. Section 2 outlines the basic concept of MDL and, in particular how to calculate the description length of the motifs. Section 3 presents the results obtained on the different data sets and Section 4 discusses the results and concludes.

II. Method

The purpose of a scoring function is to distinguish the biologically relevant motifs amongst an ensemble of motifs, that is, to approximate the biological meanings of the motifs in terms of mathematical function. In [13], we found that statistical (regression) model based on thermodynamic principles was effective to identify native folds. The attributes used captured the essential features of the motifs and helped us to summarize the data by estimating them. Herein, we investigate a pure statistical method to determine the native fold that avoids using *a priori* information.

The minimum description length principle [14], [15] was formulated in the context of computational complexity and coding theory. Its use for statistical model selection has developed over the last decades, largely as a result of the work by Rissanen (1978).

The secondary structure prediction is seen here as a scientific discovery process. Paraphrasing 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 [16]. 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.

We propose a simple encoding that follows [17] and is as close as possible to the encoding used by Seed. Let $T = \{T_1, T_2, ..., T_k\}$ be a set of k input sequences, such that $T \in \sum_{RNA}$, where $\sum_{RNA} = \{A, C, G, U\}$ is a 4-letter nucleotide alphabet. Without loss of generality, let us 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 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 significance 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 and $L(T_+) + L(T_-)$ defines the length of the data.

A. Encoding the Model

We use information theory in its fundamental form to measure the significance of different motifs [14], [15]. 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 probability distribution P is estimated using frequencies of nucleotides in the data set, $a_i \in \sum_{RNA}$ and n_{a_i} is the number of occurrences of a_i.

Figure 2 illustrates the encoding of the second part of the model M.

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

The given motif consists of two stems. It matches the sequence indicated by '`Seq`' at position 2. 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' tells 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 constructed. Bit value of '0' indicates that the base is 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 2, 6, 69, 0, 0, 0, 0, 0, 9, 12, 17, 0, 0, 0, 0, 27, 31, 17, 1, 0, 0, 1, 1, 49, 51, 17, 0, 0, 1,$ 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 by 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

$$\begin{aligned} dlen(M) = &- [\log_2 P_1(\$) + 4\log_2 P_1(s) + 4\log_2 P_1(l) \\ &+ 4\log_2 P_1(e) + 13\log_2 P_1(0) + 4\log_2 P_1(1)] \end{aligned}$$

B. 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. Figure 3 illustrates encoding a matching sequence with help of the motif.

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', as the bases with associated with the values '1' are recreated from the model, the offset k of the motif and the remaining part of the sequence. The 3'-end of the stem can be recreated 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 *G, A, U, U, G, G, U, U, C, C, G, G, C, k, G, U, A, A, A, U, U, G, G, U, C, A, C, U, G, U, C, A, A, A, A, G, A, U, G, G, U, U, C, G, A, G, C, C, C, C, C, G, C, C, A, G, $* 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 nucleotide types. 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

$$\begin{aligned} dlen(T_i, M) = &- [\log_2 P_2(\$) + \log_2 P_2(k) + 13\log_2 P_2(A) \\ &+ 16\log_2 P_2(G) + 15\log_2 P_2(C) + 13\log_2 P_2(U)] \end{aligned}$$

C. Significance of a Motif

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

$$\begin{aligned} w(M_j) = &\sum_{i=1}^{n} dlen(T_i) - (dlen(M_j) + \sum_{i=1}^{k} dlen(T_{+_i}, M) \\ &+ \sum_{i=1}^{n-k} dlen(T_{-_i})) \end{aligned}$$

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.

D. Data Sets

All the 3' UTR entries containing the keyword histone as well as an HSL3 feature were extracted from UTRdb release 19 [18]. A total of 28 sequences was obtained. The length of the sequences varies from 51 to $1,955$ nucleotides, with an average length of 701 nucleotides. See [7] for further details.

All the mammalian 5' UTR entries containing the keyword ferritin and a valid IRE motif were extracted from UTRdb release 19 [18]. A total of 14 sequences was obtained. The length of the sequences varies from 58 to 2,188 nucleotides, with an average length of 378 nucleotides.

```
          012345678901234567890123456789012345678901234567890123456789012345678901234
Seq     = GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGTCGCGC
Motif   =   NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
Struc   =   (((((..((((.........)))).(((((.......))))))).....(((..........)))))))
Bitmap  =   00000  0000              10011                 001
```

Fig. 2.

```
                    1         2         3         4         5         6         7
          01234567890123456789012345678901234567890123456789012345678901234567890123456
Seq     = GGAUUGUAGUUCAAUUGGUCAGAGCACCGCCCUGUCAAGGCGGAAGAUGCGGGUUCGAGCCCCGUCAGUCCCGCCAG
Motif   =  NNNNNNNNNNNNNNNNNNNNNNNNNCNNCCNNNNNNNGGNNGNNNNNNNGNNNNNNNNNNNCNNNNNNN
Struc   =  (((((..((((.........)))).(((((.......))))))).....(((..........)))))))
Bitmap  =  00000  0000              10011                 001
```

Fig. 3.

A tRNA dataset was assembled using a subset of the sequences from Masoumi and Turcotte [19]. Seven sequences having approximately the same length were used. These are generally challenging sequences for traditional approaches, such as MFOLD [1].

Similarly, a 5S dataset was assembled using a subset of the sequences from Masoumi and Turcotte [19]. Seven sequences having approximately the same length were used. These also are generally challenging sequences for traditional approaches, such as MFOLD.

E. Performance Measures

1) MCC, PPV and Sensitivity: 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. Offsets were not 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 [4]:

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

2) Ranking Statistics: The scoring method should be able to rank the motifs in decreasing order of their performance. In order to measure the performance of the scoring function, we used a rank based evaluation recently proposed by [20]. Rank based evaluation offer advantages over residue by being robust to outliers and providing insights about the local performance of the model. They also provide statistics that are similar to the commonly used correlation coefficients in data analysis (Spearman's ρ and Kendall's τ). Value of $\hat{\tau}$ and $\hat{\rho}$ range from -1 to $+1$, where $+1$ corresponds to a perfect model performance and -1 to making all possible errors.

III. Experimental Results

We have tested the performance of the MDL based scoring method proposed on four data sets. Different parameters were used in Seed to generate suitable stem-loop structures for each dataset. The search space contained motifs of varying degree of PPV and sensitivity. Table I shows the number of motifs found by Seed for each data set. The ranking statistics are shown in Table II. For 5S experiment, the number of motifs was large (365,505). Ten samples comprising of 10,000 motifs were picked randomly and averaged to calculate the statistics for 5S data set. Plots are drawn to show the correlation between MCC and description length of the motifs.

For single stem structure, HSL3, the top motif ranked by the scoring method has PPV and sensitivity of 100%. Also, for the second data set, IRE, we see that the high PPV/sensitivity motifs have lower description length; the top motif measuring at 100/92.7%. Figure 4 shows that the motifs 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 the spikes toward the right hand side of the diagram. The performance on HSL3 and IRE data set is also supported by high values of ranking statistics (see Table II).

We now focus on the performance of the MDL scoring method for complex structures. The motif with the lowest description length has an average PPV/sensitivity of 88.2/73.7%. We see a strong correlation between the different performance measures and the description length (see Figure 4). This also fits in nicely with the significance results of $\hat{\rho}$ (0.727 for AVGMCC). On 5S data, the average sensitivity is lower than that for the other datasets. However, the top ranked motif has a sensitivity of 40.4%, which is 7.8 percentage points less than the maximum observed sensitivity for the whole search space.

TABLE I

DETAILS OF THE EXECUTION OF SEED FOR ALL THE 4 EXPERIMENTS SHOWING THE NUMBER OF SEQUENCES (**Seqs**) PRESENT IN THE DATA SET, THE NUMBER OF MOTIFS DISCOVERED (**Motifs**), THE TOP MOTIF PICKED BY THE MDL BASED SCORING SCHEME AND ITS PERFORMANCE MEASURES. MFE IS THE SCORE OBTAINED USING FREE ENERGY FOR RANKING THE MOTIFS, MAX INDICATES THE LARGEST SCORE FOR THE WHOLE SEARCH SPACE (MAXIMUM ATTAINABLE SCORE).

Id	**Seqs**	**Motifs**	**AVGPPV**			**AVGSEN**			**AVGMCC**
			MDL	**MFE**	**MAX**	**MDL**	**MFE**	**MAX**	
HSL3	27	357	100.0	100.0	100.0	100.0	100.0	100.0	100.0
IRE	13	110	100.0	100.0	100.0	92.7	92.7	93.3	96.3
tRNA	7	5,518	88.2	96.0	100.0	73.7	71.2	76.2	80.6
5S	7	365,505	70.9	72.5	100.0	40.4	30.1	48.2	53.5

TABLE II

RANK STATISTICS. RANKING STATISTICS FOR THE SCORING METHOD FOR ALL FOUR EXPERIMENTS. AVGMCC HAS BEEN USED AS THE RESPONSE VARIABLE.

	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	0.639	0.584	0.770	0.277	0.394
AVGPPV	0.708	0.819	0.562	0.638	0.435	0.603	0.118	0.165
AVGMCC	0.686	0.813	0.561	0.639	0.539	0.727	0.198	0.288

Although the ranking statistics of 5S indicate poor ranking performance, the results for the top ranks motifs is optimal, in the sense that this is the motif with the highest MCC in our search space. The maximum PPV and sensitivity of the top 8 structurally distinct motifs are 86.4 and 50.0% respectively.

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.

IV. DISCUSSION AND CONCLUSION

In this work, we presented a scoring method for the software system Seed. The purpose was to distinguish the biologically relevant RNA secondary structure motifs from the rest. We evaluated our method on four different datasets having varying range of complexity. Two datasets we constructed consisted of selected members of UTRdb database, which contains the flanking 5' and 3' untranslated regions of genes. Others were assembled using a subset of sequences from [19].

We found that the method was able to identify high PPV motifs. For single stem structures, HSL3 and IRE, the predictions made have a high measure of PPV/sensitivity, often 100%. For complex structures, the scoring method could identify motifs with high PPV but lower sensitivity. All the results were supported by the ranking statistics.

The MDL method has its unique merits over thermodynamic based method as it successfully avoids *a priori* information. Since this method does not use any biological information, this work could be extended to compare pseudoknotted structures, for which no experimentally derived parameters are available. Future work includes developing a more theoretical/general framework expressed in terms of code length functions rather than a specific encoding.

We have introduced a scoring function formulation, implemented on the software Seed, designed to pick the best prediction(s) of RNA secondary structure motifs. The advantage of scoring functions is that they give us an intuitive mean by which to compare different motifs. This general approach of using MDL principle to evaluate RNA secondary structure can be useful for motif discovery.

REFERENCES

[1] 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.*, vol. 288, pp. 911–940, 1999.

[2] M. J. Serra and D. H. Turner, "Predicting thermodynamic properties of RNA," *Methods in Enzymology*, vol. 34, pp. 242–261, 1995.

[3] M. Zuker, "On finding all suboptimal foldings of an RNA molecule," *Science*, vol. 244, pp. 48–52, 1989.

[4] J. Gorodkin, S. L. Stricklin, and G. D. Stormo, "Discovering common stem-loop motifs in unaligned RNA sequences," *Nucl. Acids Res.*, vol. 29, no. 10, pp. 2135–2144, 2001.

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

[6] 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)*, pp. 1291–1298, 2005.

[7] M. Anwar, T. Nguyen, and M. Turcotte, "Identification of consensus RNA secondary structures using suffix arrays," *BMC Bioinformatics*, vol. 7, p. 244, 2006.

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

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

[10] E. C. Lai, "RNA sensors and riboswitches: Self-regulating messages," *Current Biology*, vol. 13, pp. R285–R291, 2003.

[11] E. Nudler and A. X. Mironov, "The riboswitch control of bacterial metabolism," *Trends Biol. Sci.*, vol. 29, no. 1, pp. 11–17, 2004.

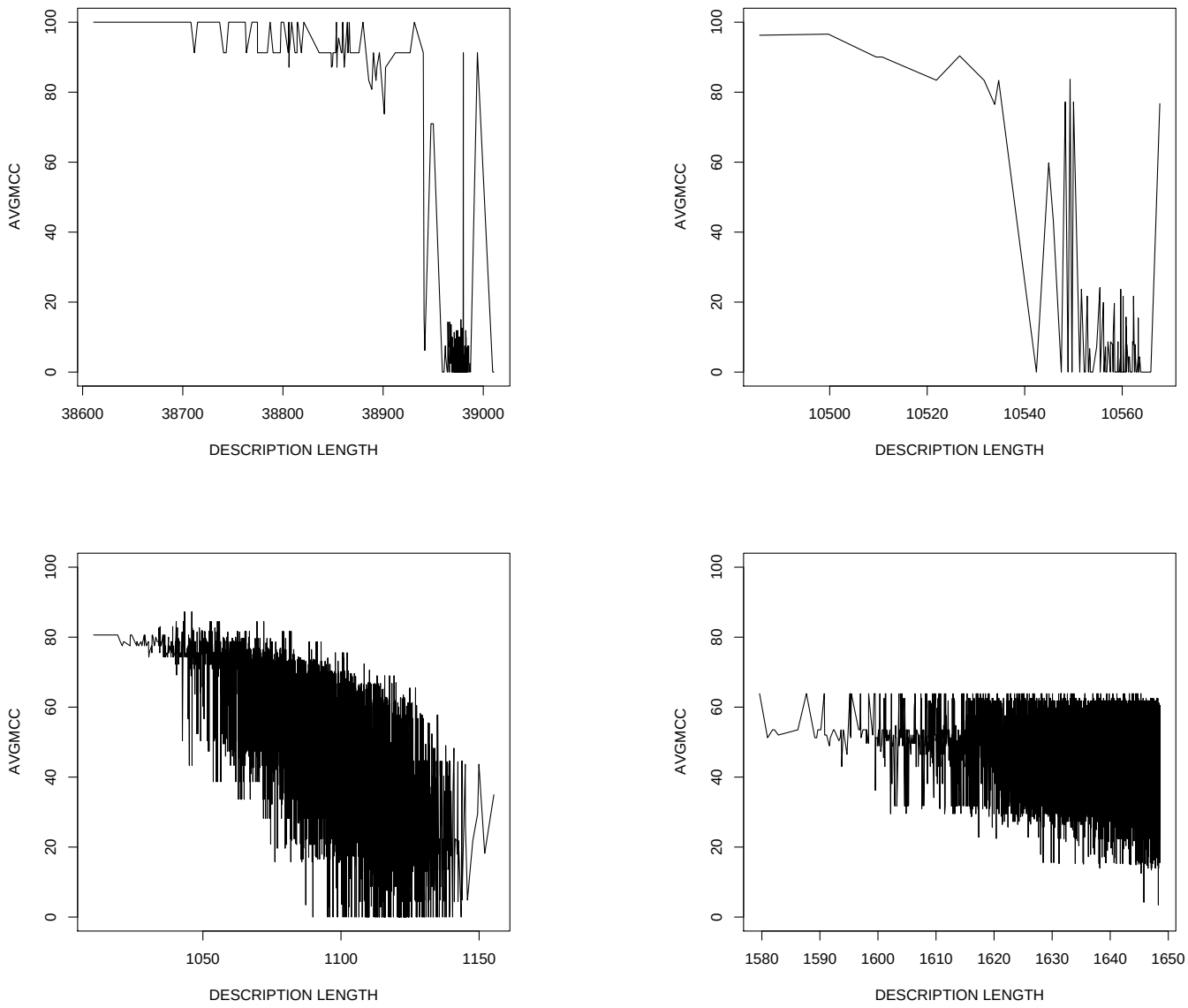

Fig. 4. Performance measure (MCC) as a function of description length. Starting from top left in clockwise direction, plots for HSL3, IRE, 5S and tRNA data sets.

[12] B. Voss, C. Meyer, and R. Giegerich, "Evaluating the predictability of conformational switching in RNA," *Bioinformatics*, vol. 20, no. 10, pp. 1573–1582, 2004.

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

[14] A. Brazma, I. Jonassen, E. Ukkonen, and J. Vilo, "Discovering patterns and subfamilies in biosequences," in *Fourth International Conference on Intelligent Systems for Molecular Biology*, 1996, pp. 34–43.

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

[16] M. A. P. P. D. Grünwald, I. J. Myung, *Advances in Minimum Description Length Theory and Applications*. MIT Press, 2003.

[17] J. T.-L. Wang, Q. Ma, D. Shasha, and C. H. Wu, "Application of neural networks to biological data mining: a case study in protein sequence classification," in *Knowledge Discovery and Data Mining*, 2000, pp. 305–309.

[18] G. Pesole *et al.*, "UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 335–340, 2002.

[19] B. Masoumi and M. Turcotte, "Simultaneous alignment and structure prediction of three RNA sequences," *International Journal of Bioinformatics Research and Applications*, vol. 1, no. 2, pp. 230–245, 2005.

[20] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-based evaluation of regression models," in *The Fifth IEEE International Conference on Data Mining (ICDM '05)*, Houston, Texas, 2005, pp. 370–377.