Chapter 5. Modelling the transcription regulatory elements of mammalian RNA polymerase III genes

Most existing ncRNA finding algorithms are designed to find structural ncRNAs. These algorithms can be regarded as being structure-dependent, because they use the potential of a particular genomic region to fold into high-order RNA structures as a signal of the existence of ncRNAs. However, structure-dependent ncRNA finding algorithms will fail to predict non-structured ncRNAs, whose functions do not depend on folding into high-order structures. In addition, a non-transcribable genomic region may be misclassified as an ncRNA locus simply because a region of structure-formation potential is predicted by structure-dependent algorithms. Therefore, to address the problem of genome-wide ncRNA finding, it is useful to consider complementary structure-independent approaches, in addition to structure-dependent algorithms. In this chapter, the possibility of using a type of structure-independent genome-wide ncRNA finding approach is explored, based on the modelling of the transcription regulatory elements.

Transcription regulatory elements have been used as a signal for finding particular classes of ncRNAs, such as tRNAs (Fichant and Burks 1991; Pavesi et al. 1994; Lowe and Eddy 1997). However, the identification of transcription regulatory elements is currently used as a screening step, not as a determination step, in genome-wide ncRNA finding. If transcription regulatory element methods are used alone for genome-wide ncRNA finding, the false-positive rate can be very high. For instance, eufindtRNA, which is an internal-promoter finding program, predicts over 1,300 candidate loci for tRNAs on human chromosome 1 (in the NCBI 35 assembly), but only less than ~10% (120) of them may be functional tRNAs based on evaluation using structure-folding potentials.

It is essentially unknown why the methods designed to predict the transcription

regulatory elements of ncRNAs appear to suffer from high false positive rates. Some possible explanations are as follows. Firstly, it is possible that existing promoter models were not built specifically for finding mammalian tRNA genes. The specificity of these tools may have been sacrificed, to a certain extent, in order to make them sensitive enough for finding tRNA genes in multiple organisms. Secondly, internal promoters may be just part of the signal required for determining the transcription specificity of tRNA genes in mammalian genomes. It is possible that other non-promoter transcription regulatory elements, such as enhancers/silencers and LCRs, may play a role in the specific initiation of tRNA transcription. Thirdly, some of the non-tRNA loci which appear to contain the internal-promoter-like patterns might correspond to novel non-tRNA ncRNA genes.

Consequently, the specific aims of this chapter include:

- Learning a new model for selectively predicting tRNAs, as well as novel ncRNA genes transcribed by RNA polymerase III (pol III genes), in the mammalian genomes.
- Finding evidence to support the functionality of the predicted non-tRNA pol III genes.

The Eponine system, described in chapter 4, appears to be suitable for these purposes. Eponine models have previously been used to predict functional sites, such as transcription start sites (TSSs) and transcription termination sites (TTSs), in complex genomes. Given a set of training sequences, the Eponine trainer can simultaneously learn the important signals, in the form of PWMs, and the "architectural" relationship (*i.e.* the distance distribution) of PWMs to a particular type of functional sites (for a detailed discussion see section 4.1, chapter 4). Eponine is one of the few systems that have been applied to learning a model capable of selectively predicting the TSSs of protein coding genes in mammalian genomes (Down and

Hubbard 2002). Given Eponine's success in modelling RNA polymerase II (pol II) TSSs, one interesting question is whether Eponine models are useful for predicting the ncRNAs transcribed by pol III in mammalian genomes. Therefore, in this chapter, the Eponine system was taken as a quick approach for modelling the transcription regulatory regions of mammalian pol III genes.

In this chapter, the Eponine Anchored Sequence (EAS) model (see section 4.1, chapter 4) was tried for creating a new model for discriminating pol III genes in the mammalian genomes.

5.1. Modelling the transcription start sites of mammalian pol III type II genes

In this section, the Eponine Anchor Sequence (EAS) model was used to model the transcription start sites (TSSs) of pol III genes. A suitable training set should consist of the genes that contain promoters with similar architectures, because the EAS model is not designed for managing a heterogeneous set of functional sites that are each associated with distinct combinations of transcription factor binding sites (TFBSs). For that reason, a brief introduction to the types of promoter architectures of eukaryotic pol III genes is given in the following.

There are three distinct types of promoter architecture that have been found in eukaryotic pol III genes, where each type of promoter is associated with a unique combination of distinct TFBSs (see Table 5-1) (for review see Paule and White 2000). The promoters of type I and type II genes are intragenic. Type I (*e.g.* 5S rRNAs) and type II (*e.g.* tRNAs) genes share an "A box" (sometimes also known as the "A block"), which is the binding site of TFIIIC. A "C box" (sometimes also as the "C block"), which is the binding site of TFIIIA, is unique to type I genes. A "B box" (sometimes also as the "B block"), which is the binding site of TFIIIB, is

unique to type II genes. Although "A boxes" for tRNAs and 5S rRNAs can be exchanged, the distances to their respective TSSs vary: it seems that the distance for tRNA genes is 10 bases, while the distance for 5S rRNAs is 50 bases. Although there are no TATA boxes for mammalian type I and II genes, the transcription factors (TFs) that interact with intragenic TFBSs seem to guide TATA-Box Binding Protein (TBP) to the upstream regions of type I and II genes and TBP can recruit pol III to the correct transcription start sites. On the other hand, promoters of type III genes are 5' to the TSS in the upstream region. Unique TFBSs of type III genes are the TATA box, the proximal sequence element (PSE), and the distal sequence element (DSE).

Type	Genes	Core TFs	TFBFs
Type I	5S rRNAs, *etc.*	TFIIIA, TFIIIC, TFIIIB, TBP, polIII	A box and C box (Intragenic regions)
Type II	tRNAs, VARNAs, 7SL, *etc.*	TFIIIC, TFIIIB, TBP, pol III	A box and B box (Intragenic regions)
Type III	U6 snRNAs, 7SK, *etc.*	TFIIIC1, TFIIIB, TBP, SNAPc, pol III	PSE, TATA box, DSE (Upstream regions)

Table 5-1. The TFs and the TFBSs associated with three distinct types of eukaryotic pol III genes

Given these distinct architectures, when creating a model that may discriminate tRNA genes as well as other pol III genes, the sources of training sequences needs to be limited to those of pol III type II genes. In the set of pol III type II genes, VARNA1 genes can be another source of training sequences, in addition to tRNAs. To date, more than 40 VARNA1 genes have been found. Although there are other pol III type II genes such as 7SL, these genes are not as numerous as VARNA1 genes. VARNA1s are encoded in adenoviruses (Weinmann et al. 1974) and they are transcribed by the mammalian RNA pol III machinery. Hence, VARNA1 genes can be considered as mammalian pol III type II genes, because there is evidence that the promoters of VARNA1 genes are similar to these of mammalian tRNA genes (Cannon et al. 1986; Wu et al. 1987).

Thus, in this section (5.1), VARNA1s and tRNAs were used as training sequences to generate an Eponine EAS model for pol III type II TSSs.

5.1.1. Materials and methods

5.1.1.1. Training and test data sets

For the purpose of creating an EAS model, one set of positive sequences and one set of negative sequences are required.

The human tRNA genes and adenovirus VARNA1 genes were used as the positive sequences. The set of mouse tRNA genes predicted by tRNAscanSE were not included because the set might contain a large number of pseudogenes (Mouse Genome Sequencing Consortium 2002). A set of negative sequences were recruited by taking random samples from the human genome. The preparation of these sequences for training and testing is described in the following subsections (subsections 5.1.1.1.1. , 5.1.1.1.2. , and 5.1.1.1.3.).

5.1.1.1.1. Preparation of human tRNA sequences

In order to avoid over fitting of a learned model to training data, validation is necessary. One type of validation is to evaluate the performance of trained models on test data that is independent of the training set. If the performance of a trained model is significantly worse than on the training data, this may indicate that this model has been over fitted to the training data.

Therefore, the recruited tRNA genes were partitioned into two groups, one for training and the other for testing. Due to the high redundancy in the set of human tRNA genes, proper partitioning became an issue. For instance, there can be as many as 20 nearly identical copies for a particular anti-codon type of tRNA genes. When using a random sampling process, it is unlikely that all the highly similar tRNA genes would be grouped into a single set. Here, I took advantage of the forty tRNA-gene subgroups already prepared in section 2.2, chapter 2,

where these subgroups were generated according to the anti-codon types and pairwise sequence identities of tRNA genes (for details see materials and methods of section 2.2, chapter 2). These forty subgroups were re-merged into two groups, group 1 and 2, based on the pairwise identities between the consensus sequences of the subgroups. The grouping process was carried out in a progressive manner, where the two groups with the highest consensus identity were merged first, and then the groups with the next highest identity were successively merged.

Group 1 and group 2 consisted of 200 and 167 human tRNA genes, respectively. The inter-dependence between the training set and the test set was further assessed by comparing the inter-group and intra-group sequence identities. Each sequence was used to search for its most similar sequences in the same group and in the other group, respectively. The results reveal that there is a clear sequence-identity difference between these two groups, since all the intra-group best pairwise identities were greater than 83% and all the inter-group best pairwise identities were smaller than 78% (Figure 5-1). The results suggest that the tRNA genes in group 1 are distinct from the tRNA genes in group 2. The tRNA genes in group 1 (group-1 tRNA genes) were used for training and the tRNA genes in group 2 (group-2 tRNA genes) were used for testing.

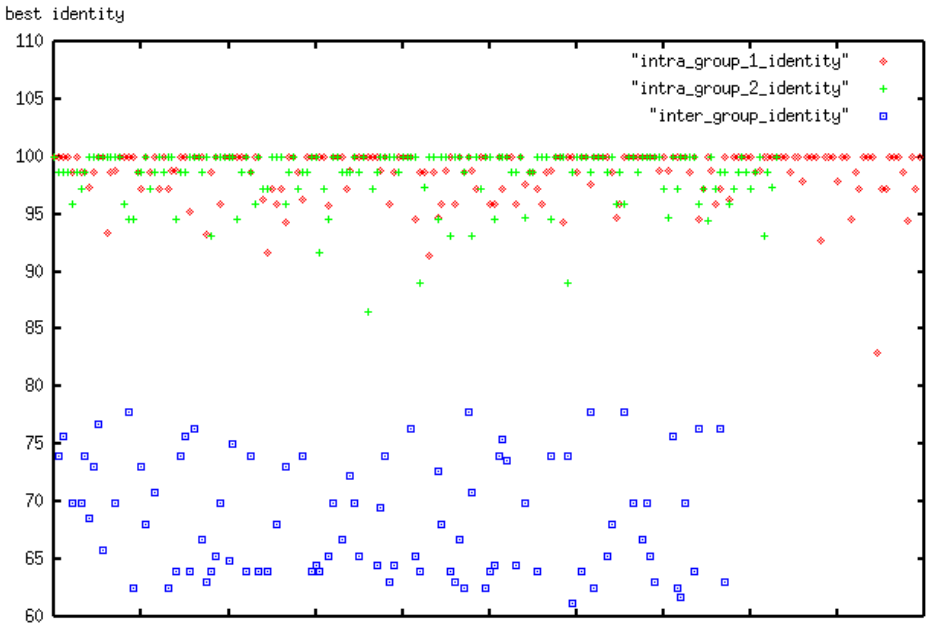


Figure 5-1. Separation of the sequence identity distributions between intra-group and inter-group sequences of tRNA genes.

When preparing the tRNA sequences for training and test, the first base of the cloverleaf-like structure of each recruited tRNA gene 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. The purpose of including the upstream and downstream flanking regions of the recruited tRNA genes in training sequences is to explore if there are motifs other than the A box and B box that can be used to model the TSSs of pol III type II genes.

5.1.1.1.2. Preparation of VARNA1 sequences

VARNA1 genes were used as another source of sequences for building a pol III type II TSS model. Forty-three regions containing VARNA1 genes were retrieved from GenBank by using the keyword "VARNA1". VARNA1 sequences were extracted from these regions by using the locations indicated in the GenBank annotation. By using TGICL (TIGR 2002-2003), VARNA1 genes were clustered into 5 subgroups (for the detailed procedure for the sequence clustering, see section 2.2, chapter 2). The 5 subgroups were further merged into two independent groups. Group 1 and group 2 consisted of 9 and 32 VARNA1 genes, respectively.

An assessment on the sequence independence, as mentioned in preparing the human tRNA genes for training, was also performed here. The results show that all the intra-group best pairwise identities were greater than 95%; all the inter-group best pairwise identities were smaller than 86%. The results suggest that the VARNA1 genes in group 1 are distinct from the VARNA1 genes in group 2.

Group-1 VARNA1 genes together with group-1 tRNA genes were used for training (Table 5-2, Training). Group-2 VARNA1 genes and group-2 tRNA genes were used for testing (Table 5-2, Testing). The 32 genes used for testing actually correspond to 9 distinct ones, because many of them have exactly the same sequences. Likewise, the 9 genes used for training correspond to 7 distinct ones.

	Training	Testing
Human tRNA genes	200 (group 1)	167 (group 2)
Adenovirus VARNA1 genes	9 (group 1)	32 (group 2)
Subtotal	209	199

Table 5-2. The training and test data sets for creating an EAS model for pol III type II TSSs

When preparing the VARNA1 sequences for training and test, the first base of each gene was used as the anchoring point; 100 bases upstream and 150 bases downstream with respect to the anchoring point in each VARNA1 gene were retrieved. The purpose of including flanking sequences for training is the same as described to prepare tRNA sequences for training in the previous subsection (see subsection 5.1.1.1.1.).

5.1.1.1.3. Preparation of negative sequences

Two sets of ten thousand random sequences were sampled from the human genome as negative training and test sequences, respectively. These random sequences were 250 bases in length.

5.1.1.2. Evaluation of the performance of EAS models against the test data set

When evaluating the accuracy of trained EAS models against the test data set prepared as described in 5.1.1.1. , 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 test 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 test sequence was predicted as a hit.

5.1.1.3. Presentation of the performances of different models

The performances of all trained models will be presented in the form of coverage-accuracy (C-A) plots. Coverage (sensitivity) is the proportion of true positive sequences that can be correctly predicted; accuracy (positive predictive value) is the proportion of true positive sequences in the set of predicted sequences. For example, with a specific threshold, if 150 out of 199 positive test sequences are successfully predicted and 5 out of 10000 negative test sequences are incorrectly classified as the pol III type II genes, the accuracy is 96.8% (150/(150+5)) and the coverage is 75.4% (150/199).

The C-A plot can be considered as an alternative presentation of Receiver Operating Characteristic (ROC) curves, except that the size of negative test sequences is not considered in the former plot. Plotting these characteristics is especially useful when comparing the performances of two competing models when using an extremely large negative data set, such as random sequences from the human genome. Suppose that there are two models, where model X predicts 150 false positives from 10,000 negative test sequences, while model Y predicts 100 false positives. Both models can predict 150 true positives from 200 positive test sequences. The false positive rates are 1.5% and 1% respectively. In contrast, the accuracies for these models are 50% (150/(150+150)) and 60% (150/(150+100)), respectively, and thus the difference between their performances can be easily seen in a C-A plot. Consequently, for

evaluating the performances of methods that are designed for finding functional sites in large and complex genomes, C-A plots are more suitable than the classic ROC curves.

5.1.1.4. Evaluation of the performance of EAS models against real genomic sequences

The performance of EAS pol III type II TSS models was also evaluated against human chromosomes 11 and 13. The human genome assembly used in this evaluation was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/).

When using EAS pol III type II TSS models to scan a chromosome, each position can be the start of a putative pol III type II gene. Consecutive hits would be clustered together if all of their scores were higher than a particular threshold. Such hits were regarded as a single record of prediction.

5.1.1.5. Determining overlapped genomic hits predicted by using different methods

An EAS pol III type II TTS model predicts the transcription start sites in genomes. By contrast, existing tRNA gene finding algorithms, such as eufindtRNA and tRNAscanSE, predict a range, namely the start and end positions for each putative tRNA gene. To determine the overlapped hits predicted using different methods, the following approach was used. If a tRNAscanSE (or eufindtRNA) predicted hit was within 100 bases downstream of an EAS pol III type II TTS model predicted site, the two hits predicted by different methods were considered to represent the same gene.

5.1.2. Results

5.1.2.1. Naïve training by using default parameters

Using the training sequences prepared as described in 5.1.1. , an Eponine Anchored Sequence (EAS) model for the mammalian pol III type II promoters was trained. Figure 5-2 is a schematic presentation of the constraint distributions relative to the anchoring point as

indicated by the blue triangle. The anchoring point in this figure corresponds to the transcription start site of pol III type II genes. The relative width of the position distributions for each hairpin is shown by the width drawn. The sequence under each constraint is motif consensus sequence. The sequence logos of the motifs in this model were presented in Figure 5-3. In the remaining part of this thesis, other Eponine models will be presented using this convention.

There were several problems with this model. Firstly, the model was unable to distinguish *bona fide* tRNA genes from random sequences (data not shown). Secondly, both the patterns of A box and B box were much shorter than what have been suggested by experimental approaches (DeFranco et al. 1980; Galli et al. 1981). Further investigation revealed that between VARNA1s and the human tRNAs, the 8th to 22nd positions, which are supposedly the "A box", are very different.

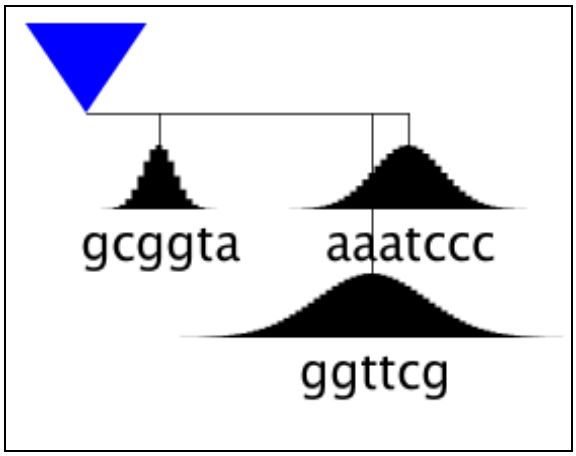


Figure 5-2. An EAS model for pol III type II promoters (naïve training)

Weight: 6.52, position: 6, width: 7.36 Weight: 11.28, position: 53, width: 3.18

Weight: 6.91, position: 68, width: 5.99

Figure 5-3. The sequence logos of position-constrained motif matrices of the naïve EAS model (Figure 5-2) for pol III type II promoters

The value of "weight" for each motif corresponds to the weight associated with each basis function in the GLM of an EAS model. The value of "position" for each motif corresponds to the mean of the discrete Gaussian distribution used to model the position of a motif relative to the reference site. The value of "width" corresponds to the width of the discrete Gaussian distribution (for other details about these parameters see subsection 4.1.2.1.1)

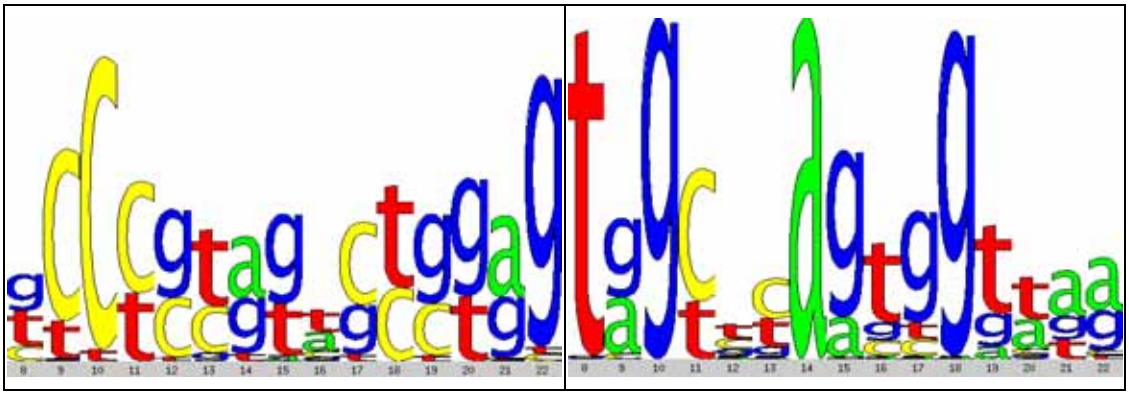


Figure 5-4. Comparison between the sequence logos of the 8th-22nd positions of VARNA1s (left) and tRNAs (right)

5.1.2.2. Optimizing the anchoring points

From the results presented above, VARNA1s, which are viral genes rather than real mammalian genes, seem to be unsuitable for training pol III type II promoter models. However, on investigation it was found that the poor training was probably due to the incorrect assignment of the anchoring points for the recruited sequences. The first base of the cloverleaf-like structure of tRNAs, is in fact not the transcription start site. The real transcription start sites of mammalian tRNAs are at the 5' regions upstream of the first base of cloverleaf-like structures. After transcription, the 5' dangling sequences of the raw tRNA transcripts must be cut off by RNase P (for review see Gopalan et al. 2002). On the other hand, transcription start sites of VARNA1s are generally used as the first bases for VARNA1 genes in the GenBank annotation.

After adjusting the anchoring points of the recruited sequences, manual alignments reveal that respective "A boxes" of VARNA1s and the human tRNAs are quite similar (Figure 5-5). These results show that when inconsistent anchoring points are provided, the Eponine trainer for the EAS models can be incapable of optimizing the PWMs.

Figure 5-5. Comparison between sequence logos of the presumable internal promoter regions of VARNA1s (left) and tRNAs (right) (after adjusting anchoring points of VARNA1s)

5.1.2.3. The EAS pol III type II promoter model

Using the sequences with correct anchoring points, a new EAS pol III type II promoter model was trained. This model is called "model 1" in the remainder of section 5.1. This model appears to be quite complex (Figure 5-6). There are five distinct motifs at the 6th, 19th, 43rd, 52nd, and 53rd positions. Respective weights for these motifs in the generalized linear models are 4.76, 8.34, 4.37, 9.01, and 12.58.

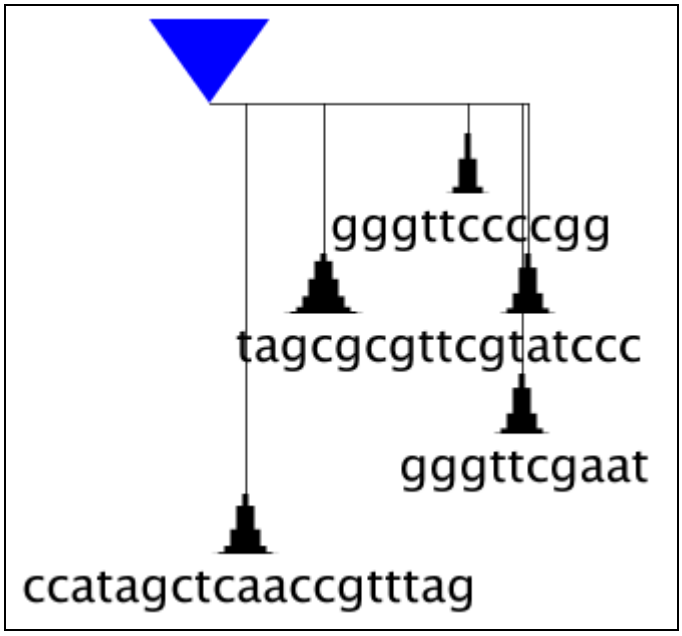


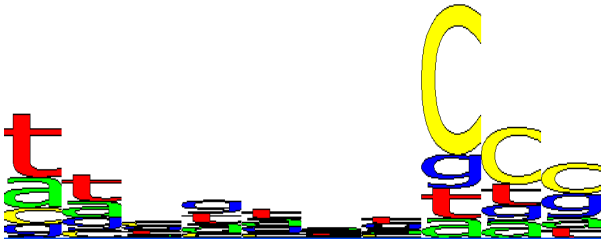
Figure 5-6. An EAS pol III type II promoter model (after adjusting the anchoring points of VARNA1s) (model 1)

Weight: 12.57, position: 6, width: 1.45

Weight: 4.37, position: 19, width: 1.87

Weight: 4.76, position: 43, width: 0.87

Weight: 9. position: 52, width: 1.30



Weight: 8.34, position: 53, width: 1.30

Figure 5-7. The sequence logos of position-constrained motif matrices of model 1 (Figure 5-6)

The annotation used in this figure follows the convention of Figure 5-3

The motifs in the new model fit the current knowledge about transcription regulation of mammalian pol III type II genes. The motifs that start at 6th and 19th positions correspond to the 5' and 3' parts of the "A box" respectively. The motifs that start at 43rd, 52nd, and 53rd positions, which are similar to one another, correspond to the "B box". The three positions represent discrete preferred sites of the "B box" in mammalian tRNA genes. The variation in the location of the "B box" is consistent with the previous reports which indicated the flexibility in distance between the "A box" and the "B box" in eukaryotic tRNA genes (Camier et al. 1990; Pavesi et al. 1994).

5.1.2.3.1. The performance of model 1 – using the recruited test sequences

The performance of model 1 was initially assessed against 199 positive test sequences recruited as described in 5.1.1.1.1. and 5.1.1.1.2. , and a set of 10,000 negative test sequences prepared as described in 5.1.1.1.3. The results reveal that model 1 can achieve 100% accuracy at 70% coverage on this data set (Figure 5-8, model 1). The high accuracy suggests that model 1 may have a low false positive rate. Besides, at this accuracy and coverage, ~50% distinct VARNA1 sequences in the test data set were successfully predicted. These results suggest that model 1 can potentially be applicable to genome-wide pol III type II gene finding. The performance of model 1 is further evaluated using real genomic sequences in the following subsection (5.1.2.3.2.).

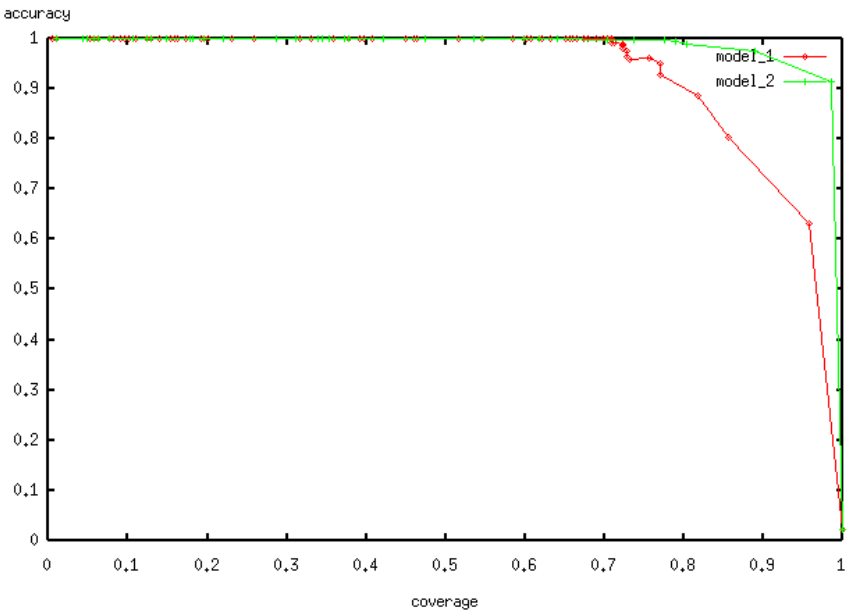


Figure 5-8. C-A plots of model 1 and model 2 on the test data set

5.1.2.3.2. The performance of model 1 – using human chromosomes 11 and 13

In order to assess the performance of model 1 in the context of real genomic sequences, this model was used to scan human chromosomes 11 and 13. In this subsection, a threshold corresponding to 100% accuracy and 66% coverage assessed against the test data set was chosen (Figure 5-8, model 1). It was found that the sizes of clustered hits were generally within the range of 1 to 3 bases, and none of them were longer than 5 bases (for definition of clustered hits see subsection 5.1.1.4.). This suggests that model 1 can detect pol III type II TSSs with good positional accuracy.

To compare the predictions made by using different methods, overlapped hits were determined as described in subsection 5.1.1.5. The methods discussed here include tRNAscanSE, eufindtRNA, and model 1. The predictions made by eufindtRNA were also compared here because eufindtRNA is a pure pol III type II promoter finding algorithm, not considering the structure-formation potential in a candidate region. In brief, eufindtRNA can be considered as an algorithm based on pure motif models. By contrast, tRNAscanSE is a hierarchical system which filters initial predictions made by other algorithms (*e.g.*

eufindtRNA, *etc*.), using structure-formation potential (for more details about how tRNAscanSE works see subsection 2.1.1.1. , chapter 2).

The results reveal that, for discriminating tRNA genes in the human genome, the performance of this model is comparable to existing algorithms (Figure 5-9 and Figure 5-10). Notably, the TSSs predicted by using model 1 and eufindtRNA frequently overlapped with MIRs. MIRs are mammalian interspersed repeats (Smit and Riggs 1995), which are tRNA-derived short interspersed repetitive elements (SINEs). The expected lengths of MIRs are ~260 bases. If the 300 bases upstream and downstream of the first base of each prediction were checked, as many as ~66% and ~51% of the TSSs predicted by model 1 on human chromosomes 11 and 13 respectively overlapped with MIRs (Table 5-3, model 1). Besides, ~57% and ~46% of the TSSs predicted by eufindtRNA on human chromosomes 11 and 13 respectively overlapped with MIRs (Table 5-3, eufindtRNA). In addition, 90.1% (20/22) and 100% (10/10) of the predictions made concurrently by both methods overlapped with MIRs (Figure 5-9 and Figure 5-10).

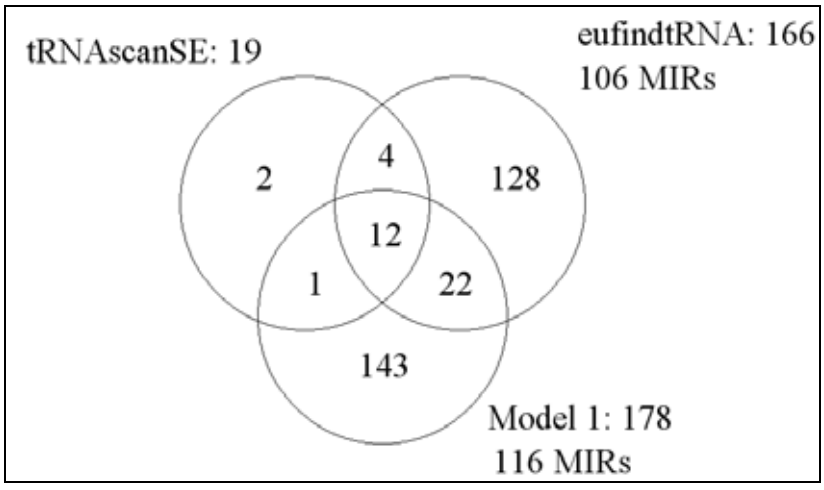


Figure 5-9. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 11

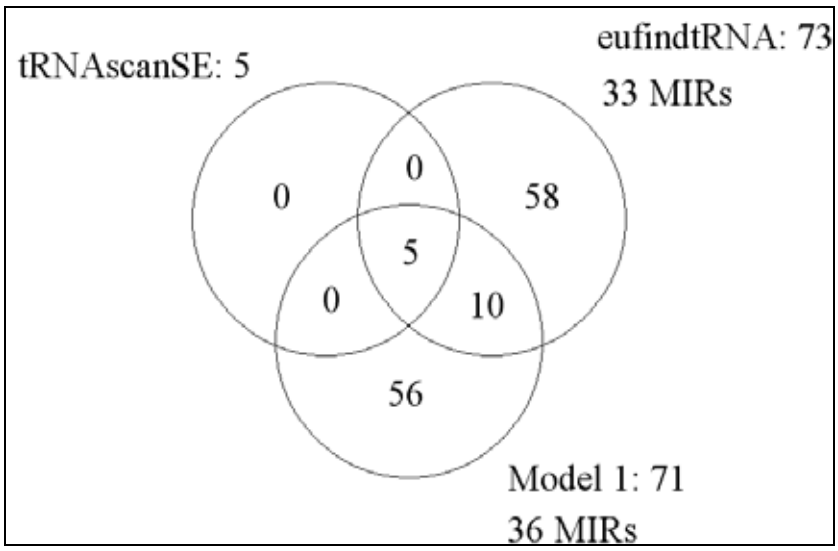


Figure 5-10. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 13

	Human chromosome 11	Human chromosome 13
EufindtRNA	63.9% (106/166)	45.2% (33/73)
Model 1	65.2% (116/178)	50.7% (36/71)

Table 5-3. Ratios of MIRs in different predictions for pol III type II genes on human chromosomes 11 and 13

MIRs – functional transcripts or pseudogenes?

It was surprising that more than half the pol III type II TSSs predicted by both model 1 and eufindtRNA are MIRs. Since less than 6% and 3% of the sequences on human chromosomes 11 and 13 respectively are MIRs, there is obviously an enrichment of MIRs in the sets of predicted TSSs.

In order to explore whether these predicted TSSs correspond to functional transcription units, two approaches were taken. Firstly, the MIRs predicted by both model 1 and eufindtRNA were used as negative sequences for training a revised EAS pol III type II TSS

model (see subsection 5.1.2.4.). If MIRs are pseudogenes, their promoters should have been at least partially degraded and thus including MIRs in negative training sequences may improve the specificity of the Eponine pol III type II TSS model. Secondly, the conservation of these MIRs in human-mouse syntenic regions was examined (see subsection 5.1.2.5.). If some MIRs are synteny-conserved, they are more likely to be functional elements.

5.1.2.4. Model 2 – using MIRs as the negative training sequences

The MIRs that were detected by both model 1 and eufindtRNA on human chromosomes 11 and 13 were added into the set of negative training sequences. The trained model (Figure 5-11) appears to be more complex than the model trained using random human genomic sequences as the only source of negative training sequences (Figure 5-6) however maintains the motifs of model 1. This new model is referred to as model 2. There are six distinct motifs at position 5, 15, 18, 18, 21, and 53. While the final motif in model 2 corresponds to the "B box", the "A box" is now represented by five motifs and there are overlaps between motifs. The performance of model 2 is slightly better than model 1 (Figure 5-8), since its accuracy is higher than model 1 when coverage is 90% ~ 100%.

5.1.2.4.1. The performance of model 2 – using human chromosomes 11 and 13

In order to compare the performance of model 2 with that of model 1 in the context of real genomic sequences, model 2 was also used to scan human chromosomes 11 and 13. In this subsection, a threshold corresponding to 100% accuracy and 55% coverage evaluated against the test data set was chosen when using model 2. Given this threshold, the number of predictions made by model 2 on human chromosomes 11 and 13 was comparable to that previously made by using model 1 (see the denominators in Table 5-4). Besides, model 2 had good positional accuracy, similar to that of model 1 (for the positional accuracy of model 1 see subsection 5.1.2.3.2.).

Using model 2 to scan human chromosomes 11 and 13, far fewer of the TSSs predicted

overlapped with MIRs than when using the previous model (model 1) (Table 5-4). Only ~16% and 10% of predictions on human chromosomes 11 (Figure 5-13) and 13 (Figure 5-14) respectively overlapped with MIRs. Besides, no MIRs on human chromosomes 11 and 13 were predicted concurrently by eufindtRNA and model 2. However, one problem with model 2 is that, the prediction coverage of tRNA genes on human chromosome 13 is decreased from 100% to 60% (Figure 5-14) and on human chromosome 11 is decreased from 68% to 63%. The result suggests that it is difficult to train a pol III type II TSS model that can completely avoid predicting TSSs which appear to be only associated with MIR elements.

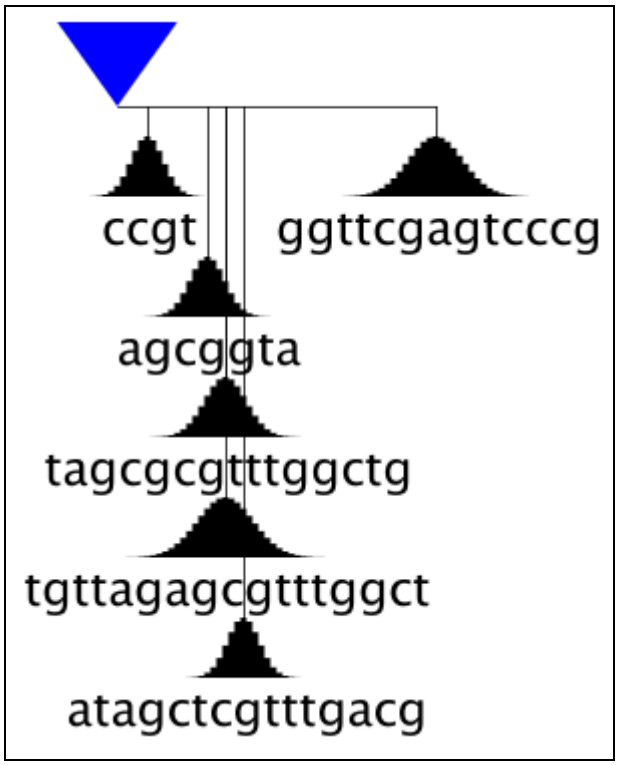
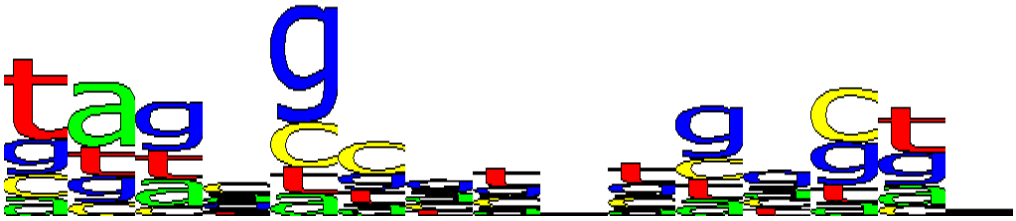


Figure 5-11. An EAS pol III type II model (using MIRs as negative training sequences) (model 2)

Weight: 4.18, position: 5, width: 2.66

Weight: 5.33, position: 15, width: 2.89



Weight: 8.11, position: 18, width: 3.44

Weight: 4.45, position: 18, width: 4.64

Weight: 18.85, position: 21, width: 2.66

Weight: 11.80, position: 53, width: 4.50

Figure 5-12. The sequence logos of position-constrained motif matrices of model 2 (Figure 5-11)

The annotation used in this figure follows the convention of Figure 5-3.

	Human chromosome 11	Human chromosome 13
Model 1	65.2% (116/178)	50.7% (36/71)
Model 2	16.0% (25/156)	10% (9/90)

Table 5-4. Ratios of MIRs in the predictions made models 1 and 2 for pol III type II genes on human chromosomes 11 and 13

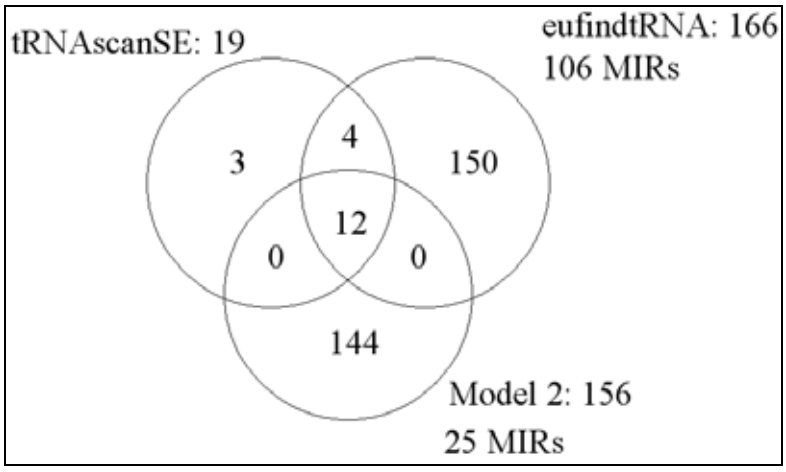


Figure 5-13. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 11

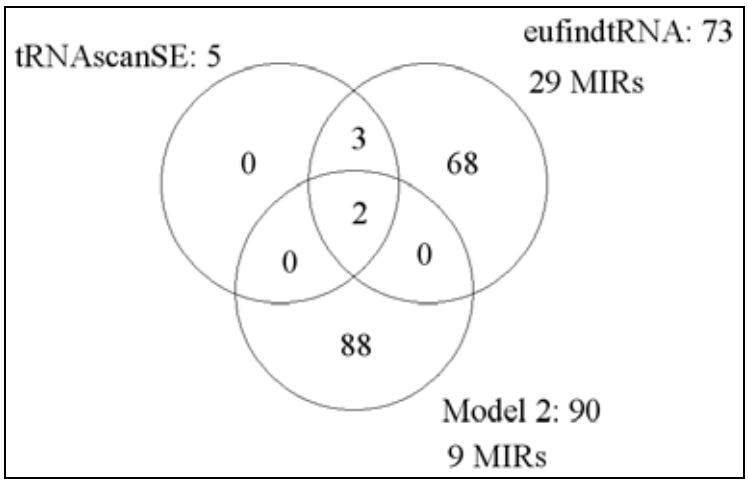


Figure 5-14. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 13

One interpretation of these results is that modelling TSSs alone, *i.e.* without considering the structure-formation potentials, is insufficient to distinguish functional pol III type II genes from inactive MIRs. However another interpretation is that the predictions are correct and that this finding implies that some MIRs are still being actively transcribed. There is evidence to suggest that transcripts of repetitive elements may not be completely non-functional. For example, mouse B2 RNAs, which are the transcripts of a class of tRNA-derived SINEs, can specifically bind RNA polymerase II holozymes to repress transcript synthesis in response to heat shock (Allen et al. 2004; Espinoza et al. 2004). The EAS pol III type II models also predict TSSs which are not associated with tRNAs or MIRs. While some of these predictions may be false positives, it is also possible that some correspond to novel functional genes.

Consequently, in the following subsection (5.1.2.5.), I explore the functionality of the predicted TSSs that do not correspond to tRNA genes. These sites may include MIRs as well as non-MIR elements. The synteny-conservation of these regions was taken as an indicator of their functionality. If regions near the predicted TSSs are conserved in the human-mouse syntenic regions, this supports the idea of them being functional transcripts.

5.1.2.5. Investigating the human-mouse synteny-conservation of the predicted pol III type II TSSs

The human-mouse synteny-conservation of the pol III type II TSSs predicted by model 1 and eufindtRNA were examined. The method used here followed the same procedures as described in section 2.1, chapter 2. The results reveal that only a few of the predicted TSSs on human chromosomes 11 and 13 are synteny-conserved (Table 5-5). Most of those that were synteny-conserved were found in the intronic regions of protein-coding genes (Table 5-6). In general, the identities between the human and mouse synteny-conserved signals are lower than 80%, except that on human chromosome 13 one pair of human-mouse synteny-conserved signals predicted by model 1 has 95% identity. Does this case represent a novel pol III type II gene? It is difficult to make this conclusion because the high identity may be evolutionarily constrained by the function of the protein-coding genes, but not necessarily by the function of any pol III type II genes. In addition, most of the alignments of the other synteny-conserved predictions in Table 5-6 contain many indels.

Therefore, the conclusion is that synteny-conservation provides no clear evidence to support the functionality of the predicted pol III type II TSSs not associate with tRNA genes.

	Methods	Non-tRNA predictions	Non-tRNA predictions in syntenic regions in the mouse genome
Human chromosome 11	Model 1	165	5 [1]
	EufindtRNA	150	5 [2]
	Model 1 and eufindtRNA	22	0
Human chromosome 13	Model 1	66	2
	EufindtRNA	68	0
	Model 1 and eufindtRNA	10	0

Table 5-5. The synteny conservation of the non-tRNA pol III type II signals on human chromosomes 11 and 13

[1]: there are 3 MIRs in these 5 synteny-conserved signals. [2]: all the 5 synteny-conserved signals are MIRs.

	Methods	Synteny-conserved signals	Overlapping with known genes	
			Protein-coding regions	Unknown
Human chromosome 11	Model 1	5	5 (introns)	0
	EufindtRNA	5	3 (introns)	2
Human chromosome 13	Model 1	2	1 (exon)	1
	EufindtRNA	0	0	0

Table 5-6. Distributions of the synteny-conserved pol III type II promoter signals in intronic and exonic regions

"Unknown" means that there are no genes annotated in the regions predicted to be pol III type II genes

5.1.3. Discussion

I attempted to model pol III TSSs using the Eponine system because of its success when applied to the similar problem of modelling RNA polymerase II (pol II) TSSs (Down and Hubbard 2002). However, the results from modelling of the TSSs of mammalian pol III type II genes have been less clear. Firstly, creating a general pol III TSS model proved impractical due to the substantially different promoter subgroups, so it was decided to concentrate efforts on modelling the largest pol III type II subgroup. It was possible to train models that could be used to scan entire human chromosomes predicting the TSSs of majority of known pol III type II genes (tRNAs) while making relatively few other predictions. However the proportion of other predictions was much higher than when Eponine was used to predict TSSs for pol II genes (Down and Hubbard 2002). Numerous TSSs predicted by using the EAS pol III type II model overlapped with MIR repetitive elements. A similar phenomenon was also observed when the tRNA-gene finder, eufindtRNA, which primarily identifies the internal promoters, was used. The biological significance of these MIRs that may have good pol III type II promoters is unknown. No evidence can be found to support the suggestion that these MIRs might generate functional transcripts.

There are a number of possible ways of explaining these results including the following:

- If we assume the majority of predictions that do not match known pol III type II genes are false positives, maybe this indicates that the Eponine system is not sufficient to model pol III type II TSSs completely. One possibility might be that the Monte Carlo method used in the Eponine trainer was unable to learn optimal PWMs representing the internal promoters of mammalian pol III type II genes with the datasets used here, which were smaller than used for pol II training.

- Alternatively, it might be that the internal promoters are insufficient for regulating the transcription of mammalian pol III type II genes, making apparently valid pol III type II predictions non functional. Other non-promoter regulatory regions, such as locus control regions (LCRs) and enhancers/silencers, might be necessary for the transcription regulation of mammalian pol III type II genes. The observation that tRNA genes tend to exist in clusters might fit with some additional regulatory process.

With respect to the first possibility, further exploration of promoter modelling using other motif-finding approaches to predict pol III type II TSSs could be considered as future work. Since the original goal of the first part of this chapter was to test Eponine as a quick approach for modelling the TSSs of mammalian pol III type II genes, a comprehensive assessment of the performances of other approaches for modelling and discovering the TSSs is beyond the scope of this chapter.

With respect to the second possibility I explored if it is possible to detect any evidence for non-promoter transcription regulatory regions associated with mammalian tRNA gene clusters. However, the initial attempt to look for signals in regions around these tRNA gene clusters (as described in section 2.2, chapter 2) was inconclusive (data not shown) and thus future work is needed.

5.2. Summary

In this chapter, an attempt was made to model the transcription regulatory regions of mammalian tRNA genes. In the first part of this chapter, the transcription start site of mammalian pol III type II genes, including tRNA genes and VARNA1 genes, was modelled by using the Eponine Anchor Sequence (EAS) model. Important findings are as follows:

- The performance of the EAS pol III type II TSS models is comparable to that of existing methods, such as eufindtRNA, for identifying the TSSs of tRNA genes.
- Both the EAS pol III type II TSS models and the internal-promoter based tRNA gene finder may predict many repetitive elements, MIRs.
- By using MIRs as the negative training sequences, the performance of the new EAS pol III type II model cannot be further improved.

One future work is to try other motif-finding approaches to predict pol III type II TSSs. Another future work is to search for non-promoter regions regulating transcription of pol III type II genes that are clustered in mammalian genomes.