Chapter 6. Finding RNA motifs in genomes

In chapter 4 of this thesis, a new RNA-motif modelling tool based on the functional-site modelling tool -- Eponine was created. This new tool is particularly designed for modelling functional sites that may be associated with local RNA motifs. In addition, the models so trained should be capable of discriminating ncRNAs in genomes. Unlike other comparative algorithms that can be used for genome-wide ncRNA finding, this tool is not dependent on sequence alignments. Thus this tool may potentially provide an alternative approach for genome-wide ncRNA finding.

In this chapter, I assessed the capability of the Eponine RNA-motif extension. Two types of capabilities are of interest:

- The capability of the Eponine RNA-motif extension to find the consensus RNA motifs, consisting of both primary-sequence and secondary-structure motifs, in a set of transcripts
- The capability of the models so learned to discriminate a particular type of ncRNAs in genomes

Three types of different ncRNAs with distinct structural features were used to perform the capability assessment. The modelling of the mammalian tRNAs is discussed in subsection 6.1.1. The modelling of the *rho*-independent transcription terminators of bacteria is discussed in subsection 6.1.2. The modelling of the pseudoknots in the 3' untranslated regions (UTR) of viral genes is discussed in subsection 6.1.3.

6.1. Using the Eponine RNA-motif extension

6.1.1. Modelling RNA-motifs of mammalian tRNAs

The set of mammalian tRNAs was chosen as the starting case for assessing the capability of the Eponine RNA-motif extension, since the consensus clover-leaf secondary structure features of tRNAs have been studied for decades. tRNAs are also widely used as a data set for evaluating the performances of RNA secondary-structure prediction programs and ncRNA classifiers.

In this subsection, further assessment is made of the performances of the stringent and the fast modes of the Eponine RNA-motif extension (for definitions of the stringent mode and the fast mode, see Figure 4-3 and subsection 4.2.2.1.). It was shown that when identifying the canonical secondary structures of tRNAs, the stringent mode was better than the fast mode (see Table 4-1). An issue which was not investigated is the effect of using different structure-scanning modes on performance in the context of discriminating ncRNAs in genomes. If the models trained using the fast mode do not perform significantly worse than the models trained using the stringent mode, maybe the fast mode could be sufficient for the purpose of discriminating ncRNAs in genomes.

Consequently, there are two purposes of this subsection. Firstly, the performances of pure structural-motif models trained using the stringent mode and the fast mode, respectively, are compared. Secondly, I demonstrate that the Eponine RNA-motif extension can be used to train a discrimination model consisting of both primary-sequence patterns and RNA secondary-structure motifs.

6.1.1.1. Materials and methods

6.1.1.1.1. Recruiting the genomic sequences for training and testing

The sets of human tRNA genes created in section 5.1, chapter 5, were used for assessing the capabilities of the Eponine RNA-motif extension. The human tRNAs of group 1 were used for training models, and the tRNAs of group 2 were used for testing the performances of these trained models (Table 6-1, positive sequences). In order to realize the effect of using genomic sequences on modelling consensus RNA motifs, the flanking regions of human tRNA genes were included. The first base of the cloverleaf-like structure of each tRNA was used as the anchoring point; 100 bases upstream and 150 bases downstream with respect to the anchoring point in each human tRNA gene were retrieved. Two thousand random sequences and ten thousand random sequences were sampled from the human genome as negative training sequences and negative test sequences, respectively (Table 6-1, negative sequences). The human genome assembly used for random sampling was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/). These random sequences were 250 bases in length.

	Positive sequences	Negative sequences
Training data	200 genomic sequences of human tRNAs (group 1)	2000 random sequences from the human genome
Test data	167 genomic sequences of human tRNAs (group 2)	10,000 random sequences from the human genome

Table 6-1. The training and test data sets for modelling the human tRNAs

6.1.1.1.2. Determination of the performance of EAR models against the test data set

The training sequences described in the previous subsection were used to train the Eponine Anchored RNA-motif models (the EAR models, see subsection 4.2.2.3.1, chapter 4). When evaluating the performance of trained models, the 100th base of each test sequence was taken as the anchoring point. A true positive was determined if any region within 5 bases away

from the anchoring point of a positive sequence was predicted as a hit. A false positive was determined if any region within 5 bases away from the anchoring point of a negative sequence was predicted as a hit.

6.1.1.1.3. Setting the parameters of the Eponine RNA-motif extension

The size of windowed regions for predicting the local RNA structural motifs was set to 50 bases when running the Eponine RNA-motif extension. As a result, only the base pairs within each windowed region of 50 bases would be considered in the trained models. The windows were limited to 50 bases in this subsection for several reasons. Firstly, finding a consensus global RNA structure in a set of sequences is not the objective of designing the Eponine RNA-motif extension. It is instead designed to use consensus local RNA motifs for discriminating a particular type of ncRNAs in genomes. Secondly, one purpose of this subsection is to compare the performances of different RNA-motif scanning modes, *i.e.* the stringent mode and the fast mode (for the details of these two modes, see section 4.2, chapter 4). If evidence strongly suggests that long-range canonical base pairs are essential for discriminating a particular type of ncRNAs, the size of windowed regions can certainly be increased at the cost of computational time.

6.1.1.2. Results

6.1.1.2.1. Pure secondary-structure models of human tRNAs

By using the stringent mode, an EAR model consisting of eight hairpins was trained (Table 6-2 and Figure 6-1 A). While it might seem that too many hairpins were found, the eight hairpins can be grouped into five distinctly positioned hairpins, namely, hairpins that start at 10^{th}, 15^{th}, 27^{th}, 49^{th}, and 59^{th} positions respectively in tRNA molecules. Among these predicted consensus hairpins, hairpins that start at 10^{th}, 27^{th}, and 49^{th} positions clearly correspond to three well-known hairpins, D arm, anticodon arm, and T arm, respectively in tRNAs. The hairpin that starts at 59^{th} position can be viewed as a shifted T arm, because some

tRNA genes contain intronic sequences and the distance between the first base of cloverleaf-like structure and T arm is therefore longer than that in the tRNAs without introns.

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
2.05	10	0.48	10	1.2	3	0.7
2.13	10	0.41	8	0.5	4	0.2
1.83	15	0.33	6	2.6	3	0.3
2.51	26	1.07	9	0.0	4	0.2
2.32	27	1.96	7	0.1	5	0.5
2.08	49	1.00	7	1.0	3	0.6
1.54	50	10.14	7	0.3	5	0.1
1.68	59	0.00	5	1.0	4	0.2

Table 6-2. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the stringent mode for locating local hairpins

The titles, "Weight", "Position", and "Width", are used as described in Figure 5-3. "Loop size" is the mean of the discrete Gaussian distribution used to model a loop region. "Stem size" is the mean of the discrete Gaussian distribution used to model a stem region.

A fast-mode EAR model consisting of ten hairpins was also trained (Table 6-3). Just as the hairpin groups in the stringent-mode EAR model, these ten hairpins can be categorized into four distinctly positioned hairpin groups, namely, hairpins that start at 3^{rd}, 10^{th}, 27^{th}, and 47^{th} positions respectively in tRNA molecules. The latter three correspond to three well-known hairpins, D arm, anticodon arm, and T arm respectively in tRNA molecules.

It seems that the model trained using the stringent mode for locating local hairpins is slightly simpler than the model trained by using the fast mode, although most likely this is caused by chance. In the current implementation of the Eponine RNA-motif extension, similar sub-models of individual hairpins are not merged and in different training runs the numbers of hairpins found may differ. In brief, the difference between the numbers of hairpins found by

two models does not suggest that one of the models may be better than the other one.

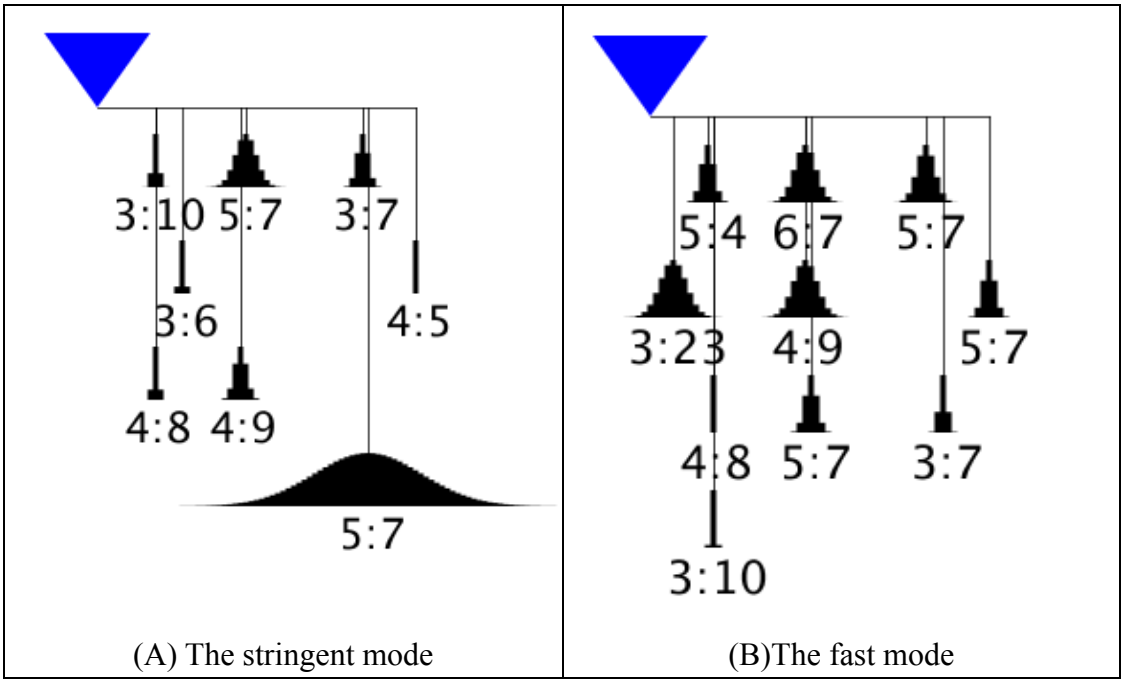
Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
1.97	3	2.27	23	0.7	3	0.1
2.69	9	1.05	4	0.2	5	0.3
2.78	10	0.08	8	0.7	4	0.1
2.34	10	0.23	10	0.8	3	0.5
1.51	26	1.83	7	0.1	6	1.1
1.35	26	2.06	9	0.0	4	0.1
1.82	27	1.00	7	0.7	5	0.1
2.89	47	1.52	7	0.0	5	0.2
1.73	50	0.59	7	2.9	3	0.0
1.38	58	1.00	7	1.2	5	0.0

Table 6-3. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the fast mode for locating local hairpins

The titles used in this table follow the convention of Figure 5-3 and Table 6-2.

Evaluating the performances of the fast mode and the stringent mode

By using the test data set recruited as described in 6.1.1.1.1. the performances of the models trained respectively using the fast mode and the stringent mode of the Eponine RNA-motif extension were evaluated. The results suggest that the performance of the fast mode can be as good as that of the stringent mode (Figure 6-4, fast mode and stringent mode). Although using the fast mode risks missing important hairpins, it can still be used for finding consensus RNA structural motifs in sequences when sufficient positive sequences are used for training. Since by using the fast mode the CPU time is about 40%-60% of the time taken by using the stringent mode, all models in the following were trained by using the fast mode, unless otherwise indicated.



(A) The stringent mode (B)The fast mode

Figure 6-1. Two Eponine anchored RNA structural models for mammalian tRNAs

The diagrams were prepared following the convention used in Figure 5-2, except that the motifs shown here are RNA structural motifs. The constraints drawn with two numbers under them correspond to RNA hairpins. These numbers are used to describe the dimension of a consensus hairpin. Each dimension consists of the stem size and the loop size that are separated by a colon. For example, in the right most hairpin in (A), 4:5 means that the size of this stem is 4 base pairs and the length of the loop is 5 bases.

6.1.1.2.2. A mixed primary-sequence and RNA secondary-structure model

Here, the capability of the Eponine RNA-motif extension to model both primary-sequence and RNA secondary-structure motifs was evaluated by using the human tRNAs recruited as described in 6.1.1.1.1. The results reveal that the EAR model is capable of finding both primary-sequence and RNA secondary-structure motifs of tRNAs (Figure 6-2). Such models that contain both primary-sequence and RNA structural motifs are referred to as mixed models in this thesis.

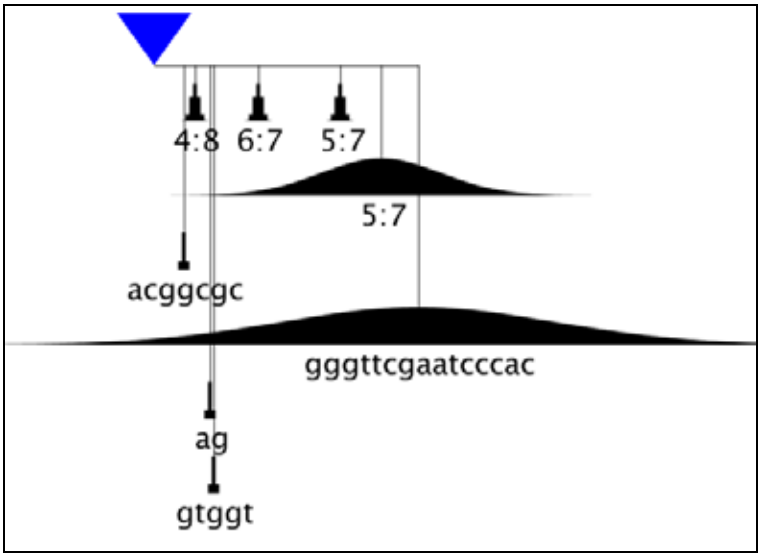


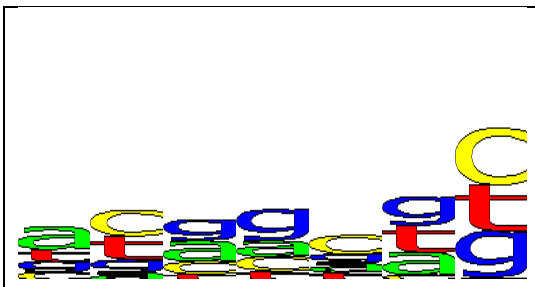
Figure 6-2. An Eponine anchored and mixed (primary-sequence and RNA structural) model

This figure is drawn following the convention used in Figure 6-1.

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
5.06	8	0.45	Not available (a PWM of 7 columns)			
1.97	11	1.00	8	0.15	4	0.01
1.76	15	0.45	Not available (a PWM of 2 columns)			
4.15	16	0.45	Not available (a PWM of 5 columns)			
1.48	28	1.00	7	0.46	6	2.39
2.19	50	1.00	7	0.39	5	0.52
2.40	61	16.11	7	0.04	5	0.05
31.88	71	36.50	Not available (a PWM of 15 columns)			

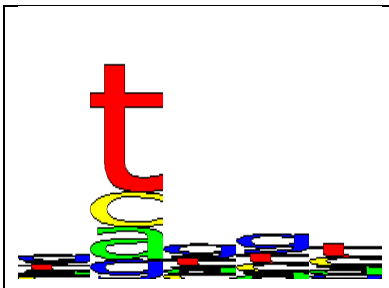
Table 6-4. The trained parameters of the EAS mixed model presented in Figure 6-2

The titles used in this table follow the convention of Figure 5-3 and Table 6-2

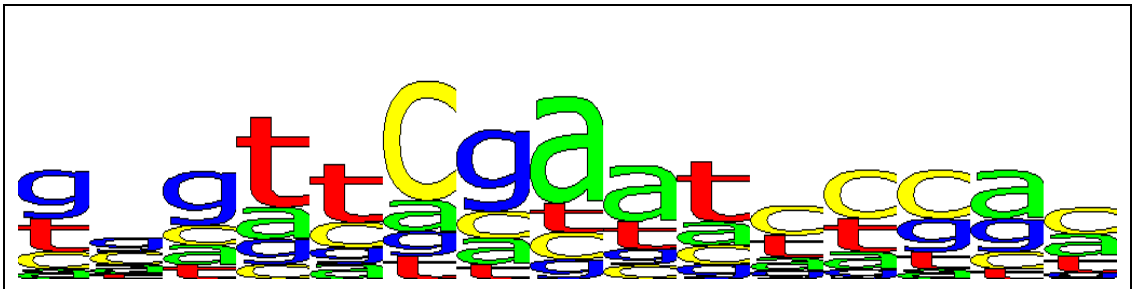


Position: 8

Position: 15



Position: 16



Position: 71

Figure 6-3. The sequence logos of position-constrained motif matrices in the Eponine EAS mixed model presented in Figure 6-2 and Table 6-4.

"Position" corresponds to the "Position" column in Table 6-4.

Evaluating the performances of the mixed model of human tRNAs

The capability of the trained mixed model to differentiate human tRNAs from random genomic sequences was also evaluated using the test data set recruited as described in 6.1.1.1.1. The results reveal that a mixed model ("mixed model, fast mode", Figure 6-4) can perform better than models consisting of only RNA structural motifs ("structure-only" models,

Figure 6-4). For discriminating tRNAs in the human genome, the false positive rate of the mixed model should be much lower than that of the models consisting of only RNA secondary-structure motifs (comparing the "structural-only" models with the mixed model, Figure 6-4).

For comparison, a pure primary-sequence model, which did not consist of RNA motifs, was trained taking the training data set as described in 6.1.1.1.1. The performance of this pure primary-sequence model was also evaluated using the test data set recruited as described in 6.1.1.1.1. However, in this evaluation, the accuracy of the mixed model for human tRNAs ("mixed model, fast mode", Figure 6-4) was not as good as this pure primary-sequence model ("pure primary-sequence model", Figure 6-4) when the coverage (sensitivity) was set to be higher than 90%. There were 10 false positives predicted by the mixed model, while only 2 false positives were found by using the pure primary-sequence model.

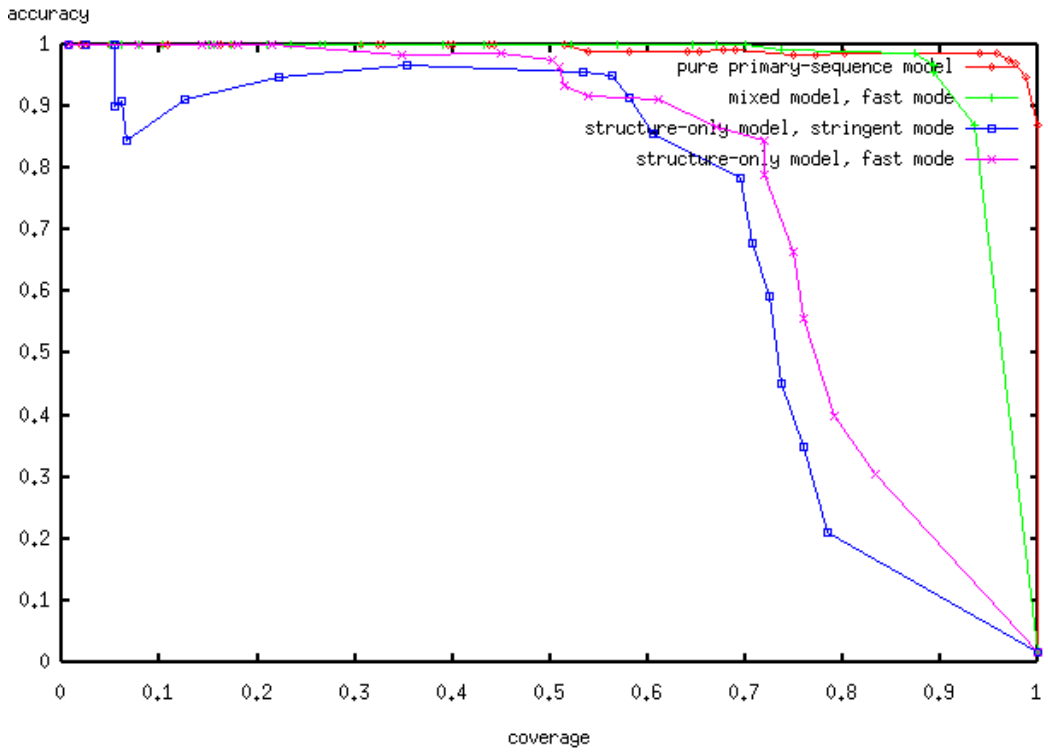


Figure 6-4. Comparison of performances among models trained by different modes for classifying human tRNA genes from random genomic sequences

6.1.1.3. The false positives predicted by using the mixed model

To explore why a mixed model discovered more false positives, the features of the 10 high-scoring false positives were examined in detail. The conservation of the internal promoter in each sequence, and the conservation of local RNA motifs corresponding to the D arm, anticodon arm, and T arm in the canonical tRNA clover-leaf like structures were evaluated.

The results reveal that most of the false positives predicted by the mixed model of human tRNAs contain only a subset of the motifs in the canonical tRNA structures (Table 6-5). In summary these false positives can be characterised as:

- A sequence with a strong internal promoter (as determined by eufindtRNA) can be identified as a tRNA.
- A sequence with a partial set of weak motifs, either in a combination of a weak internal promoter and a local RNA structural motif, or in a combination of two or more local RNA structural motifs, can be identified as a tRNA.
- Most of the false positives overlap with repetitive elements.

Serial ID	Internal promoters[1]	D arm	anticodon arm	T arm	Repeat
1	+	-	-	-	SINE/MIR
2	-	-	-	+[2]	LINE/L1
3	+	-	+	-	LINE/L1
4	-	-	+ (ss) (offset)	+	SINE/MIR
5	+	-	-	-	LTR/MaLR SINE/Alu
6	+	-	+ (ss)	+ (offset)	SINE/Alu
7	-	-	+ (ss) (offset)	+ (offset)	LINE/L1
8	-	-	+ (ss)	+ (ls)	LTR/MaLR SINE/Alu
9	+	-	-	+ (offset)	LINE/L1
10	+	-	-	+	(not available)

Table 6-5. The high-scoring false positives predicted by using the mixed model of human tRNAs

[1]: the internal promoters were determined by using eufindtRNA with a relaxed parameter set

[2]: there is an additional hairpin at the 3' side of the T arm. This additional hairpin also contributes to the final score.

(ss): a stem which is smaller than the corresponding canonical local RNA motif.

(ls): a stem which is longer than the corresponding canonical local RNA motif.

(offset): a hairpin is a few bases away from the best positions in the canonical tRNA structure.

(not available): not overlapping with repetitive elements

Due to the scoring scheme used in Eponine, these findings are not really surprising. Given a GLM-based RNA-motif model such as the mixed model of human tRNAs, the final score of a genomic locus is actually a transformed weighted sum of PWM scores and RM scores. Thus, a mixed model consisting of many local motifs may be apt to identify truncated ncRNAs and other ncRNA-derived sequences. In fact, such behaviour is not unique to the Eponine RNA-motif extension. A similar observation has been made in the development of tRNAscanSE (Lowe and Eddy 1997), where the tRNA covariance model was shown to discover some truncated tRNAs and tRNA-derived SINEs which could not be identified by using promoter-based methods (such as eufindtRNA), and hierarchical and rule-based systems (*e.g.* tRNAscanSE) for genome-wide tRNA finding.

6.1.2. Modelling *rho*-independent transcription termination

The modelling of human tRNA genes partially demonstrates the capability of the Eponine RNA-motif extension. Since many existing ncRNA-finding algorithms have also been shown to be capable of detecting the cloverleaf-like structures, the result of the modelling of human tRNAs only reveals that the Eponine RNA-motif extension has a function similar to other tools. Consequently, in this subsection, a more difficult case (for reasons see the discussion in the next two paragraphs), the *rho*-independent transcription terminators, was used to evaluate the capability of the Eponine RNA-motif extension.

The *rho*-independent transcription terminator, which consists of both primary-sequence and RNA structural motifs, is an important functional element for regulating the transcription termination of bacterial genes (Uptain and Chamberlin 1997). Unlike modelling tRNA genes, finding *rho*-independent transcription terminators is a topic that has received less investigation. Apparently, only *ad hoc* algorithms can find *rho*-independent transcription terminators in the bacterial genomes (d'Aubenton Carafa et al. 1990; Ermolaeva et al. 2000; Lesnik et al. 2001; de Hoon et al. 2005). Up to this point, no general-purpose RNA-motif finding algorithms have been used to find the consensus RNA motifs in these regions of transcription termination.

One reason that makes *rho*-independent termination signals an unpopular data set is that the boundaries of *rho*-independent termination signals are not so well defined as known ncRNA genes (such as tRNA genes). It is difficult to adequately align these regions. The identities of pairwise alignments of the regions around transcription termination sites are generally low. Fewer than 0.5% of pairwise alignments have identities greater than 60% (data not shown), if the alignments are generated by randomly choosing raw sequences that have been used by de Hoon *et al*. (de Hoon et al. 2005). Whether these low-identity alignments can reveal the structural relations among sequences cannot be confidently determined. However,

as has been discussed previously (see section 2.1, chapter 2, and section 4.2, chapter 4), most existing algorithms would not be expected to have good performance in finding structural signals in such data set.

Some *ad hoc* algorithms were claimed to have high specificity and high sensitivity in detecting *rho*-independent transcription terminators. However, there must be some doubt about the generality of such results given the training and optimisation processes used. Firstly, some models were actually tested with exactly the same sequences that have been used for training respective models (d'Aubenton Carafa et al. 1990; Lesnik et al. 2001; de Hoon et al. 2005). These models may be over fitted and unable to generalise to new data, something that has not been tested for because of the use of a non-independent test data set. Secondly, some algorithms discard all predictions in intragenic regions (Ermolaeva et al. 2000), even though the scores of these predictions exceed the computationally defined threshold. The eradication of this major source of false positives makes it impossible to properly estimate the accuracy and specificity of the predictions made by these algorithms.

6.1.2.1. Materials and methods

6.1.2.1.1. The data sets for training and testing the Eponine anchored RNA-motif model

In order to train and test the EAR models for *rho*-independent transcription terminators, 423 transcription terminators that have been used by de Hoon *et al*. (de Hoon et al. 2005) were divided into two data sets for training and testing respectively. Each sequence consists of 20 bases upstream and 50 bases downstream of the respective transcription termination site annotated by Hoon *et al*. (de Hoon et al. 2005).

Two sets of 2,000 negative sequences for training and testing models, respectively, were randomly taken from the *B. subtilis* genome (GenBank accession number: AL009126). These negative sequences were 70 bases in length.

6.1.2.1.2. Determination of the performance of EAR models against the test data set

When evaluating the performance of EAR models for *rho*-independent transcription terminators against the test data set, the 20th base of each sequence was taken as the anchoring point. A true positive was determined if any region within 5 bases away from the anchoring point of a positive sequence was predicted as a hit. A false positive was determined if any region within 5 bases away from the anchoring point of a negative sequence was predicted as a hit.

6.1.2.1.3. Scanning for rho-independent transcription terminators in genomes

When an EAR model for *rho*-independent transcription terminators was used to scan genomes, both strands of genomes were scanned. Each position in a genome can be the first base of a *rho*-independent transcription terminator. Consecutive hits would be clustered together if all of their scores were higher than a particular threshold and considered as a single prediction.

Determination of putative terminators of genes

For each gene, if a predicted *rho*-independent TTS on the same strand is within the range starting from 50 bases upstream of the stop codon, continuing till the 500 bases downstream of the stop codon, this TTS is considered as a putative terminator, unless if this TTS is within the coding region of the next gene. If there were more than one candidate hit for a particular gene, the one that was closer to the stop codon was used.

Determination of intragenic terminators

If an intragenic predicted hit is more than 50 bases from the stop codon of a gene, it is regarded as a true intragenic hit.

6.1.2.1.4. The data set for training and testing the Eponine Windowed RNA-motif model

To assess the capability of the Eponine Windowed RNA-motif model (the EWR model,

see subsection 4.2.2.3.2, chapter 4) to find consensus RNA motifs in a set of sequences where no reference points are known, a set of 423 *B. subtilis* genomic sequences that contain *rho*-independent transcription terminators was prepared. In order to make the assessment more challenging, the positions of *rho*-independent transcription terminators in respective sequences were randomly distributed between 1 and 100 (Figure 6-5). These sequences were randomly divided into a training set (212 sequences) and a test set (211 sequences). The negative sequences recruited for training and testing models were the same as described in subsection 6.1.2.1.1.

When evaluating the performance of EWR models for *rho*-independent transcription terminators, a true positive was determined if any position in a positive sequence was predicted as a hit. A false positive was determined if any position in a negative test sequence was predicted as a hit.

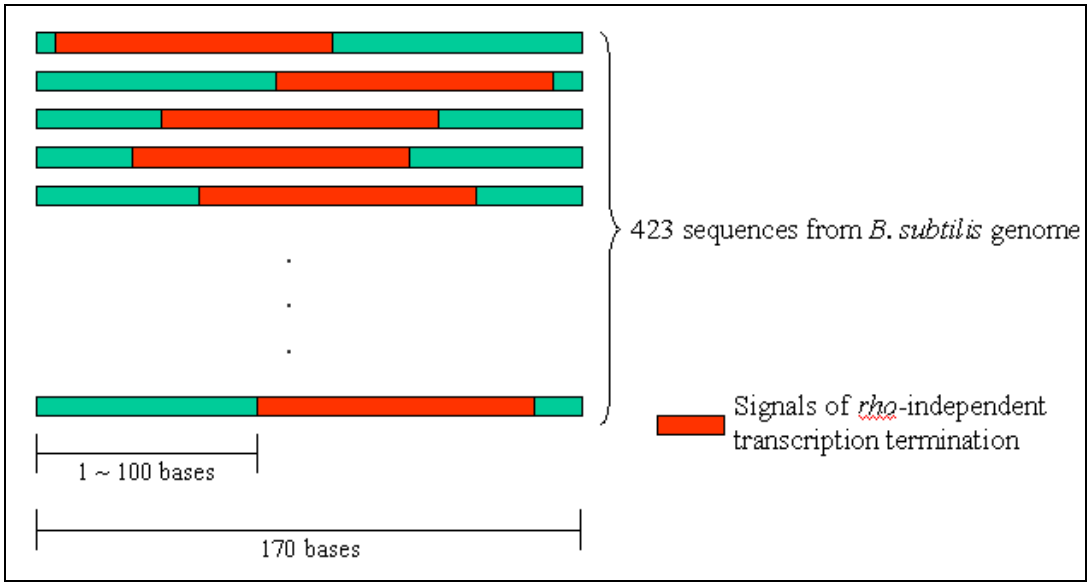


Figure 6-5. Preparation of a set of unanchored sequences that contain *rho*-independent transcription terminators at random positions

6.1.2.2. Results

6.1.2.2.1. The Eponine anchored RNA-motif model (EAR model)

The EAR mixed model for the *rho*-independent transcription terminators of *B. subtilis* consisted of five motifs (see Table 6-6 and Figure 6-6). This model is basically consistent with the current knowledge of the composition of the *rho*-independent terminators (For details see Lesnik et al. 2001), where the first two motifs (weights 0.85 and 5.30, Table 6-6) correspond to an A-region (adenosine-rich region); and a stable hairpin (weight 6.03, Table 6-6) is followed by a T-region (weight 13.62, Table 6-6) (thymidine-rich region in genome, corresponding to uridine-rich region in transcripts). An additional motif is at positive 5 (weight 4.17, Table 6-6). However, its importance is not clearly understood. Since it overlaps with the hairpin motif it may be capturing sequences preference within the hairpin of *rho*-independent transcription terminators. The Eponine sub-model for the hairpin of *rho*-independent transcription terminators is at position 5 (weight 6.03, Table 6-6); the stem size is 9 base pairs in length and the loop size is 12 bases in length. The standard deviation for the distribution of loop size is 16.5 bases, which is obviously larger than the mean loop size (12, Table 6-6). The heavy tail in the distribution of the loop size is consistent with the previous models of the *rho*-independent terminators of either *E. coli* or *B. subtilis* (d'Aubenton Carafa et al. 1990; Ermolaeva et al. 2000; Lesnik et al. 2001; de Hoon et al. 2005).

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0.85	-3	0.60	Not available (a PWM of 3 columns)			
5.30	1	0.63	Not available (a PWM of 5 columns)			
6.03	5	4.46	12	16.5	9	2.13
4.17	5	1.38	Not available (a PWM of 4 columns)			
13.62	29	17.96	Not available (a PWM of 7 columns)			

Table 6-6. The trained parameters of an EAR model for *bacillus rho*-independent transcription terminators

The titles used in this table follow the convention of Table 6-4.

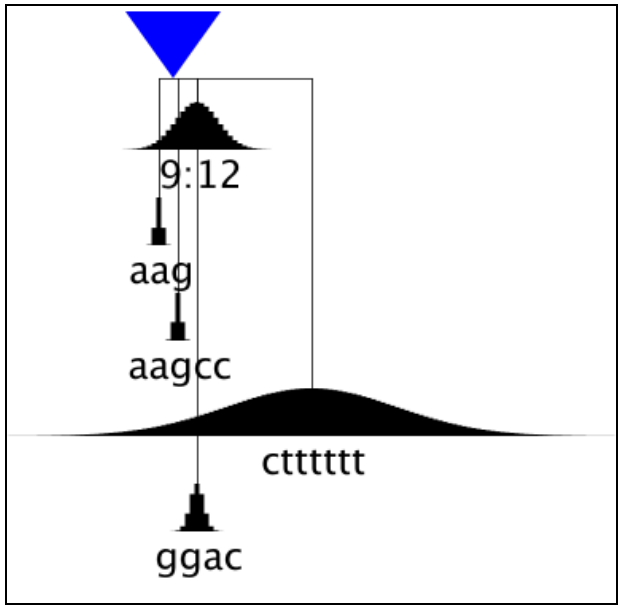
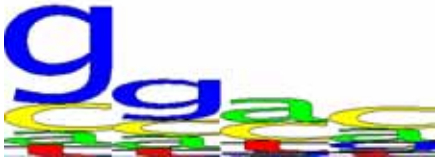


Figure 6-6. An EAR model for *rho*-independent transcription terminators

This figure is drawn following the convention used in Figure 6-1.

Position: -3

Position: 1



Position: 5

Position: 29

Figure 6-7. The sequence logos of the position-constrained motif matrices presented in Figure 6-6 and Table 6-6

"Position" corresponds to "Position" column in Table 6-6.

For comparison, a pure primary-sequence model, which did not consist of RNA motifs, was trained taking the training data set as described in 6.1.2.1.1. A structure-only model, which did not consist of primary-sequence motifs, was also trained using the same data set. C-A plots of different models for the *rho*-independent transcription terminators were calculated using the test data set of 211 positive sequences and 2000 negative sequences. The result reveals that the performance of the mixed model (see Table 6-6 and Figure 6-6) is better than that of the pure primary-sequence and structure-only models (Figure 6-8).

Discriminating the *rho*-independent transcription terminators in real bacterial genomes

In order to further assess the performances of the EAR mixed model and other algorithms, the sensitivities and specificities were estimated by using the result of scanning the full-length

genomic sequences of *B. subtilis* and *E. coli* K-12 (GenBank accession number: U00096) (Table 6-7). The predictions that overlap with experimentally verified *rho*-independent transcription terminators were counted as true positives. In order to avoid bias in the evaluation, only known terminators that were not used for training the respective algorithms/models were used to estimate sensitivities. Predictions in intragenic regions were taken as false positives for estimating false positive rates. Although some of the *rho*-independent transcription terminators may possibly reside in intragenic regions, the location distribution of true terminators should be greatly biased towards intergenic regions. While it is likely that some of the predictions that fall in intergenic regions are false positives, the ratio of intragenic predictions over all predictions provide at least an estimate of the false positive rate.

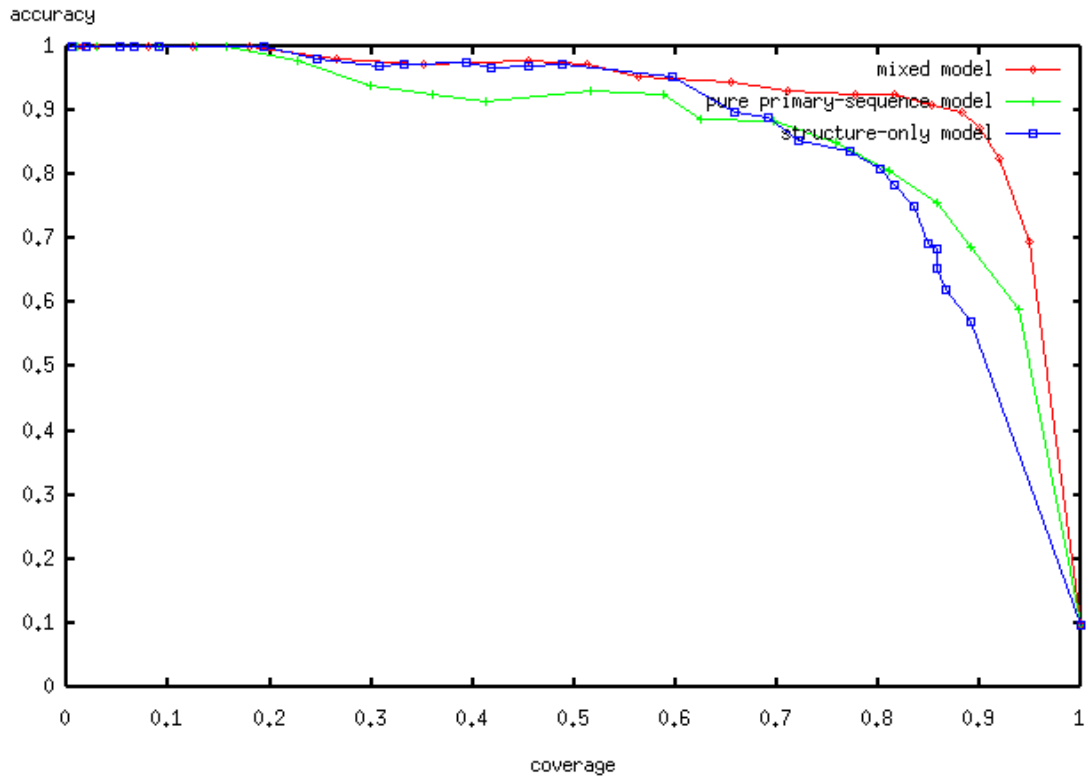


Figure 6-8. Comparison between the C-A plots of the mixed, the structure-only, and the primary-sequence-only models of *rho*-independent transcription terminators

(A) Performance for finding *rho*-independent transcription terminators in *B. subtilis*

Reference	The name of the software	Independent test data	Sensitivity	False positive rate	Intragenic hits
(Ermolaeva et al. 2000)	TransTerm	Yes[1]	86.2% (399/463)	NA[2]	NA[2]
(Lesnik et al. 2001)	RNAMotif	No	NA	NA	NA
(de Hoon et al. 2005)	NA	No	NA[3]	NA[3]	NA
This thesis, 2006	EAR mixed model	Yes	85.3% (180/211)	14% (766/5477)	766

(B) Performance for finding *rho*-independent transcription terminators in *E. coli*

Reference	Sensitivity	False positive rate	Intragenic hits
(Ermolaeva et al. 2000)	89%-98%	NA[2]	NA[2]
(Lesnik et al. 2001)	80%-100%	39% (2586/6635)	2586
(de Hoon et al. 2005)	67%	NA	NA
This thesis, 2006	81% (119/147)	16.6% (431/2604)	431

Table 6-7. Comparison of the performance of different algorithms in finding *rho*-independent transcription terminators in *B. subtilis*

(A) The performances of different algorithms for finding *rho*-independent transcription terminators in *B. subtilis*. (B) The performances of different algorithms for finding *rho*-independent transcription terminators in *E. coli*. Numbers in parentheses are the values that are used to estimate the sensitivities and the false positive rates for different algorithms. The sensitivities are the ratios of experimentally verified terminators that can be successfully predicted by different algorithms. The numbers of predictions that are in intragenic regions are taken as the numbers of false positives. The false positive rates are estimated by dividing the numbers of false positives with the numbers of all predictions. The statistics for TransTerm is estimated by using the results retrieved from http://www.cbcb.umd.edu/software/TransTerm/. The statistics for RNAMotif is retrieved directly from its original paper (Lesnik et al. 2001). The statistics for de Hoon et al.'s algorithm is taken directly from its original paper (de Hoon et al. 2005).

[1]: no negative sequences are used for estimating accuracy and specificity; only sensitivity is estimated by using positive sequences that are not used for training.

[2]: not available because intragenic hits are considered as background and invalidated in final output. For realizing the meaning of this table, see text for details.

[3]: not available because de Hoon *et al.*'s algorithm was trained by using *rho*-independent transcription terminators of *B. subtilis* as the positive training sequences.

NA: not available from respective papers and cannot be estimated by using results retrieved from related websites.

The results reveal that the EAR mixed model is competitive for predicting *rho*-independent transcription terminators in the bacterial genomes. Although the parameters of the EAR mixed model were trained using sequences from *B. subtilis*, this model can find *rho*-independent transcription terminators in *E. coli* with a reasonable sensitivity (81%, this thesis, Table 6-7 B) and a similar estimated false positive rate (16.6%).

In order to compare the EAR mixed model with other algorithms, each case is discussed separately because there are specific considerations associated with each algorithm. Firstly, the sensitivity, 81% (this thesis, Table 6-7 B), is obviously higher than the sensitivity (67%, de Hoon *et al*., Table 6-7 B) for finding *rho*-independent transcription terminators of *E. coli* by using de Hoon *et al*.'s algorithm. The latter was also trained by using sequences from *B. subtilis*. Although de Hoon *et al*.'s algorithm was claimed to have a specificity of 94% for finding *rho*-independent transcription terminators of *B. subtilis*, the high specificity was actually estimated by using only 567 non-terminating sequences (de Hoon et al. 2005), but not random intragenic regions in *B. subtilis*. In addition, the 567 negative sequences, which have been used for training the algorithm, are re-used for testing (de Hoon et al. 2005). The real specificity and false positive rates of de Hoon *et al*.'s algorithm should therefore be regarded as unknown.

Secondly, although the sensitivity (81%, this thesis, Table 6-7 B) of the EAR mixed model for predicting *rho*-independent transcription terminators of *E. coli* seems to be not as good as the sensitivity (80% ~ 100%, Table 6-7 B) of RNAMotif, the false positive rate of the EAR mixed model is estimated as only 14.7%, which is much lower than that (39%) of RNAMotif, calculated in a similar way. It should also be noted that the sensitivity of RNAMotif was estimated with exactly the same positive sequences that had been used for training. No predictions made for other bacterial genomes using RNAMotif can be found in original papers or on related websites.

Thirdly, the sensitivity (85.3%, this thesis, Table 6-7, A) of the EAR mixed model for finding terminators of *B. subtilis* was comparable to that (86.2%, Table 6-7, A) of TransTerm, even though it is impossible to estimate the false positive rates of TransTerm due to its peculiar way of estimating the confidence of predictions (Ermolaeva et al. 2000) (For details see discussions in the 5th paragraph in the introduction of this subsection, 6.1.2.).

Consequently, among the algorithms mentioned above, the EAR mixed model is the only *rho*-independent transcription terminator finding approach for which reasonably robust indicators of both sensitivity and specificity are available.

6.1.2.2.2. The Eponine windowed RNA-motif model (EWR model)

rho-independent transcription terminators should still be considered an easy case when evaluating ncRNA-finding algorithms, since there is a clearly definable reference point, namely the transcription termination site, in each sequence. When no obvious reference points are known, finding consensus RNA motifs is difficult for most available computational approaches. The Eponine windowed RNA motif model (EWR model) is specifically designed for such situations.

The results presented here (Figure 6-9) reveal that the EWR models are capable of finding key signals, corresponding to A-region (the motifs at offset 0 in sensors 1 and 2, Table 6-8), the stable hairpin (the motif at offset 26 in sensor 1, and the motif at offset 16 in sensor 2, Table 6-8), and T-region (the motif at offset 58 in sensor 1, and the motifs at offsets 42 and 79 in sensor 2, Table 6-8), for *rho*-independent transcription terminators in unanchored sequences (see subsection 6.1.2.1.4.). Although the performance of this EWR model (Figure 6-11) is not really comparable to the EAR mixed model, nearly 70% accuracy could be achieved when the coverage is 70%.

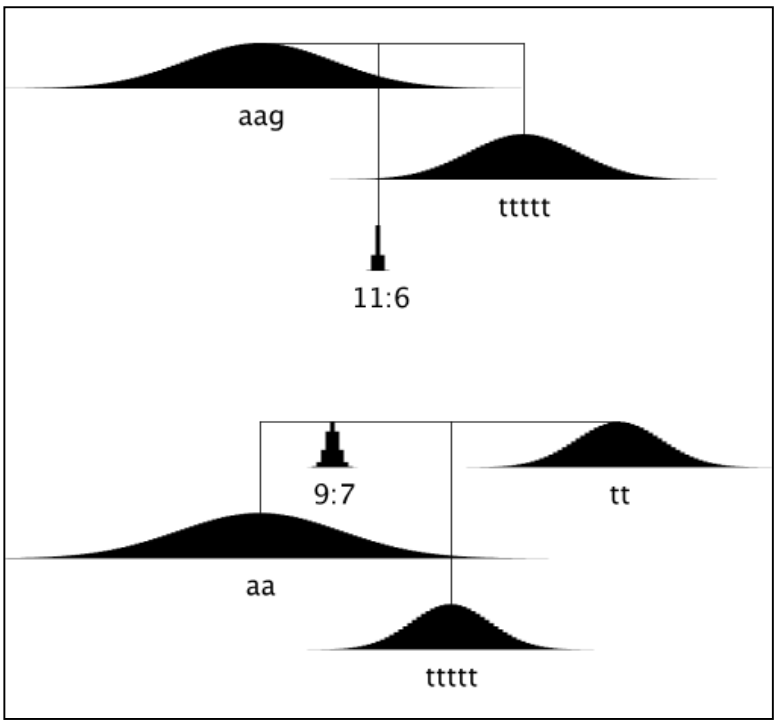


Figure 6-9. An EWR model for *rho*-independent transcriptional terminators

There are two convolved sensor basis functions (CSBFs, see subsection 4.1.2.1.2.) in the GLM of the EWR model for *rho*-independent transcription terminators. The upper one is referred to as sensor 1 and the lower one is referred to as sensor 2 in the following text.

Sensor 1:

Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0	16.25	Not available (a PWM of 3 columns)			
26	0.58	6	7.02	11	0.08
58	11.99	Not available (PWM, 5 columns)			

Sensor 2:

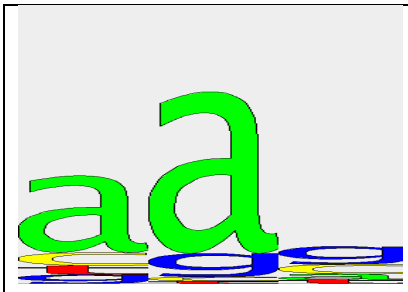
Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0	17.92	Not available (a PWM of 2 columns)			
16	1.39	7	8.69	9	0.13
42	8.62	Not available (a PWM of 5 columns)			
79	9.36	Not available (a PWM of 2 columns)			

Table 6-8. The trained parameters of an EWR model for *bacillus rho*-independent transcription terminators

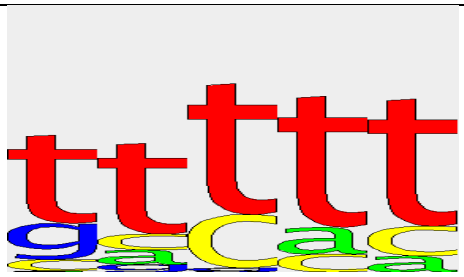
Sensor 1 is the convolved sensor basis function (CSBF) presented in the upper half of Figure 6-9 and sensor 2 is the CSBF presented in the lower half of Figure 6-9

"Offset" refers to the mean of the discrete Gaussian distribution used to model the distance between each motif and the first motif. Other titles follow the convention of Table 6-4.

Sensor 1:

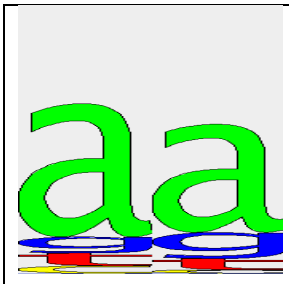


Offset: 0

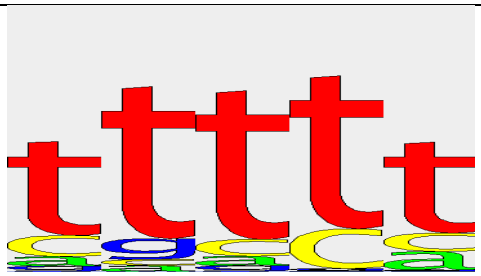


Offset: 58

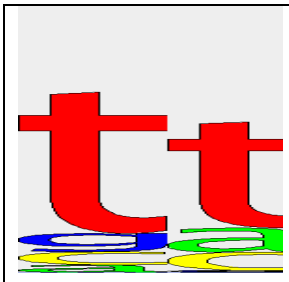
Sensor 2:



Offset: 0



Offset: 42



Offset: 79

Figure 6-10. The sequence logos of position-constrained motif matrices presented in Table 6-8 and Figure 6-9

"Offset" corresponds to "Offset" column in Table 6-8. Sensors 1 and 2 correspond to the sensors in Table 6-8 and Figure 6-9

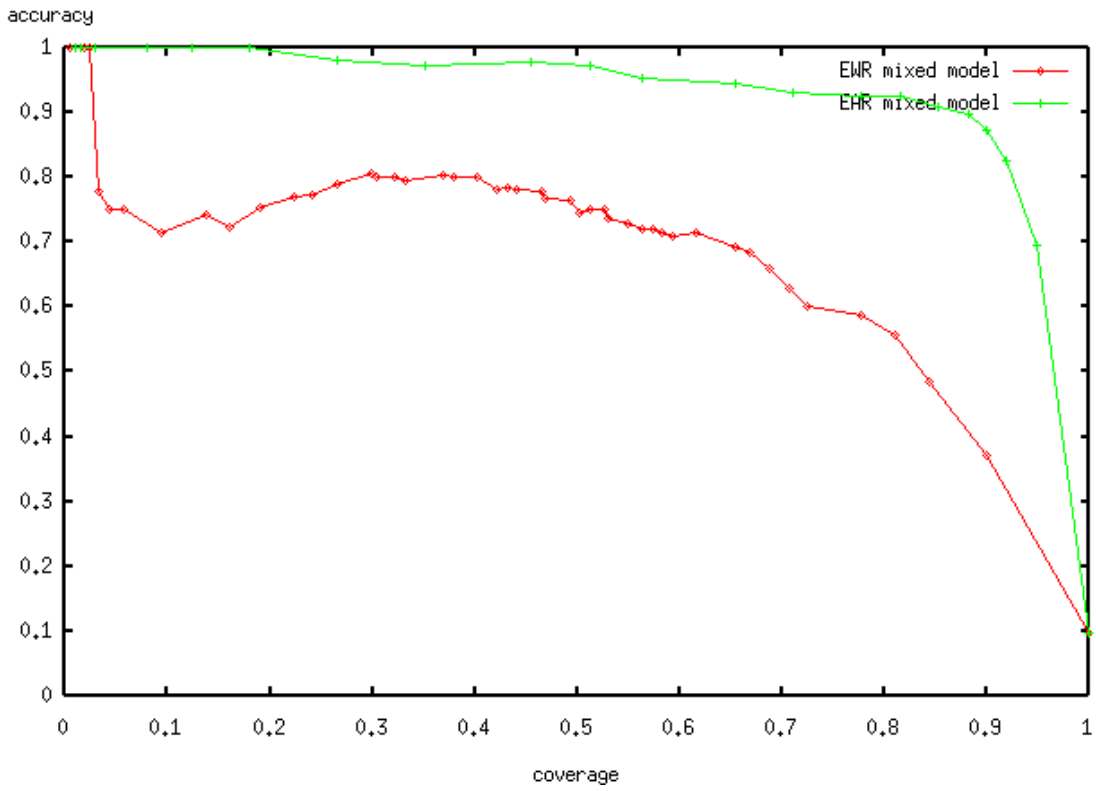


Figure 6-11. Comparison of the C-A plots of an EAR mixed model and an EWR model for *rho*-independent transcription terminators

6.1.2.3. Discussion

One obvious question about using the Eponine RNA extension to model *rho*-independent transcription terminators is the wide distribution of motif positions. For example, in the EAR mixed model (see subsection 6.1.2.2.1.), the width of the position distribution of the T-region is 17.96 (weight 13.62, Table 6-6). In the EWR model (see subsection 6.1.2.2.2.), there are also heavy tails for position distributions of both the A-region and the T-region (Figure 6-9). It seems that both of the EAR and the EWR models for *rho*-independent transcription terminators are inconsistent with the current view that the stable hairpin is immediately followed by the T-region. However, it should be noted that in the Eponine RNA-motif

extension, the first base of the respective hairpin is used as the position of each RNA structural motif. Consequently, in the EAR mixed model, the distances between the reference point (presumably the first base of the transcription termination signal) and the T-region in different sequences varies in response to the variations in the dimensions (loop size and stem size) of the stable hairpin in *rho*-independent transcription terminators. For similar reasons, it is not surprising that the wide position distributions of the T-region were also found in the EWR model of *rho*-independent transcription terminators. Consequently, the current implementation of the Eponine RNA-motif extension may not model ideally the proximity of motifs to their 5' adjacent structural motifs.

The inadequacy in modelling the exact relations between motifs and reference points separated by variable length structural motifs is a current weakness of the Eponine RNA-motif extension. For the purpose of modelling the relation between the hairpin and the T-region in the *rho*-independent transcription terminators, using the last base of the stem region as the location (reference point) for each structural motif might be helpful. However, switching the reference point for structural motifs is not expected to be a solution in all the situations, especially when the ncRNAs of unknown types are modelled as the most suitable reference points for a hairpin may vary from case to case. For example, in modelling the RNA motifs where the loop regions are responsible for the specific interaction with proteins, the most suitable anchoring point for hairpins could be the centre of the loop regions.

6.1.3. Modelling pseudoknots

Pseudoknots are seldom used for testing algorithms for finding consensus RNA motifs. Algorithms that were claimed to be capable of finding consensus pseudoknots in a set of sequences include GPRM (Hu 2002), ILM (Ruan et al. 2004), and comRNA (Ji et al. 2004). There are certain restrictions in using these algorithms. For example, GPRM and comRNA

cannot find primary-sequence motifs; users of GPRM must assign the expected number of hairpins in sequences; ILM requires pre-aligned sequences.

Although the Eponine RNA-motif extension is not specifically designed for finding consensus pseudoknots in sequences, it is not prohibited from finding consensus hairpins that overlap with each other, such as non-juxtaposed and non-nested stem regions in pseudoknots. In other words, the Eponine RNA-motif extension has the potential to find consensus pseudoknots in a set of sequences. The additional advantage of using a classification machine, such as the Eponine RNA-motif extension, is that the trained model may be applicable to finding new functionally related pseudoknots in genomes.

6.1.3.1. Materials and methods

To assess the capability of the Eponine RNA-motif extension for finding consensus pseudoknots, 18 sequences of 3' UTRs of genes of soil-borne rye mosaic viruses and soil-borne wheat mosaic viruses, which were also used by Hu (Hu 2002) for assessing GPRM, were recruited from the PseudoBase database (van Batenburg et al. 2001) as positive training sequences. Five hundred sequences of 40 bases in length were randomly sampled from the human genome and used as negative training sequences. The human genome assembly used for random sampling was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/).

These training sequences were used to train an EAR model as well as an EWR model. When the EAR model was used to model these pseudoknots, the first base of each sequence was used as the anchoring point.

6.1.3.2. Results

The resulting EWR model for the 3' UTRs of viral genes consisted of two consensus hairpins (Figure 6-12). The stem regions of these two hairpins were neither juxtaposed nor

nested. The distribution of the first base of the second hairpin peaks (offset: 5, hairpin ID 2, Table 6-9) at the end of the 5' stem of the first hairpin (stem size: 7, hairpin ID 1, Table 6-9). The most probable positions of the two hairpins were consistent with the configuration of the pseudoknots in these 3' UTRs of viral genes that were used for training. The result shows that the EWR models are capable of finding consensus pseudoknots in a set of sequences.

An EAR model for the pseudoknots in 3' UTR of viral genes was also trained. This EAR model also consisted of two hairpins (data not shown), which is consistent with the non-nested configuration of pseudoknots as shown in the EWR model.

Hairpin ID	Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
1	0	2.7	4	8.8	7	0.8
2	5	2.7	9	4.1	4	0.2

Table 6-9. The trained parameters of an EWR model for pseudoknots in 3' UTRs of viral genes

The titles used in this table follow the convention of Table 6-8.

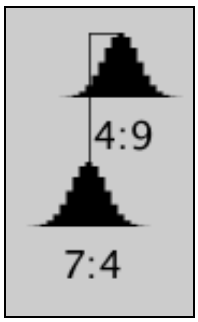


Figure 6-12. An EWR model for the 3' UTRs of viral genes

The notation used to describe RNA hairpins follows the convention of Figure 6-1.

6.2. Discussions

6.2.1. Considerations of using the Eponine RNA-motif extension

In order to train an Eponine RNA-motif model, a number of positive training sequences are required. For example, a set of ten sequences is insufficient for finding the pseudoknots in the 3' UTRs of viral genes with the current implementation and parameter settings of the Eponine RNA-motif extension. Training an Eponine RNA-motif model may require tens of positive sequences. In terms of finding functional RNA motifs, this requirement seems to be a weakness of the Eponine RNA-motif extension, compared to algorithms that can predict optimal RNA structures using only few sequences. Nonetheless, by using only a few sequences or even one sequence, available RNA-motif finding algorithms may also have difficulty in finding consensus structures in a set of unaligned sequences (Gardner and Giegerich 2004). Even though the algorithms that take pre-aligned sequences seem to have a good performance, none of them have been tested on alignments of real genomic sequences. Existing tests have generally been performed on alignments of well-trimmed sequences (Hofacker et al. 2002; Knudsen and Hein 2003; Coventry et al. 2004; Gardner and Giegerich 2004; Ruan et al. 2004). A similar situation is also true for the ncRNA classifying algorithms that utilise pre-aligned sequences (see also subsection 2.1.3.5. , chapter 2).

Another issue around using the Eponine RNA-motif extension is the computer time required for training a model. For example, it may take ~7 hours (24,108 seconds) and ~22 hours (79,661 seconds) to train an EAR mixed model and an EWR mixed model respectively for human tRNAs (Table 6-10). Within the trainer, predicting all local hairpins in each training sequence is not the most time-consuming step when using the Eponine RNA-motif extension. With the current implementation of the fast model of the Eponine RNA-motif extension, it takes less than 3 seconds by using an x86-64bit machine (3.2 Ghz Pentium IV EMT64, 64-bit

Linux) to predict local hairpins for a sequence of 250 bases in length. A significant proportion of time is actually spent using the Monte Carlo method to optimise parameters of PWMs and RNA motifs. For example, it is estimated that three-fourths of the CPU time used for training an EAR mixed model of tRNAs is spent in learning parameters of motifs, while only one-fourth of the CPU time (~6000/24108) is spent in predicting local RNA secondary structures (Table 6-10, tRNAs, EAR mixed model, CPU time).

	Training type	Sequence length	Number of positive sequences	Number of negative sequences	CPU time (x86-64bit) (seconds)
tRNAs	EAR mixed model	250	200	2000	24108.83
	EWR mixed model	250	200	2000	79661.45
Rho-independent transcription terminators	EAR mixed model	170	212	2000	15162.24
	EWR mixed model	170	212	2000	47300.76

Table 6-10. The execution time for training the EAR and the EWR models of tRNAs and rho-independent transcription terminators

"CPU time" is the CPU time of a 3.2 Ghz Pentium IV EMT64 machine which runs the 64-bit Linux OS.

When a trained model is applied to finding a particular type of RNA motifs in genomic sequences, most of the time will be spent on folding all windowed regions of genomic sequences. Using the Eponine RNA-motif models to scan the whole genome for searching RNA motifs can be very time-consuming. For example, using the EAR model to search for transcription termination terminators in the bacterial genomes took as long as one-week CPU time on an x86-64bit machine (3.2 Ghz Pentium IV EMT64, 64-bit Linux), scanning ~4-megabases x 2 (Table 6-11).

Organism	Genome length	CPU time (Pentium-4) (secs)
B. subtilis	4,214,630 x 2 strands	589755.91
E. coli	4,639,675 x 2 strands	638613.26

Table 6-11. The execution time for using the EAR model of *rho*-independent transcription terminators to scan the genomes of *B. subtilis* and *E. coli* respectively

6.2.2. Towards creating general EWR models of vertebrate ncRNAs

The scoring scheme of the Eponine RNA-motif extension is designed to allow a dynamic recruitment of relevant features. By using the Monte Carlo methods and the RVM strategy, theoretically the Eponine RNA-motif extension can determine the differential degrees of significance of various structural features for a particular hairpin and then choose the most relevant features for modelling it. In this project, however, this capability has not yet been evaluated. Lengths of stems and loops are currently the only features that have been recruited to model ncRNAs. It is possible that under certain circumstances, other features could significantly contribute to the model. While the hairpins of different classes of ncRNAs may vary in their stem and loop sizes, a recent report suggests that ncRNAs tend to have more stable structures than do random sequences (Clote et al. 2005). Although folding stability alone proved to be insufficient for identifying ncRNAs in genomes (Rivas and Eddy 2000), certain combinations of different structural features might be useful for genome-wide ncRNA finding.

One unfinished piece of work in this project is using the Eponine RNA-motif extension to create a general EWR model of vertebrate ncRNAs. There can be at least two approaches to fulfil this goal. Firstly, the EWR model can be used to find the consensus features of various classes of ncRNAs. In order to evaluate the performance of the trained model, a *k*-fold cross validation can be used. ncRNA classes can be divided into *k* groups and each group of ncRNAs is left out when training that particular model. The trained model could then be evaluated by using these ncRNAs. This process would be repeated until the *k* models had been evaluated.

Another possible approach for creating an EWR vertebrate-ncRNA-model is taking human-mouse syntenic alignments as the training sequences. The proposed approach can be,

not only a potential way to create a general ncRNA model, but also a useful strategy to look for undiscovered ncRNAs in mammalian genomes. The development of the Eponine RNA-motif extension provides a way to test hypotheses with regard to genome-wide ncRNA finding. The capability of this tool in genome-wide ncRNA finding is worthy of further exploration.

6.3. Summary

In this chapter, using three types of ncRNAs with distinct RNA structural motifs, I have demonstrated the capability of the Eponine RNA-motif extension to model the RNA motifs in transcripts. The applications of this extension include the following:

- When a particular type of functional sites is known for a set of sequences, Eponine anchored RNA-motif models can be used.
- When a functional site is suspected but the anchoring point in a set of transcripts is unknown, Eponine windowed RNA-motif models can be used.
- Eponine RNA models can be used for prediction, *i.e.* to search for novel sites of a particular type of ncRNAs in genomes.

There are some limitations of the tentative applications of the Eponine RNA-motif extension:

- The Eponine RNA-motif extension is designed to learn discrimination models consisting of local RNA motifs. This tool may not be capable of modelling the global consensus RNA secondary structure.
- For the purpose of discriminating novel functional sites in genomes, the trained model may be apt to find false positives that consist of only a subset of functional motifs.

There are some special issues that need to be taken into consideration in using the Eponine RNA-motif extension:

- A number of sequences are required for training the models.
- In training the models, significant amount of time may be spent in learning the parameters of PWMs and RNA motifs, due to the use of the Monte Carlo methods in optimization.