MODELLING TRANSCRIPTION TERMINATION SIGNALS

3.1 Introduction

My initial objective here was to develop a transcription termination model and thereby determine the correct gene 3'-end. In general when annotating a genome, poly(A) signals and/or cleavage sites are generally considered to be the end of a gene. Almost all the gene prediction methods available so far use this principle. However predictions of poly(A) signals along the chromosomes are likely to be dense, as the probability of occurrence of the hexamer in 3 billion human sequences is high. Existing gene prediction programs screen out false poly(A) predictions by conditioning it on numerous other parameters like coding potential, exon-intron structure and ORF length. As here the objective is to construct an *ab initio* gene model purely based on gene regulatory signals without considering such parameters, it is imperative to consider other gene 3'-end motifs apart from the poly(A) signal to limit false predictions.

In vivo and *in vitro* experiments to identify such 3'-end consensus motifs responsible for transcription termination have not been so far successful as discussed in chapter 1. However these experiments have established that the poly(A) signal and auxiliary sequences are essential for efficient termination of RNA polymerase II.

Extensive computational analysis to identify transcription termination signals has not previously been attempted. Gene 3'-end identification programs, such as *ERPIN*, *Polyadq* and *PolyAH* depend only on poly(A) variants and auxiliary motifs. Here I discuss the results of the algorithms with which I tried to detect additional RNA polymerase II transcription termination signals.

3.2 Datasets

Given the difficulty in experimentally annotating the poly(A) signal and cleavage site of each gene in the human genome, extracting a large dataset with precise location of gene 3'-ends has been a challenge. However with the recent publication of the latest version of human chromosome 22 (Collins *et al.,* 2003), a set of 422 genes with high quality

annotation was obtained. The methodology adapted for annotating chromosome 22 included three main approaches –

First, a *variety of programs* with different statistical methods were used to predict gene structures cautiously, as these methods are likely to produce incorrect and over predictions. Second, a match of *Expressed Sequence Tags* (EST) of transcribed genes with the genomic sequence gave direct evidence of expressed genes. Third, *comparative genomics* with related species helped to identify conserved gene structures.

Gene structures were identified using evidence from transcribed sequences across their entire length. Full length cDNAs or assembled ESTs were aligned to genomic DNA to resolve the splice sites and confirm 3'-ends. A 3' end was judged confirmed if it had a run of at least four adenine residues at the 3' end of cDNA/EST not present in the genomic sequence. Thus the 3'-ends with the processing signals were manually verified and confirmed for *in vivo* biological function.

Based on further evidence, this set of entries with confirmed 3'-ends were classified into *complete protein coding genes*, *partial genes*, *non-coding genes* and *pseudogenes* with the following definitions.

(1) A *complete protein-coding gene* has sequence identity to human cDNAs or ESTs across its entire length and a predicted ORF of at least 300 bases.

(2) A *partial gene* had sequence similarity to cDNA, EST or peptide sequence but did not comply with complete gene criteria.

(3) *Non-coding RNA genes* included small RNAs and published complete genes that did not contain ORF of at least 300 bases.

(4) A *pseudogene* had similarity to a known gene or protein but had evidence of disrupted function.

Using these criteria, 393 complete protein coding genes, 153 partial genes, 31 non-coding transcripts, 234 pseudogenes and 125 IGLV and J gene segments (Ig gene segments) were annotated for chromosome 22. Among all these categories, 376 protein coding genes, 56 partial genes and 15 non-coding transcripts are found to have confirmed 3'-ends with the hexamer variant and cleavage site. Out of these 447 genes (376+56+15), I extracted sequences from 422 genes in the interval of -200 to +2000 bases relative to the cleavage site to form the *positive dataset*. The remaining 25 sequences (447-422) had an overlapping transcript within the 2000 base pair downstream sequence of the cleavage site. A set of 22 sequences from these 422 entries were set aside and used as an independent test set, leaving the remaining 400 sequences for training purposes.

For training Eponine Anchored Sequence (EAS) models, the RVM requires a *negative dataset* that does not have 3'-end processing signals. Choosing an appropriate negative set for training purposes is a determining criterion for making sensible Eponine models. So I extracted sequences from different sources and will briefly explain these sets while discussing the Eponine sequence models.

3.3 Nucleotide composition analysis

Figure 15 shows the average base composition of 422 sequences for 200 bases upstream and 2000 bases downstream of the cleavage site. The undulations in the graph are seen concentrated near to the cleavage site then in other regions of the sequences. The zoomed figure (Figure 16) with base compositions for -100 to +50 bases from the cleavage site shows two significant peaks for adenine nucleotide distribution. The first broad peak spans -30 to -5 base pairs while the second, peaks at position 0. Followed by the cleavage site the adenine concentration suddenly decreases with increase in thymine composition. The thymine level represents the U-rich sequence observed near cleavage sites. The guanine and cytosine concentration is generally relatively low and more equal to the background distribution.

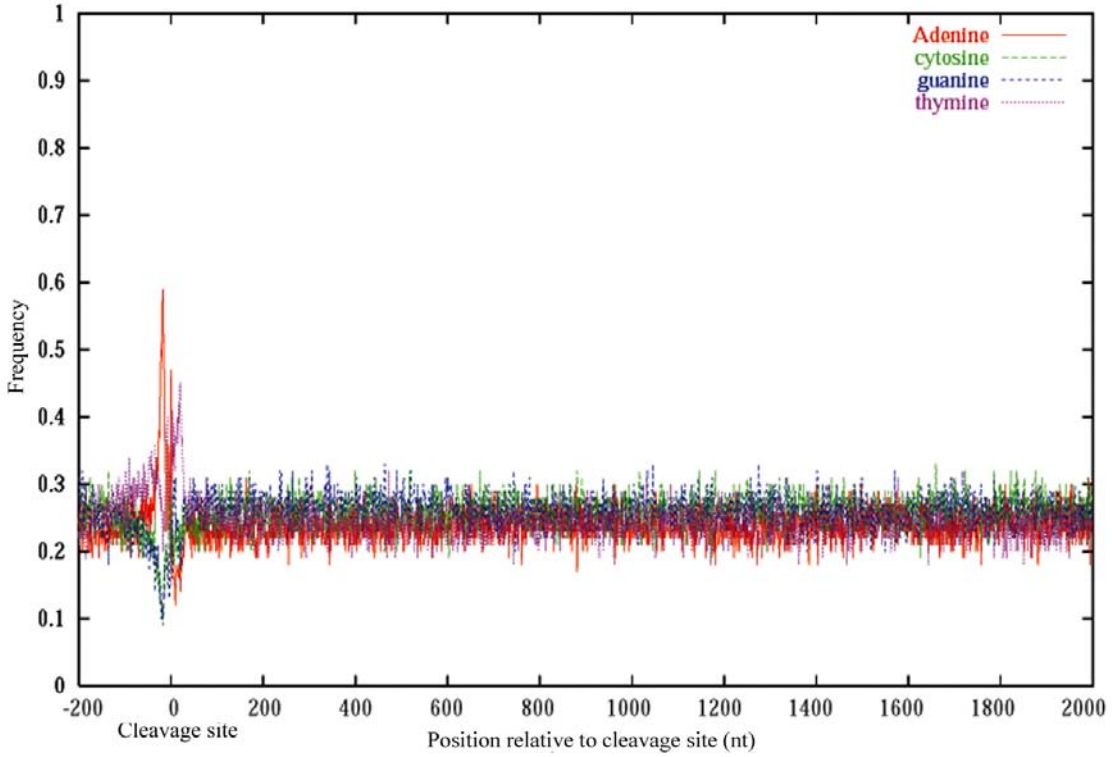

Figure 15. Nucleotide composition spanning -200 to 2000 bases relative to the cleavage site

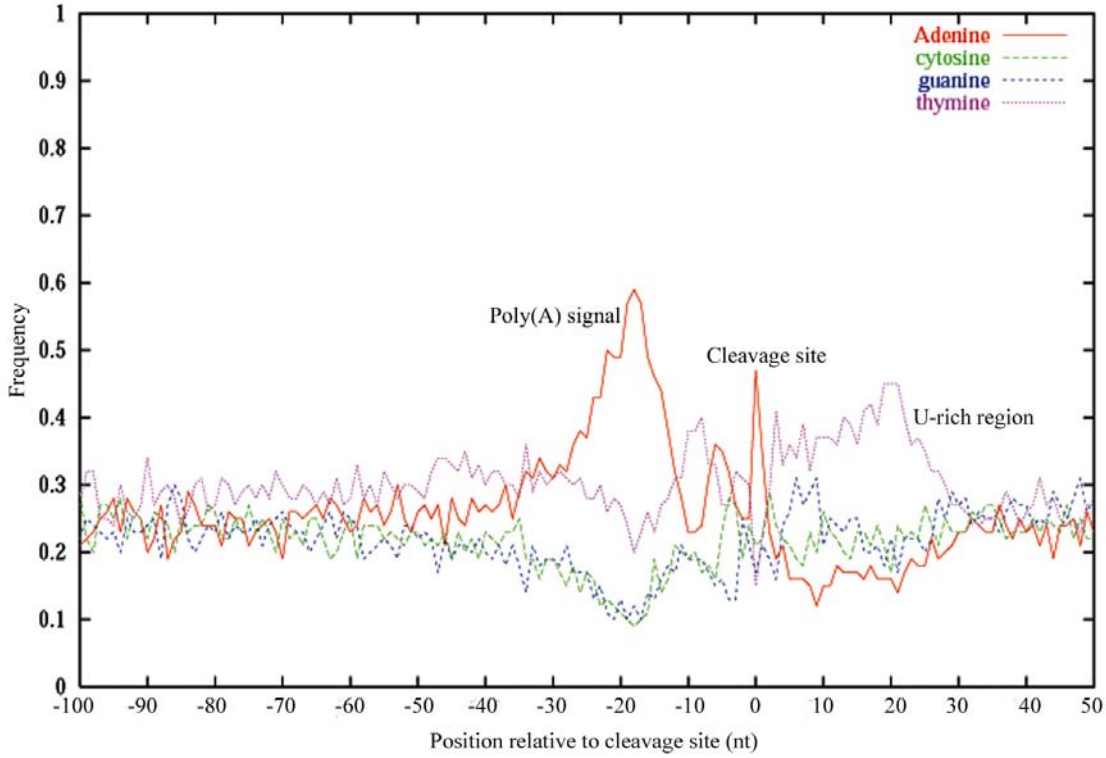

Figure 16. Nucleotide composition spanning -100 to 50 bases relative to the cleavage site

There is no difference in base composition in the sequences 50 to 2000 bases downstream of the cleavage site (Figure 15). The termination signals responsible for RNA polymerase pause and release are expected to be present in this region. However the graph shows no difference in the nucleotide distribution from background. This might be due to the un-alignable nature of termination motifs as they are known to occur at varied positions for different genes (Dye and Proudfoot, 2001). Even if there is a small bias in nucleotide composition for single sequences, by averaging, no bulk effect is seen.

Thus the nucleotide composition analysis has identified the 3'-end processing signals known previously. However I find no significant compositional variation in the downstream region where the polymerase pauses before terminating.

3.4 Secondary structure analysis

As the simple nucleotide composition analysis does not reveal the pause signal at the downstream region, I decided to search for stem-loop secondary structures known to play a major role in prokaryotic transcription termination (Henkin, 1996). In the eukaryotic genome, stem-loop structures and their role in terminating transcription has been established for histone genes. Now the question is, whether similar structures are present in eukaryotic protein coding genes or not. If present, stem-loop structures are likely to be found at the downstream sequences of the cleavage site where actual pausing of RNA polymerase II occurs. This differs from histone stem-loops as in these genes the structure was found upstream of 3'-end processing signals. As explained in chapter 1, special proteins like SLBP bind to these structures, to stabilize them and hinder poly(A) tailed histone mRNA formation. Unlike this, a potential structure at the downstream region in the protein coding gene might explain the drag, pause and queuing of the polymerase before termination. So to search for any stem-loops I used two simple algorithms developed by *Ruth Nussinov* and *Zuker* (Nussinov, 1978; Zuker, 1994).

3.4.1 Nussinov algorithm

Nussinov algorithm (Nussinov, 1978) is one of the simplest approaches to predict secondary structure of RNA molecules. The method is based on finding the configuration with the greatest number of paired bases. Apart from normal Watson-Crick base pairing used for

calculation, *G:U* pairing is also allowed. Since testing and scoring each possible structure is computationally expensive, the algorithm uses a dynamic programming to find an appropriate solution. For details of the algorithm and implementation refer to chapter 2.

I scanned the 422 sequences in the positive dataset with the overlapping window sizes of 15, 25, 35 and 50 bases with an interval of 5 and 10 bases between each window. As explained in chapter 2, the initial algorithm was modified to consider the loop length parameter. The base pair metrics score was calculated for each window allowing at least 3, 5, 7 and 10 bases in the loop region. Figure 17 shows the average metrics score of the sequences scanned with overlapping window sizes of 60 bases, 10 bases between each window and loop length of 5 nucleotides.

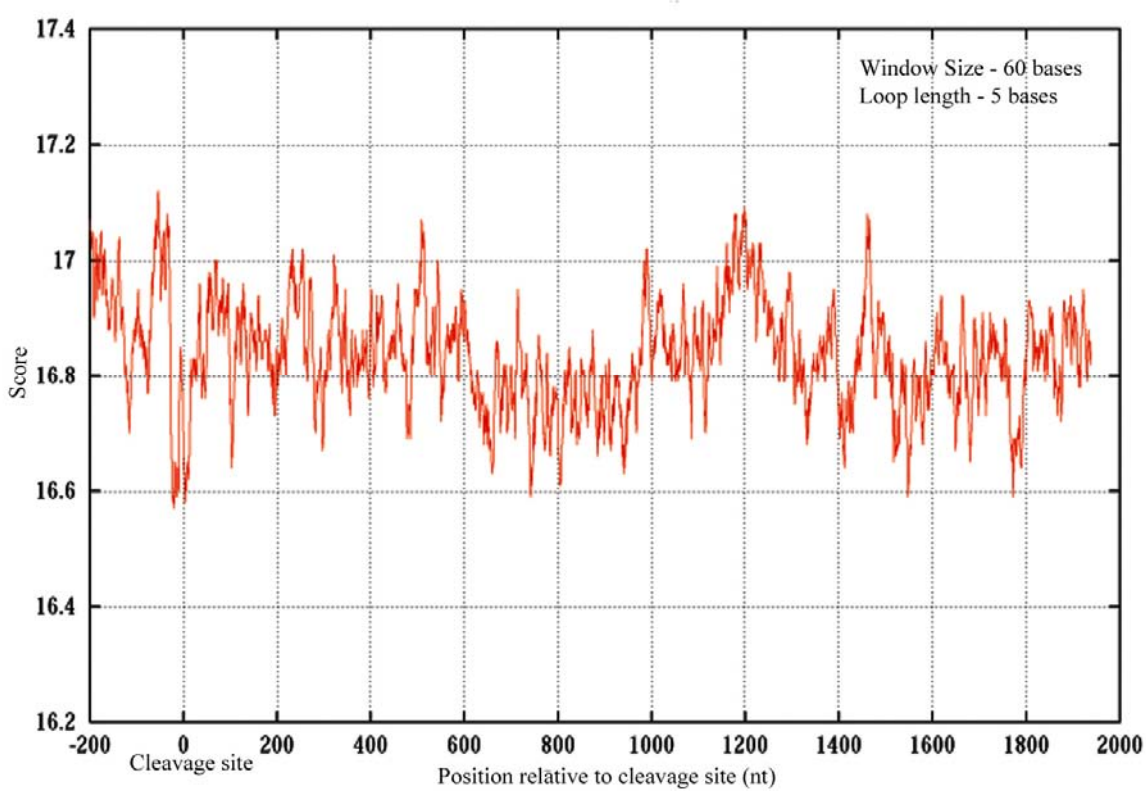

Figure 17. Averaged score values of sequences around cleavage site calculated using Nussinov algorithm

The X and Y-axis represents sequence length and score respectively with the cleavage site referred to as position 0. The graph shows the scores are spread out evenly with no significant peaks or troughs. This might be due to the following reasons –

(a) no stem-loop structure in protein coding genes.
(b) the algorithm is simple and observing the base pair maximization metrics alone might not be sufficient to detect any secondary structure.
(c) the algorithm does not consider non-Watson-Crick pairing except *G:U* base pairing.
(d) base stacking metrics that explains the structure stabilization is not used.
(e) position of stem-loops might be varied for each gene and averaged metrics score did not show any bulk effect.

The result emphasis the need for an algorithm with additional metrics and thus I used the one published by Zuker.

3.4.2 Zuker algorithm

Zuker algorithm (Zuker and Stiegler, 1981) is one of the most well-known algorithm to predict RNA secondary structures based on the free energy minimization principle. The algorithm is optimal for predicting secondary structures of RNAs with no pseudoknots. According to the algorithm, all the structures of RNAs can be decomposed into either sequential or nested structures. This algorithm is implemented in the well known RNA secondary structure programs like *RNAfold*, *mfold* and *ViennaRNA*.

Here I used the *ViennaRNA* implementation to look for the presence of secondary structures in the positive dataset. The algorithm calculates free energies for all possible structures and determines the one with the least value to be the most probable structure. The lower the predicted free energy value, the more likely the structure is thermodynamically stable and likely to persist.

I scanned the 2200 base sequence with overlapping window sizes of 20, 30, 40, 50, 60, 70 and 80 bases, skipping 5 or 10 bases between each window. The free energy scores for each window size (with a gap of 10 bases between windows) along the length of the sequences are plotted in Figure 18.

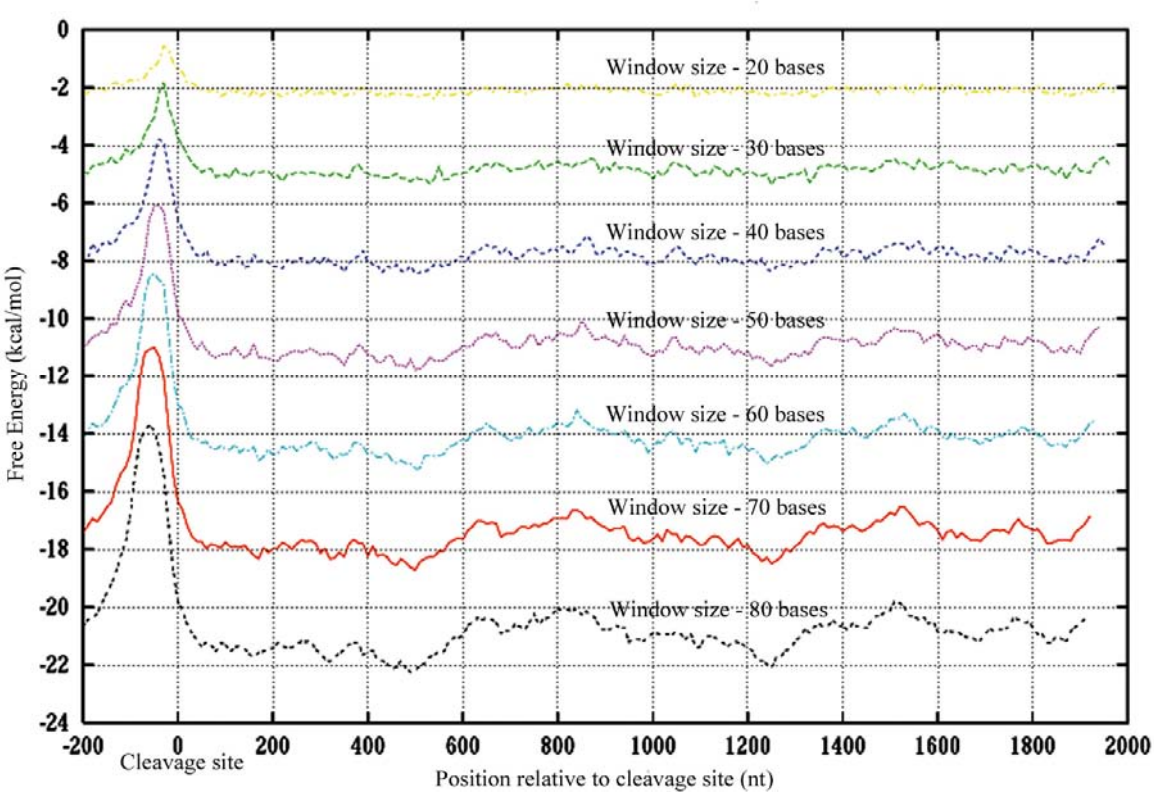

Figure 18. Averaged free energy values of sequences around cleavage site calculated using Zuker algorithm

The Y-axis represents free energy values measured as *kcal/mol* while the X-axis shows sequence length as *nucleotides*. Here, I analyse the plot by splitting it into four regions and comparing the energy values to the average value of the negative set of sequence.

(i) Region 1 extending from -200 to 0 bases shows a statistically significant peak at the poly(A) region. The higher energy value means there is less probability of a secondary structure at this region.

(ii) Region 2 from 0 to +100 bases from the cleavage site covering the U-rich sequences of the 3'-end processing signals has a free energy value less than the average at 95% confidence level. However the significance of the energy scores is not prevalent when the window size is reduced to 20 bases.

(iii) Region 3 comprising +100 to +650 bases from the cleavage site shows the energy values are less than the average value at 99% significance level when the sliding

window parameter was fixed at 60 bases. The condition was found true even at lower window sizes. This shows there is possibility of a RNA secondary structure in this region. However the broad distribution of free energy scores indicate the possible stem-loops are distributed at varied positions for different genes. On averaging energy values for the 422 sequences at this region, the overall distribution is likely to get flattened rather than appearing as a sharp trough. This agrees with earlier understanding of the presence of termination related pause signals at varied distances from cleavage site (Dye and Proudfoot, 2001).

(iv) Region 4 from +950 to +1350 shows a decrease in free energy values from the average at 95% significant level in the window size of 60 bases. However the significance score reduced with diminishing window sizes indicating there might not be any secondary structure in this region.

The free energy parameters used for calculating the scores were derived from recent experiments and the values are likely to be influenced by sequence artifacts. Hence, I have plotted the GC and GT densities along the sequence calculated with the 60 bases sliding window (Figure 19 and Figure 20). GC and GT richness in the sequence affect DNA base pairing and emit low free energy values. These low energy values based on the sequence composition bias need not necessarily mean there is a RNA secondary structure. So to avoid this misinterpretation, I did correlation studies of GC and GT density with free energy values of the corresponding regions.

For region 1, I found a strong negative correlation between the free energy values and GC and GT density. Guanine and cytosine compositions are expected to be less around this region as the poly(A) signal and the cleavage site increases richness in A and somewhat in T density. This increased adenine and reduced guanine and cytosine concentration might be the reason why the free energy values in this region are remarkably high.

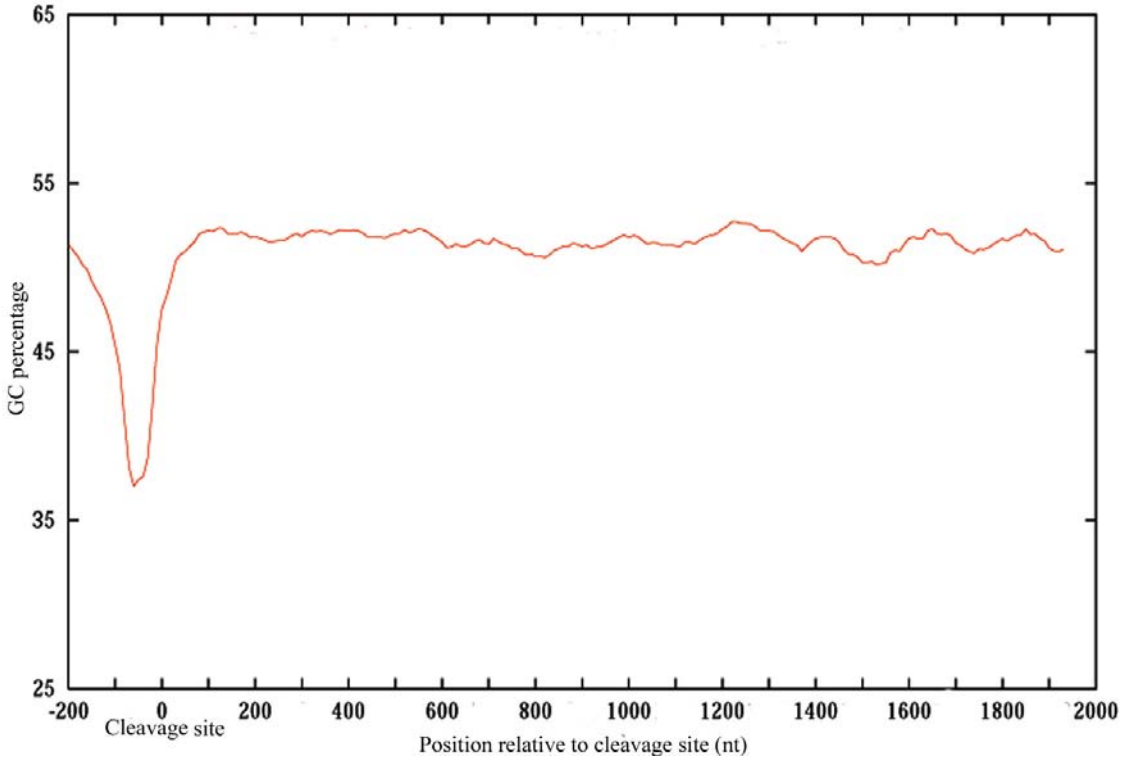

Figure 19. Percentage of GC residues in the sequences around cleavage site

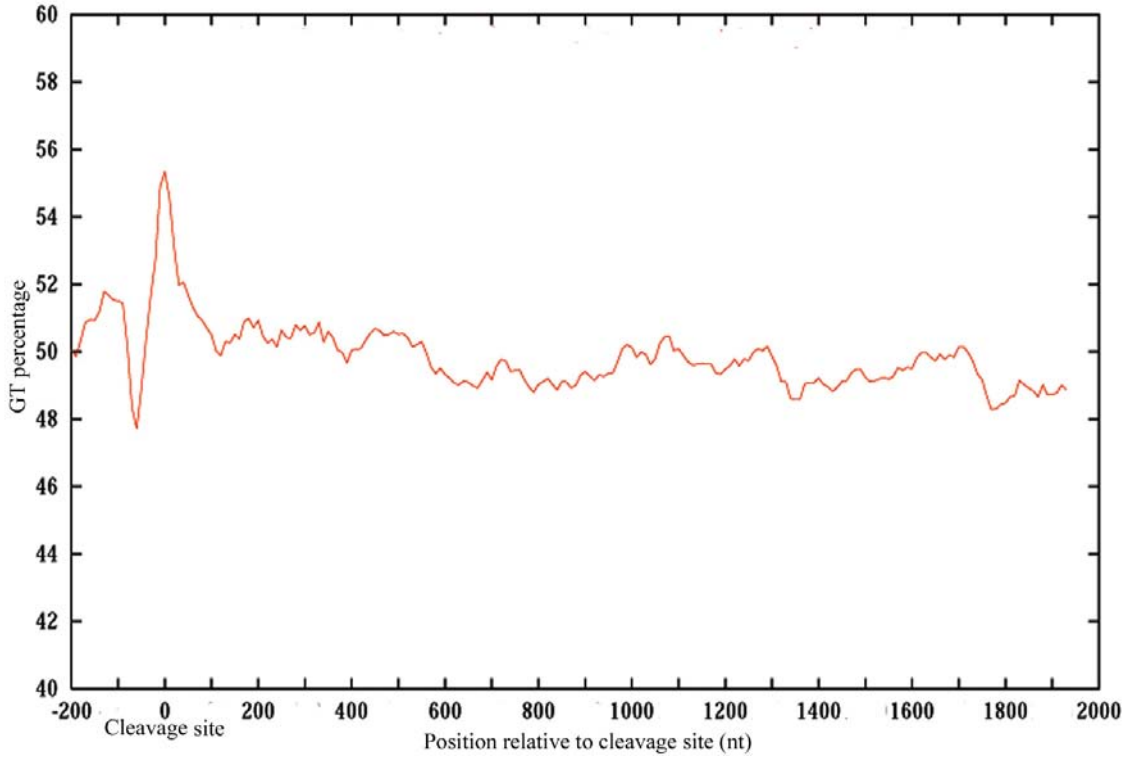

Figure 20. Percentage of GT residues in the sequences around cleavage site

Region 2 does not have any correlation between GC densities and free energy values, whereas GT density has a strong positive correlation coefficient. This is due to the U-rich sequence of the 3'-end processing signals followed by the cleavage site. Thus the energy values at this region are influenced by the base composition.

Region 4 has positive correlation between GC density and free energy values, whereas the GT density does not show any. Thus the lower than average free energy found at a 95% confidence level in this region might be due to the influence of GC base stacking.

Unlike other regions, sequences in region 3 have no correlation between GC/GT densities and free energy metrics scores. So the likely secondary structure present between +100 to +650 bases may not be due to nucleotide composition bias. However the significance of lower free energy values decreases with the size of the sliding window. Hence the results show there is scope for stem-loops in the region, however detailed biochemical experiments are required to confirm their existence.

Thus the Zuker algorithm predicts there is a possibility of RNA secondary structure from 100 to 650 bases from the cleavage site and that they are less likely to be caused by sequence artefacts. Further experiments can confirm the presence of any such structure and help us to understand the polymerase pause and release from the DNA.

3.5 Eponine transcription termination model

With the results from nucleotide composition and secondary structure analysis, I then resorted to Eponine described in chapter 2 to train a transcription termination model. Eponine is a probabilistic sequence classifier based on a relevance vector machine and requires a set of positive and negative sequences to learn informative basis functions to classify them. As explained earlier, the major positive dataset for training an Eponine model was derived from human chromosome 22. Similarly the major negative dataset was extracted from random sequences from the transcription units. However, apart from these datasets, various others were also used. Here I explain in detail how the model was constructed and cross-validated. Following it, I compare the performance of the model to the poly(A) prediction program, *ERPIN*. The Eponine model, apart from detecting the annotated gene 3'-ends, made other predictions, which are referred to as *false positives*.

Towards the end of this chapter I explain the distribution of these false positives and two hypotheses explaining the possible role of such predictions in gene regulation.

3.5.1 Training the transcription termination model

One important criterion while learning any classification model is to choose an appropriate negative set. This set of sequences forms the basis for differentiating it from the positive set provided for training. Here I attempted different sets of negative sequences which are explained below –

(i) I extracted sequences of 2200 bases each from chromosome 1 by choosing random points. A pseudo-random number was generated using PERL *rand()* function and 2200 bases from that point was dumped from ENSEMBL database in FASTA format. Similarly, another set of random sequences from chromosome 20 was extracted by generating a number between 1 and 62 x 10^6, as the length of chromosome 20 is approximately 62 mega bases.

(ii) Another set of random sequences were extracted from chromosome 1, as described earlier, after repeat masking of the whole chromosome using *Repeat Masker* (Smit and Green, 1996). Random regions of the chromosome are chosen in such a way as that no repeat masked bases were part of the 2200 base negative sequence.

(iii) RNA polymerase is not expected to terminate in exon sequences and thus another set of negative sequences was derived from chromosome 22 exons. With the quality annotation available, sequences of all the exons annotated in chromosome 22 were extracted after leaving 100 bases near the donor and acceptor splice sites. These sequences were then concatenated together to form a single sequence. Then as explained previously for random sequences, a set of sequences of 2200 bases each were randomly dumped from this single sequence.

(iv) Similarly intron sequences were extracted from all annotated last introns in chromosome 20 and 22. Intron sequences from these two chromosomes were dumped after removing the 100 bases near the donor and acceptor splice sites. This is done to

avoid representing the splice signals in the negative dataset. A set of sequences, each of 2200 bases in length, was randomly picked from the concatenated intron sequences.

(v) With the gene annotation of chromosome 22, sequences from the transcription units (including exons and introns) were extracted after leaving 250 bases near to the gene 3' end. The extracted sequences were concatenated and from it, 422 random sequences of 2200 bases each were dumped to form a negative set. Although weak terminators or pause signals are likely to be present in the transcription unit and thus in the negative set, they formed one of the best training sets for learning transcription termination models.

These different sets of negative sequences along with the positive dataset were used for training the transcription model. In each case of training an equal number of positive and negative sequences were used. The cleavage site formed the anchor point for the EAS model. The models were trained for approximately 10000 cycles using the VRVM trainer as described in chapter 2. At various points in the training process, I dumped 'checkpoint' models to identify the basis functions learnt. Initial models picked more basis functions and as training progressed they gradually converged leaving fewer basis functions explaining the dataset. One such final model is shown in Figure 21. The models have generally two sets of basis functions. The positively weighted position constraints, represented in black and negatively weighted position constraints, coloured in blue. The positive constraints are cases where the motif presence is likely to determine the termination site, whereas negative constraint makes the possibility less likely.

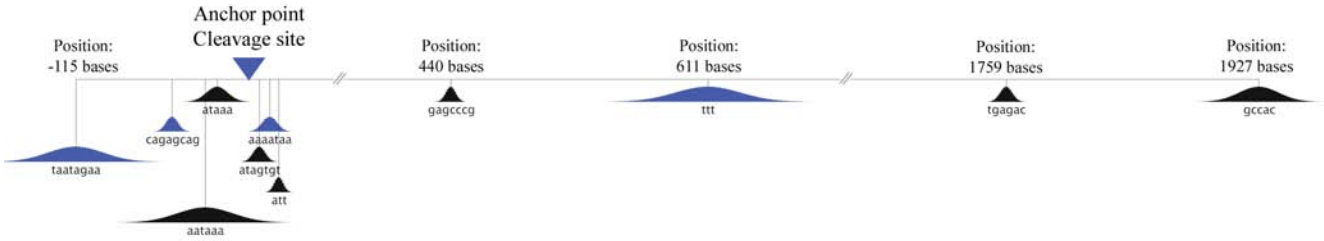

Figure 21. Transcription termination model trained from chromosome 22 sequences

Various models gave a consistent set of motifs learnt from the datasets. The poly(A) hexamer, AAUAAA was represented in the model around -20 to -30 bases as expected.

Similarly the GT/T-rich region found earlier through various studies can be found in the model immediately after the anchor site. The most interesting part I found in the models is the consistent appearance of the downstream motifs, which hereafter I refer as *Pause Elements* (PE), as such elements had earlier been shown experimentally to pause transcribing polymerase (Aranda and Proudfoot, 1999; Enriquez-Harris *et al.*, 1991; Yonaha and Proudfoot, 1999). The motif's positions varied with different models and training, however the sequence of the motifs did not change much. Various motifs found in these regions are listed in Table 1. Although there is diversity in the motifs, the positive constraints can be generalized into two types - Poly-C variants (CCCC, C with intermittent A or G) and GGAGG variants (GGAGG with intermittent A and C). Negative constraints as explained above are motifs not expected in the termination region and thus a non-T rich motif or its variants (with intermittent A/C) are likely to be present downstream of the cleavage site. The stretch of non-T residues in the downstream region is interesting as prokaryotic and polymerase I termination are likely to occur with a run of T residues. However the model indicates polymerase II termination requires the opposite and this agrees with results from experiments done on human β-globin and α2 globin genes (Dye and Proudfoot, 2001; Enriquez-Harris *et al.*, 1991). Table 2 lists the occupancy value of top 15 motifs learnt in different runs of training in a scale of 0 to 1. A value more than 1 indicates the motif is represented more than once in few models.

Table 1. Consensus motifs found in sequences between 50 and 2000 bases from cleavage site

Motifs found in sequences where RNA polymerae is likely to terminate	CACCC AGCCAG CACCC AGGGACAC GCCACC CCCGCCC CCCCC GAGCCG CACCCG AGAGGG GACTC CCCGG GCAGGG CCGCCC GGGC GCGC GGG GGGGGGA AGTGTG
Motifs unlikely to be found in sequences where RNA polymerae is likely to terminate	TTTT TTATTT TTTACA TTGCAA

Table 2. Occupancy value for motifs detected in the transcription termination models.

Number of models considered - 106	
Occupancy value for motifs between -200 and 0 bp	
Motifs	**Occupancy Value**
aataaa	1.09
aataa	0.22
ataaa	0.09
attaaa	0.07
caataaa	0.04
attaa	0.04
aataaaa	0.04
taaa	0.03
taataaa	0.03
aaaaaaa	0.03
aaaaga	0.03
aaaaaa	0.03
aataaag	0.03
aaataaa	0.03
agagca	0.03
Occupancy value for motifs between 0 and 30 bp	
tgtgt	0.05
tgtgtc	0.04
agtgt	0.04
aatt	0.04
aaaataa	0.04
att	0.03
ct	0.03
aaaaaaaaaaaaa	0.03
tt	0.03
gtctg	0.02
aa	0.02
tttg	0.02
tcgtgtgt	0.02
tgtgtgtgtctt	0.01
tgtgtcga	0.01
Occupancy value for motifs above 30 bp	
tt	0.19
ttt	0.11
ttttt	0.05
tgagaa	0.03
cc	0.03
tttt	0.03
att	0.02
ccag	0.02
taa	0.02
tttc	0.02
cctcct	0.02
cg	0.02
gtttt	0.02
atttt	0.02
at	0.02

Also the position constraints found in the sequences 2 kb downstream of the cleavage site are repetitive (Table 1). This emphasize that the signals are present in multiplex and effective termination might depend on the cumulative effect of all the signals, agreeing with the experimental results found earlier (Aranda and Proudfoot, 1999).

3.5.2 Window size

I focused on a 2 kb window downstream of the cleavage site for modeling as the detailed experiments in human β-globin and ε-globin genes showed 2000 bp is enough for termination of polymerase II (Dye and Proudfoot, 2001). However I also tried sequences of varied lengths - 500, 1000, 1500, 2000, 3000 and 4000 bases downstream of the cleavage site. As I increased the downstream window size from 500 to 2000 bases the performance of the models improved. However there was no improvement when the window size moved from 2000 to 4000 bp. This suggests that 2000 bases are enough for termination in most cases. What signals found downstream in the 2000 to 4000 base sequences appear to be repetitions of the positive and negative constraints explained before. Even if a larger window is included, the trainer did not learn any position constraints beyond 2600 bases from cleavage site and thus model a compact transcription termination region.

3.5.3 Cross validation

As explained above, the inherent problem with most comparison based classifiers is the use of a proper negative dataset sequence while training. The motifs detected by the models may represent biases in the negative dataset used, rather than signals in the positive dataset. So to cross check if the position constraints discussed in the earlier models are consistent with signals in the positive dataset, I used different negative datasets for the training. Different training parameters gave similar motifs and the positions of the motifs are also found to be conserved. This suggests that the motifs learnt by the classifier are not biased by the training datasets used and are conserved in the sequences and may have some biological function.

The positive dataset for the model discussed above is derived from chromosome 22. To cross validate the PE detected by the earlier model and to provide evidence that they are not just due to some strange distribution of chromosome specific sequences; new models were trained using a positive dataset from chromosome 20. Gene annotation from VEGA

database was used to extract 200 bases upstream and 2000 bases downstream of the cleavage site from chromosome 20. One of the models trained with this dataset is given in Figure 22. The signals detected by this training are similar to the motifs described in Figure 21 and Table 1. The similarities between these two independently trained models suggest that these PE motifs may be a general feature of human gene 3'-ends.

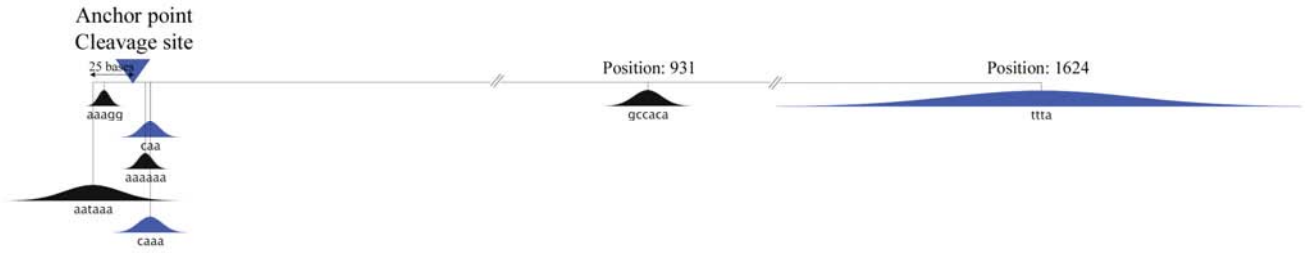

Figure 22. Transcription termination model trained from chromosome 20 sequences

To investigate whether the PE motifs detected by the model were likely to be due to repetitive elements found in the human genome, the positive and negative datasets were repeat masked using *Repeat Masker* and training was done on a masked set. Models from various training cycles still learnt the PE discussed above. Thus the pause elements are not part of any repeat sequence although they occur multiple times in the 2000 bases downstream of the cleavage site.

Chromosome 22 has a higher GC content than the average for human chromosomes and the PE represented in the model might mimic the CpG island in the training datasets. To rule out this possibility, the positive dataset was scanned for CpG island using *CpG Report* (Micklem) with the default parameters of 100 bases sliding window, minimum length of 200 bases CpG island, 0.6 ratio for observed-expected value and 50% of G and C composition in the window. The scanning found that only 62 sequences out of 422 sequences had CpG island-like sequences even at a conservative threshold score of 80. Thus with the maximum of only 14.69% of the positive dataset containing CpG island-like sequences, the likelihood of learning a CpG motif is less. Hence the PEs are not mimics of CpG island signals.

3.5.4 Model refinement

Different training cycles showed the PE sequence motifs are present at varied positions downstream of the cleavage site. Hence I decided to try to refine the models by optimizing parameters to better by capture the position distribution of the PE. Important among the parameters modified is the Gaussian distribution width. As explained in chapter 2, the position of the basis function relative to the anchor point is captured as a Gaussian distribution. Since the default width allowed for the maximum expansion of the Gaussian distribution was found not to be ideal for learning the multiplex motifs of the PE, the width was optimized with various training cycles. Models trained with the new parameters gave PE with wider distribution and higher constraint weights. One such model is shown in Figure 23. The parameter file used to create the model is given in Appendix C. The new optimized parameters also removed the negative constraints present in earlier models and replaced them with positive constraints. Table 3 lists the positional constraints and Gaussian widths of the motifs seen in Figure 23.

Table 3. Position contraints learnt while training chromosome 22 sequences

MOTIFS	POSITION	CONSTRAINT WEIGHT	GAUSSIAN WIDTH
	-29	13.78	17.24
	-22	7.43	7.12
	8	9.72	0.73
	309	20.80	123.22

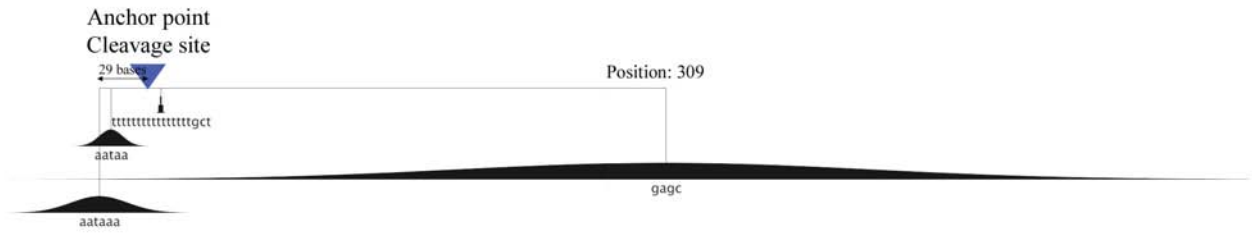

Figure 23. Transcription termination model trained from chromosome 22 sequences with modified parameters

Eponine package is available in different flavours apart from the anchored sequence model widely used in this project. I tried three different versions for learning transcription termination signals and those are –

(i) Windowed Sequence Model – In this method, an 'anchor point', positioned in EAS training on the cleavage site, is not used. Position constraints that are learnt from the set of training sequences using the same principles of the EAS are only *between* basis functions. I used an equal number of sequences (422 sequences) from confirmed 3'- gene ends of chromosome 22 and transcription unit as positive and negative set for training respectively. The trainer was allowed to run for approximately 4000 cycles and a model for every 500 cycles was dumped and tested for its performance. Interestingly, the motifs represented by the basis functions are different to those of the EAS model and the model performed poorly. This indicates that the windowed sequence model might not be suitable for the problem under consideration.

(ii) Hierarchical Sequence Model – This method works similarly to EAS except that position constraints are allowed to form anchor points for other position constraints in a hierarchical manner. This extension to the basic classifier may be helpful to better model situations where there are minor positional differences between features for given sets of sequences. For instance, TATA box might be used as a major constraint to classify promoters from other sequences. However, there are variations within different types of promoter sequences and few constraints can in turn be used to capture different promoter elements, say different transcription factor binding sites. Each transcription factor binding site has variations between different cell lines and few constraints can distinguish them (Figure 24). Thus the overall model with major constraints classify the dataset; and each

major constraint is made from minor position constraints; and each minor constraint is made of other constraints capturing the few differences within the different set of sequences, making a hierarchical structure. This method was used to train models from the 422 positive and negative sequences as before, allowing the trainer to run for 9000 cycles. The model captured similar signals as the anchored model and a hierarchy structure was learnt near the cleavage site. However no hierarchical structure was found in the sequences downstream of the cleavage site.

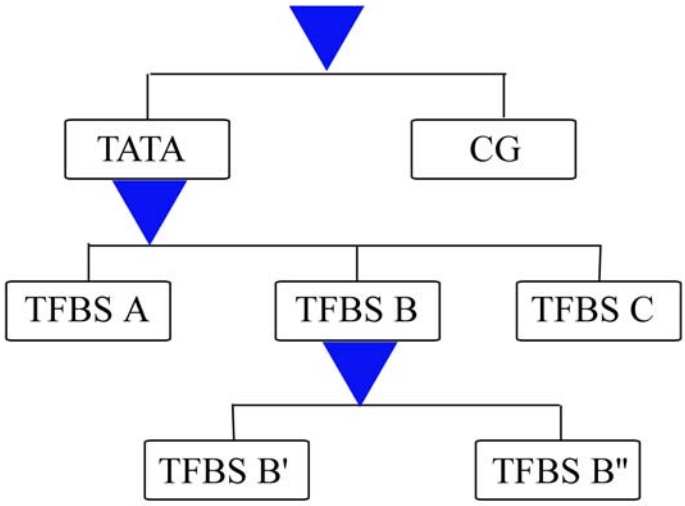

Figure 24. Schematic representation of hierarchical sequence model. Most promoter sequences have TATA box and CpG islands. Each promoter element can have specific DNA binding site for transcription factors to dock (represented as TFBS). Variations in each TFBS can in turn be modeled giving a hierarchical view of classifying different promoter types.

(iii) Dinucleotide Sequence Model – This classifier model capture features using the same strategy as that used in EAS except that the basis functions can also be derived based on dinucleotide compositions. The model found similar signals as those identified earlier by the anchored model.

(iv) Stem-Loop Model – As noted earlier, there is a possibility of there being secondary structure in the downstream region where RNA polymerase is likely to terminate and the algorithms have difficulty in capturing them as they are likely to vary its position for different genes. To address this I exploited the Eponine positional variance constraint by implementing the Nussinov algorithm with the Eponine trainer to detect secondary

structures. However, training on the chromosome 22 datasets with this combination turns out to be very computationally expensive. However, the combination worked for histone genes where a stem-loop structure at the end of the gene is known to terminate transcription. I collected 100 histone stem-loop structures from Rfam database (Griffiths-Jones *et al.*, 2003) and trained it against random sequences. The Eponine trainer with the use of Nussinov algorithm basis function successfully picked a constraint that distinguished histone 3'-ends from other sequences.

Likewise, in the earlier analysis, I tested the model with a toy dataset of 100 sequences of 180 bases each. Each sequence is designed by using a known sequence of 60 bases with potential to form stem-loop structure, flanked with random sequences extracted from chromosome 20. This set of sequence is trained against a random sequence dataset to learn stem-loop constraints that can classify both the data. The trainer picked 3 constraints – 1 positive flanked by 2 negative constraints on training. The positive constraint is due to the higher base pair metrics score for the known sequence (than the random sequence) in the positive dataset. However the base-pair metrics score flanking the known sequence is lower than the random sequence and hence these differences between positive and negative sequences were captured as two negative constraints by the trainer.

Although the stem-loop model was successful in picking secondary structure constraints that were able to distinguish positive and negative sequences in the two cases described, the method was found to be computationally expensive for training on transcription termination datasets. Scanning 2000 bases downstream of cleavage site to identify a secondary structure constraint that can classify it from random sequences was found to be computationally expensive when implemented using Nussinov algorithm. Future work to improve the algorithm or adapt a different algorithm for scanning might prove to be useful in constructing higher order models.

3.6 Performance of the model

To test the performance of the model, it is necessary to define true positives and false positives. Any predictions with higher score value among the predictions made on the transcription unit and lying within 2500 bases from the cleavage site in the same strand as the gene are considered to be *True Positives (TP)*. Any predictions within the transcription

unit in the same strand as the gene but not within 2500 bases from the cleavage site are considered as *False Positives (FP)*. Internal predictions on the reverse strand, intergenic predictions and predictions within 2500 bases of the cleavage site but in the reverse strand are ignored (Figure 25).

With these definitions, I tested the EAS model on chromosome 20 annotations and Figure 26 shows the performance as a Receiver Operating Characteristics (ROC) curve (ROC-Curve). ROC curve was plotted using the values given in Table 4.

Table 4. Coverage and accuracy values of transcription termination model along chromosome 20. ROC curve was constructed using these values.

Threshold	Coverage (%)	Coveargе values	Accuracy (%)	Number of True positives	Number of False positives
0.99	90.38	188/208	11.58	140	1069
0.992	87.01	181/208	14.43	127	753
0.994	79.80	166/208	15.34	95	524
0.996	71.15	148/208	16.50	67	339
0.998	36.53	76/208	16.91	23	113
0.999	12.5	26/208	18.18	6	27
0.9992	6.73	14/208	17.64	3	14
0.9994	4.32	9/208	10.0	1	9
0.9996	0.96	2/208	0.0	0	2

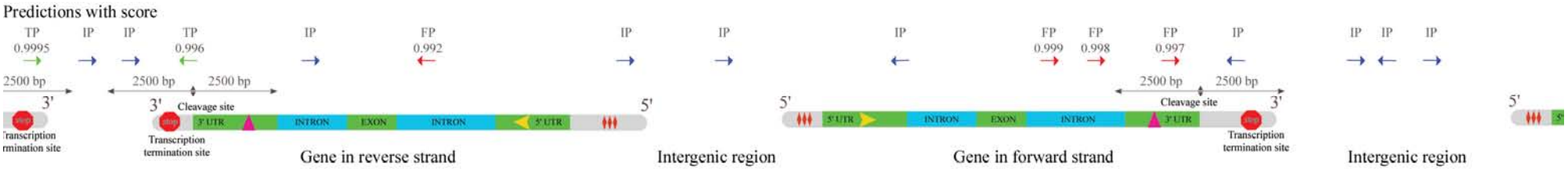

Figure 25. Schematic diagram showing criteria used for determining True positives (TP), False Positives (FP) and Ignored Predictions (IP).

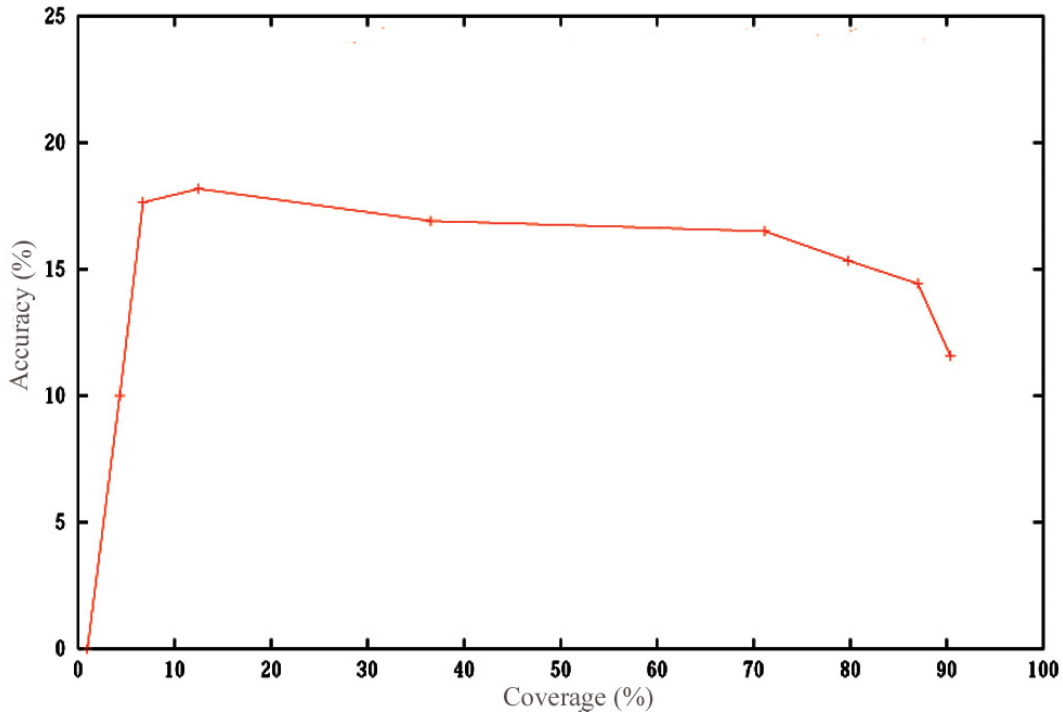

Figure 26. ROC curve on transcription termination sites in chromosome 20 for Eponine model.

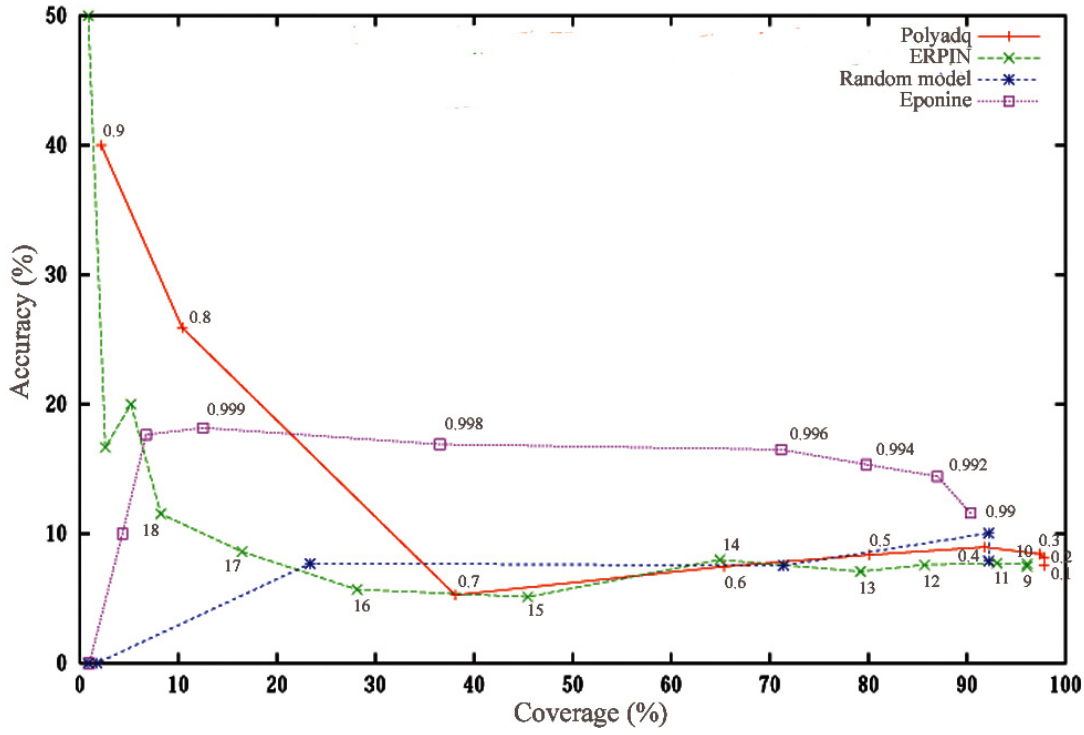

Figure 27. ROC curves on transcription termination sites for Polyadq, ERPIN and Eponine in comparison with random model.

I compared the performance with two other existing programs (ERPIN and Polyadq) to predict 3'-processing signals and a random model (Figure 27). Before discussing the performance I will briefly explain how random predictions are made.

Chromosome 20 without contig gaps is roughly 59 Mbp in length and hence 59,000,000 pseudo-random numbers between 0 and 1 were generated using PERL *rand()* function. The number generated is the score for each nucleotide in the chromosome and if the value exceeded 0.99 (threshold used in EAS model), it was recorded as a random prediction and the corresponding position in the chromosome was dumped. The strand for the predictions are generated using another *rand()* function. This procedure resulted in 294485 forward strand and 295229 reverse strand random predictions being collected. All the predictions are dumped in GFF format and compared with annotations of chromosome 20.

Figure 27 shows the ROC curve for Eponine, ERPIN, Polyadq and the random model. Both Polyadq and to lesser extent ERPIN have the highest accuracy at low coverage, however their performance drops to random at about 20% and 40% respectively. The accuracy of Eponine predictions seems less correlated with score, with no high accuracy peak at low coverage, however Eponine predictions remain at twice the accuracy of random at about 70% coverage. For comparable level of coverage, Eponine makes approximately 41% and 38% less false positives compared to Polyadq and ERPIN respectively, since at this level of coverage it is the only algorithm still performing better than random.

To determine if the genes identified by ERPIN, Polyadq and Eponine are same or different; I took predictions of the models at cut-off values of 9 (default threshold), 0.1 (default threshold) and 0.99 (comparable Eponine threshold) respectively. The result showed most of the gene termination sites were identified by all the three programs (might due to high coverage) and Eponine has detected all the predictions of EPRIN and misses only 14 genes predicted by Polyadq.

Among the positional constraints learnt by the EAS model, the 3'-end processing signals (poly(A) signal and GT rich motif) play an important role as can be seen from the constraint weight in Table 3. This is consistent with the results from experiments where upon deletion of the poly(A) signal the elongating polymerase failed to terminate and caused a run-over

(Edwalds-Gilbert *et al.,* 1993; Yeung *et al.,* 1998). To investigate the poly(A) signal requirement, I made a few models without DNA sequences spanning the 3'-end processing signals. Comparison of models developed from DNA sequences spanning 100 to 2000, 300 to 2000, 500 to 2000, 1000 to 2000 and 300 to 3000 bases from the cleavage site showed all the models performed less well than the models with 3'-processing signals. Thus the poly(A) signal appears to be a significant constraint in the model and required to make valid predictions along the chromosome. However the PE found by the model should not be underestimated as they are found to improve prediction accuracy.

3.7 Positional accuracy of the model

The density of predictions along the chromosome with respect to the annotated cleavage sites is shown in the Figure 28. As it can be noted, most of the predictions are associated with the annotated sites. Apart from the huge peak, there are predictions on either side of the peak with a distribution equal to the background prediction density. The model is good in detecting the directionality of the transcription termination site and the figure shows only the predictions matching the same strand as that of the annotation. Likewise, positional accuracy of ERPIN and Polyadq are shown in Figure 29. These methods are also good in predicting the transcription termination sites accurately.

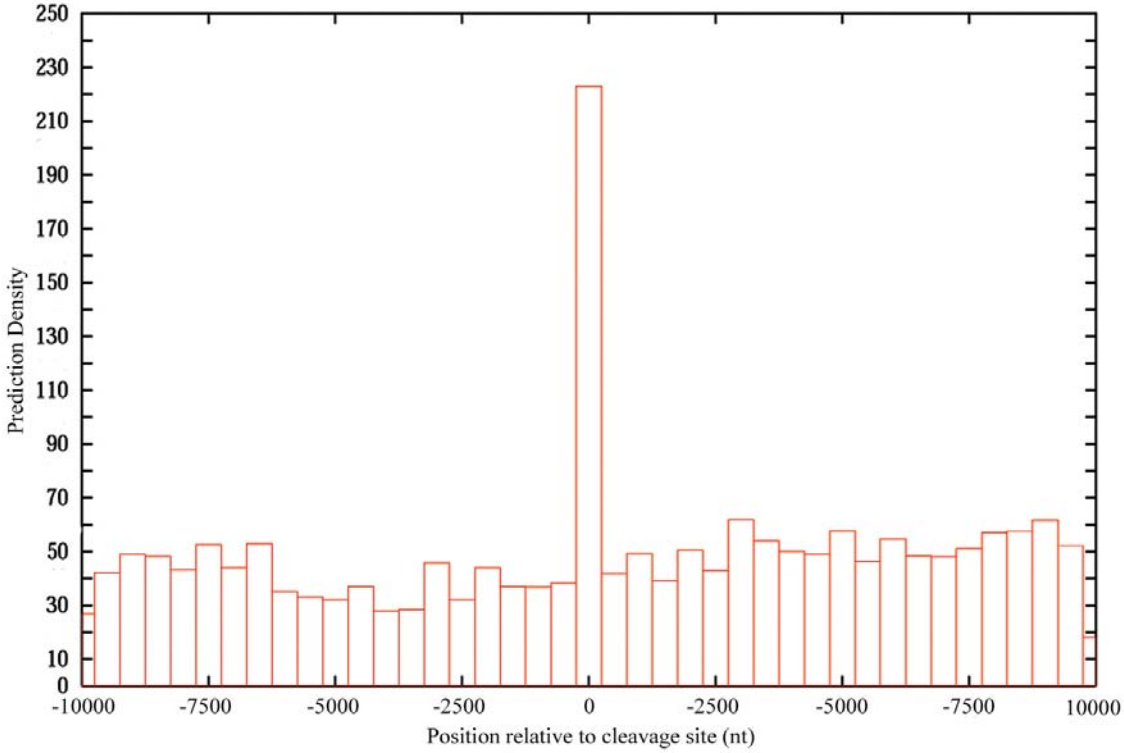

Figure 28. Prediction density for transcription termination model along chromosome 20.

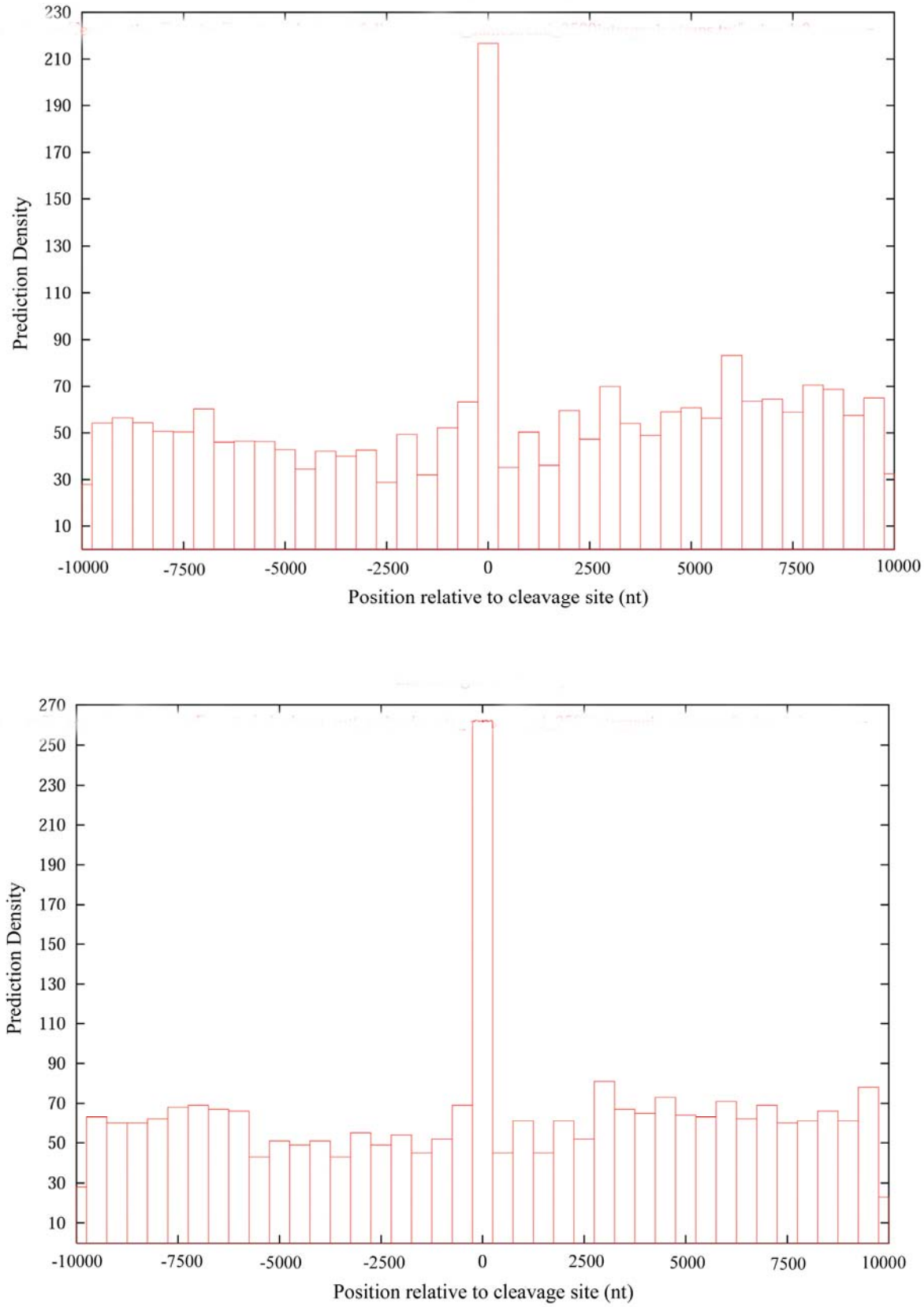

Figure 29. Prediction density along chromosome 20 for (a) ERPIN and (b) Polyadq.

3.8 Internal predictions

I hereafter concentrate on the distribution of predictions made within the gene as termination-like signals lying within the transcription unit are likely to challenge the transcription machinery leading to premature termination. Understanding the features of these predictions will help to differentiate them from correct gene ends and increase the accuracy of the model. Nearly 10% of total predictions along the chromosome are found within the gene in the same strand. To investigate whether the internal predictions made with Eponine model are linked with its learning protocol, an independent program, ERPIN was used for comparison. To facilitate this comparison, values are reported as predictions per 100 kb of genome sequence, subdivided into gene feature categories (Table 5).

Table 5. Distribution of false positives within transcripts, exons and introns.

False Positives per 100 kb	***Eponine***	***Erpin***
In Transcripts	19.86	32.00
Exons	3.66	5.41
Introns	20.76	33.48
Single Exons	0.00	0.00
First Exons	1.19	1.19
Internal Exons	6.11	8.55
Last Exons	1.60	3.72
Single Introns	16.54	30.38
First Introns	27.88	40.58
Internal Introns	21.35	35.84
Last Introns	9.69	11.77

The predictions per 100 kb of gene are 19.86 and 32.00 for Eponine (threshold: 0.99) and ERPIN (default parameters) respectively indicating the good performance of Eponine model assuming most of these predictions represent true false positives. The predictions were then subdivided into those present in exons and introns of the gene. The predictions per 100 kb of intron were found significantly higher than in the exons. The reason for this huge bias towards introns is unknown. On further classifying the distribution of predictions in exons between single (gene with just one exon), first, internal and last exons, showed internal exons have more predictions. However, since the number of predictions per 100 kb of exons is low, deriving conclusions from these small figures holds no importance.

In the same way, distribution of predictions found in introns, were classified between single (genes with just one intron), first, internal and last introns. The numbers indicate that first introns have significantly more predictions compared to internal and last introns. This holds true for ERPIN program predictions as well. The bias towards first introns is puzzling and so far there is no experimental evidence to explain the phenomenon. The table shows last introns have dramatically lower prediction rate as I excluded any predictions within 2500 bases from the cleavage site for this analysis. The average ERPIN predictions per 100 kb of single and first intron was calculated to be 39.26 and this value is equal to the 39 false positives per 100 kb specificity reported by the authors of the program earlier (Gautheret and Lambert, 2001).

Thus almost all internal predictions found within gene are present in introns and first introns have more predictions than internal and single introns.

Interesting experimental results were found in a recent chromatin immuno-precipitation assay by Affymetrix using high-density oligonucleotide arrays representing all nonreptitive sequences on human chromosomes 21 and 22 for Transcription Factor Binding Sites (TFBS) (Cawley *et al.*, 2004). The assay was designed to identify TFBS for Sp1, cMyc and p53 factors and found a minimal of 12,000 sites for Sp1, 25000 sites for cMyc and 1600 sites for p53 in the human genome. Only 22% of these predictions are found near 5’ termini of the gene while 36% lie within or near 3’ to well characterized genes and remaining 24% in intergenic regions. The TFBS not linked to 5’ termini of the gene are found correlated with noncoding RNAs. A significant proportion of these RNAs are co-regulated with

protein coding genes and activated by retinoic acid. The unexpected number of TFBS with just 3 transcription factors observed under one environmental induction condition suggests that there may be a large number of transcription units still to be identified in the genome. As the novel transcripts identified are found to be regulated by the same mechanism as protein-coding genes, they are expected to have similar transcription termination sites indicating that some of the excess predictions made by EAS model along the chromosome might be biologically significant.

The Sp1 transcription factor is known to bind G-rich elements (resembling PE) and is able to pause the elongating polymerase (Yonaha and Proudfoot, 1999). Identification of Sp1 TFBS within genes supports the idea that some of the internal predictions may be functional pause sites (Cawley *et al.,* 2004). Likewise, the PE in the EAS model resemble the MAZ (Myc-Associated Zinc Finger Protein) binding site (GGGGAGGGGAC) and MAZ sites have been shown to pause polymerase better than Sp1 sites. Also, MAZ protein has been found to be necessary but not sufficient for efficient 3' end formation (Yonaha and Proudfoot, 1999).

All of the experiments described above support the idea that some of the internal predictions made by the EAS model may not be false positive predictions but may be functional *in vivo*.

The EAS model has poly(A) signal as its major constraint and an internal prediction means similar signals are present within the transcription unit. A SELEX experiment to determine the branch point sequence from HeLa cell nuclear extract yielded a sequence motif AAUAAAG, that proved to be functional both as polyadenylation and branch site in a competitive manner (Lund *et al.,* 2000). Earlier experiments have also shown the competition between spliceosome and polyadenylation factors while the RNA polymerase is elongating the gene (Takagaki and Manley, 1998; Takagaki *et al.,* 1996). The complexes compete for the branch point signal in the acceptor site and depending on the local concentration of factors and strength of the signal either splicing or polyadenylation occurs. Thus the conserved branch point signal and its associated poly(T) tract might mimic the poly(A) signal and the poly(T) tract of the 3'-processing signals. I suspected that this could have caused the EAS model to make internal predictions. However I found no significant increase in the density of predictions near branch point signals.

I also suspect that in at least in a few cases, the internal predictions made by the model are not really false positives and may instead act as terminator signals for alternative transcripts of the same gene. As the number of alternative transcripts is difficult to quantify, even for well annotated chromosomes, further analysis will be needed to clarify this.

3.9 GO correlation

Initial observations of lists of genes having internal predictions showed they are enriched in a subset of genes. To investigate the nature of genes having high number of internal predictions, I used the Gene Ontology (GO) database (Gene Ontology Consortium, 2004). From the annotations of chromosome 20, 176 genes were mapped to a GO identifier and 45 of these genes have 10 or more internal predictions so I used GO to find their biological role. The analyses showed, 32 out of 45 genes have *Cell growth and or maintenance* function corresponding to GO identifier, *GO: 0008151*. Thirty two (32/45) is higher than the random expectation of 24 from 91 genes with the same GO id in the 176 genes dataset. There were much smaller differences in the use of more specific GO terms, and hence no functional annotation of genes with large number of internal predictions could be determined.

Figure 30 shows the wide variation in the density of internal predictions with an average of 10-18 internal predictions per 100 kb of transcription unit. This number is less than the prediction made by a comparable program, ERPIN. I did another GO ontology search for a set of 27 genes that have 20 or more internal predictions per 100 kb but found no common functional annotation. Unsurprisingly, the number of internal predictions was found to be correlated with transcript and intron length. Figure 31 shows transcript length and internal prediction rate has a linear correlation. However Figure 32 shows shorter introns of less than 1000 bases have high propensity to have more internal predictions. In introns of less than 1 kb, an average 18 internal predictions per 10 kb is present. This is significantly higher than 2.5 internal predictions per 10 kb of large introns. The reason for such bias is not known although manual investigation of some predictions suggested that they might be terminators of alternative transcripts.

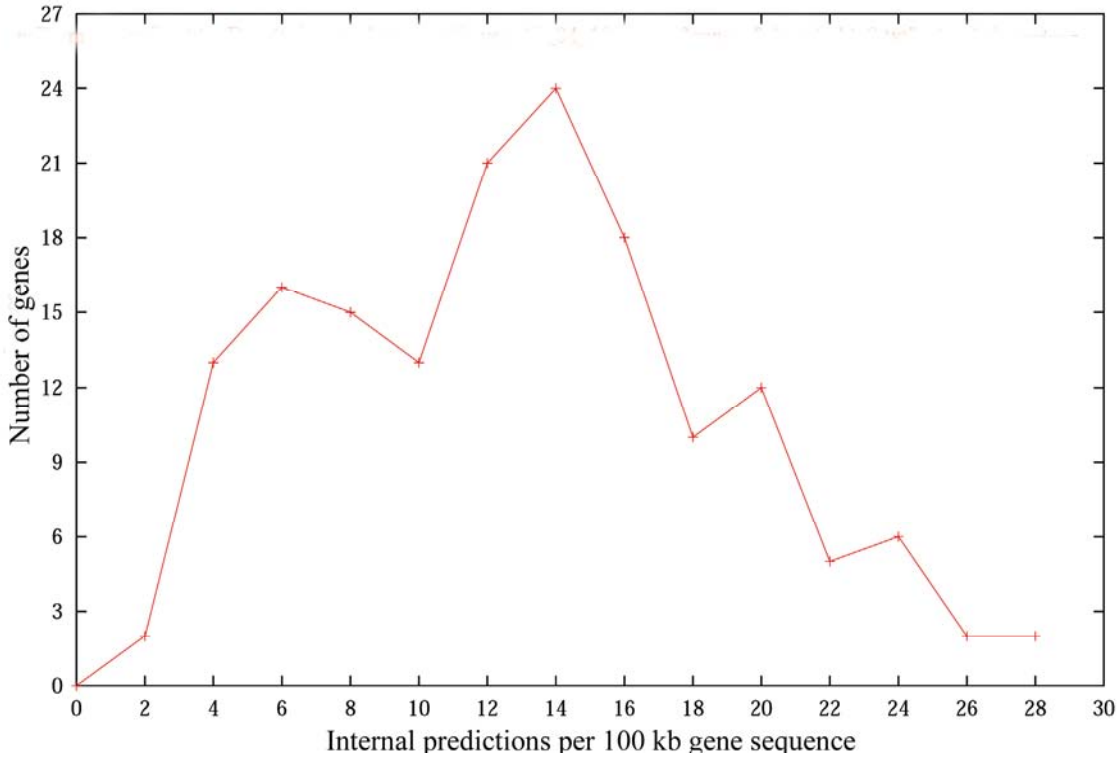

Figure 30. Internal predictions per 100 kb of gene sequence in chromosome 20 for Eponine model.

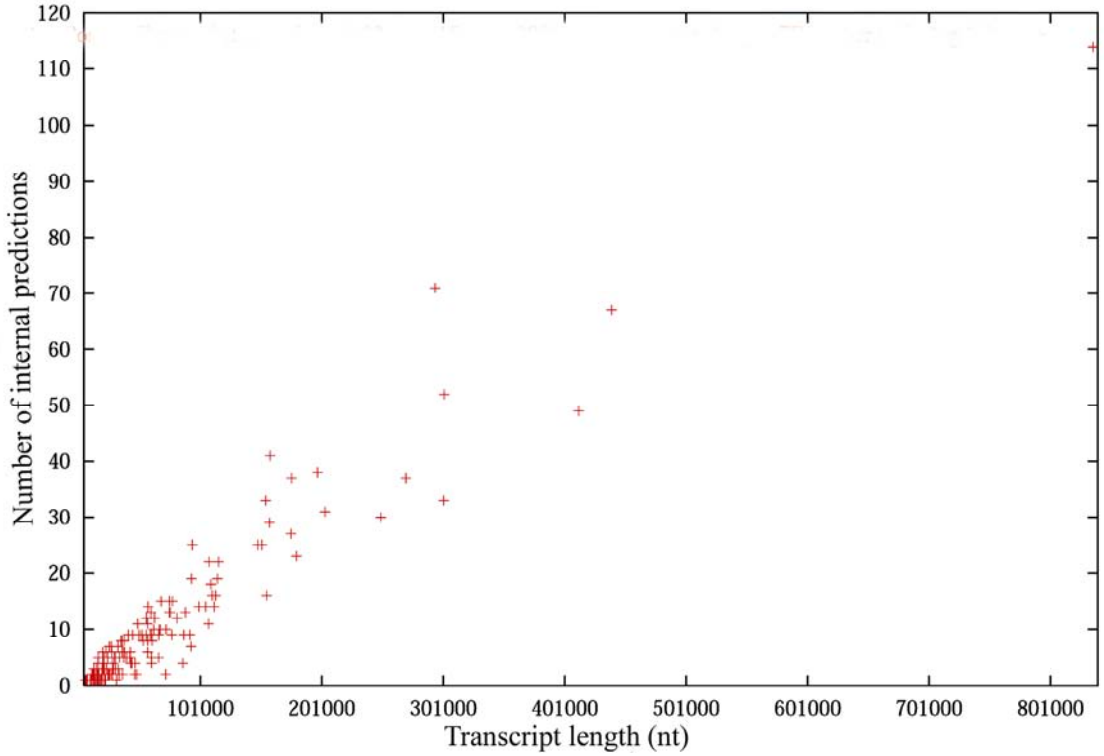

Figure 31. Internal predictions per transcript in chromosome 20 for Eponine model.

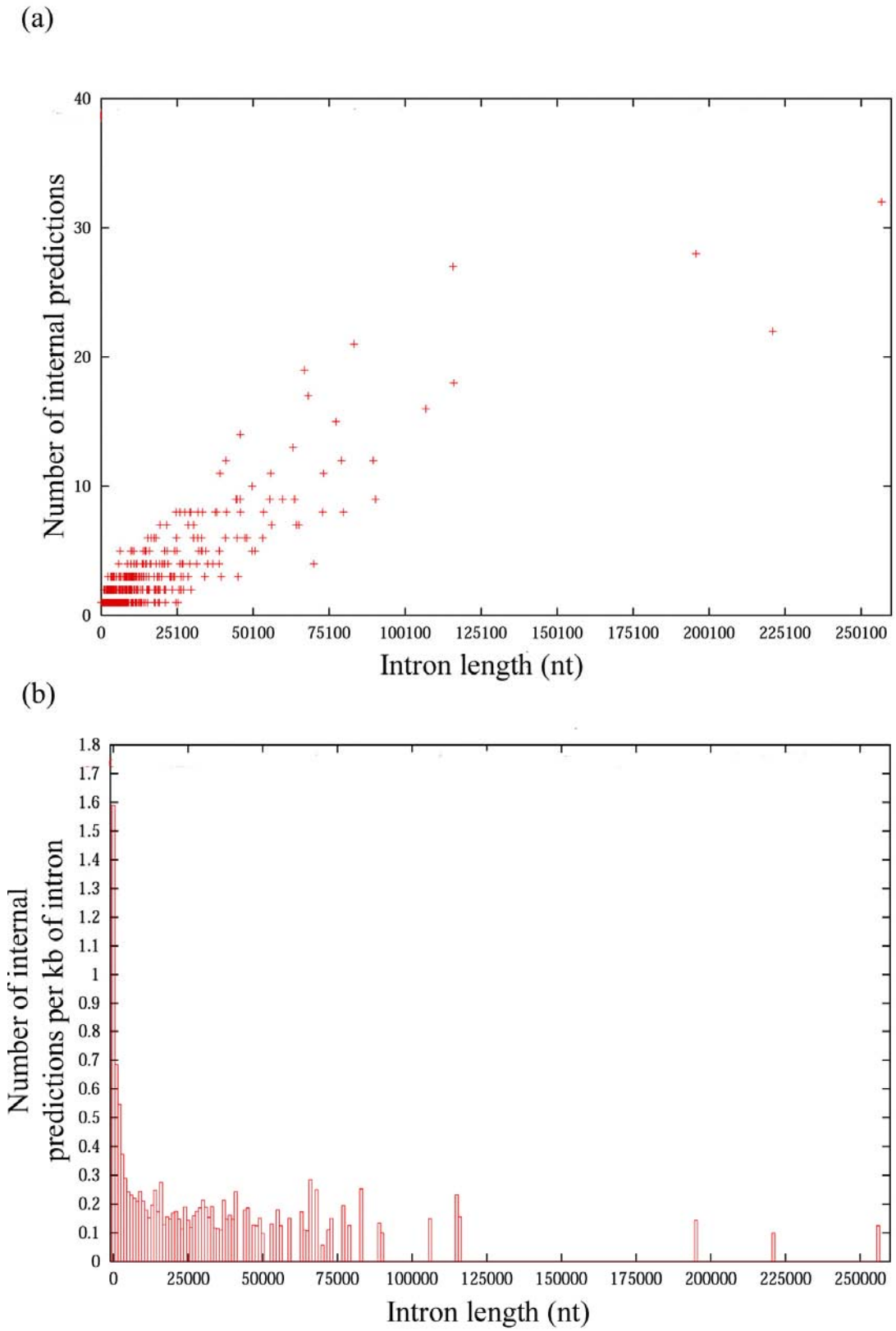

Figure 32. Internal predictions of Eponine model in introns of chromosome 20 (a) Number of internal predictions versus intron length (b) Number of internal predictions normalised over intron length.

3.10 Predictions near annotated gene start sites

Analysis of the density of predictions with respect to different gene features shows an unexpected concentration near the start of genes. The predictions found near the annotated gene start site are found to be close to its promoter elements and are unlikely to be a transcription termination site of the preceding gene. This was confirmed by considering only genes with no known annotated genes in the 2500 bases upstream of the start site and a minimum transcription unit length of 2500 bases. With this criterion, I selected 399 genes from the 584 annotated genes in chromosome 20, and 639 genes from the 1003 annotated genes of chromosome 6. The position density graphs shown here are calculated for these subsets of annotated genes. Figure 33 shows the density of predictions near annotated gene start sites of chromosome 20 in the same and opposite strand of the gene. A significant number of predictions are found just upstream (within 500 bases) of gene start site in the same strand orientation as the gene and the density is generally higher upstream than within the transcription unit (Figure 33a). Density predictions in the opposite strand show a significantly reduced number of predictions at the start site (position 0) and no significant increase in predictions elsewhere (Figure 33b). These results were generally consistent across different trained Eponine models, however predictions of two other transcription termination models also showed a number of predictions in the opposite strand just downstream of the annotated start site (Figure 34). Similar results were obtained on a different independent dataset of chromosome 6 (Figure 35a and Figure 35b). Also to investigate whether the predictions near gene start site are specific to models derived using the Eponine package, or more generally predicted, results form the two independent programs, Polyadq and ERPIN were similarly analysed for chromosome 20 sequences (Figure 35c; Figure 35d; Figure 35e; Figure 35f). Although there is no significant same strand prediction peak in the upstream region as seen for Eponine models, the density of predictions do appear to be relatively higher in the upstream region than in the transcription unit.

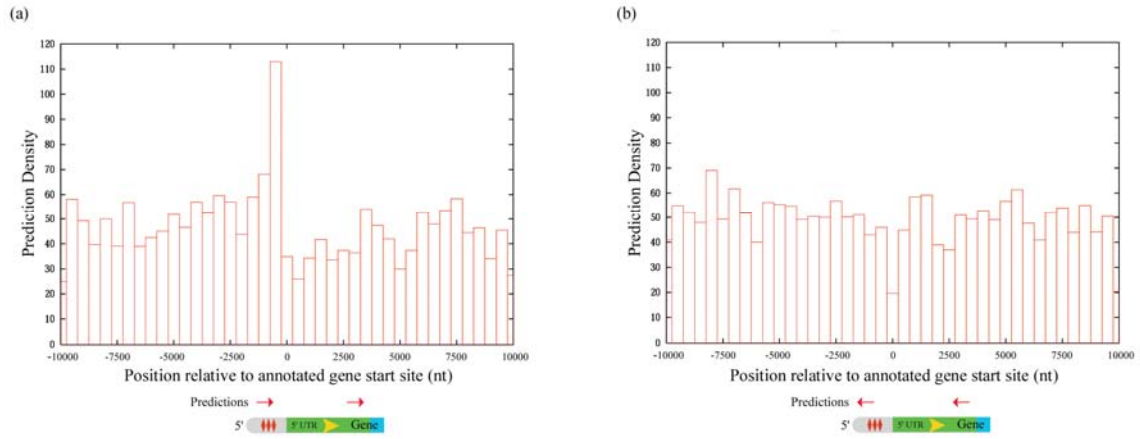

Figure 33. Prediction densities for transcription termination model near chromosome 20 annotated gene start sites. (a) Density of predictions in the same strand as of the gene (b) Density of predictions in the reverse strand as of the gene.

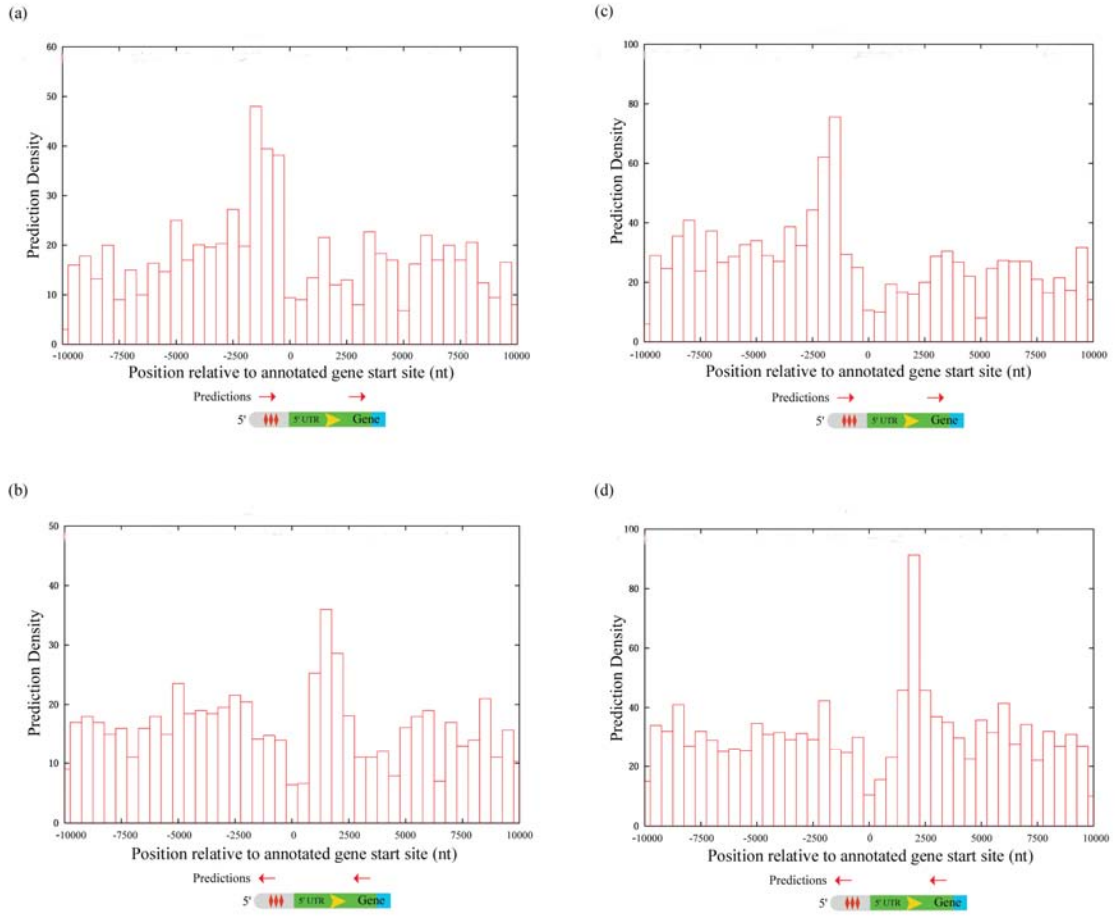

Figure 34. Prediction densities for two other transcription termination models near chromosome 20 annotated gene start sites (a), (c) Densities of predictions in the same strand as of the gene. (b), (d) Densities of predictions in the reverse strand as of the gene.

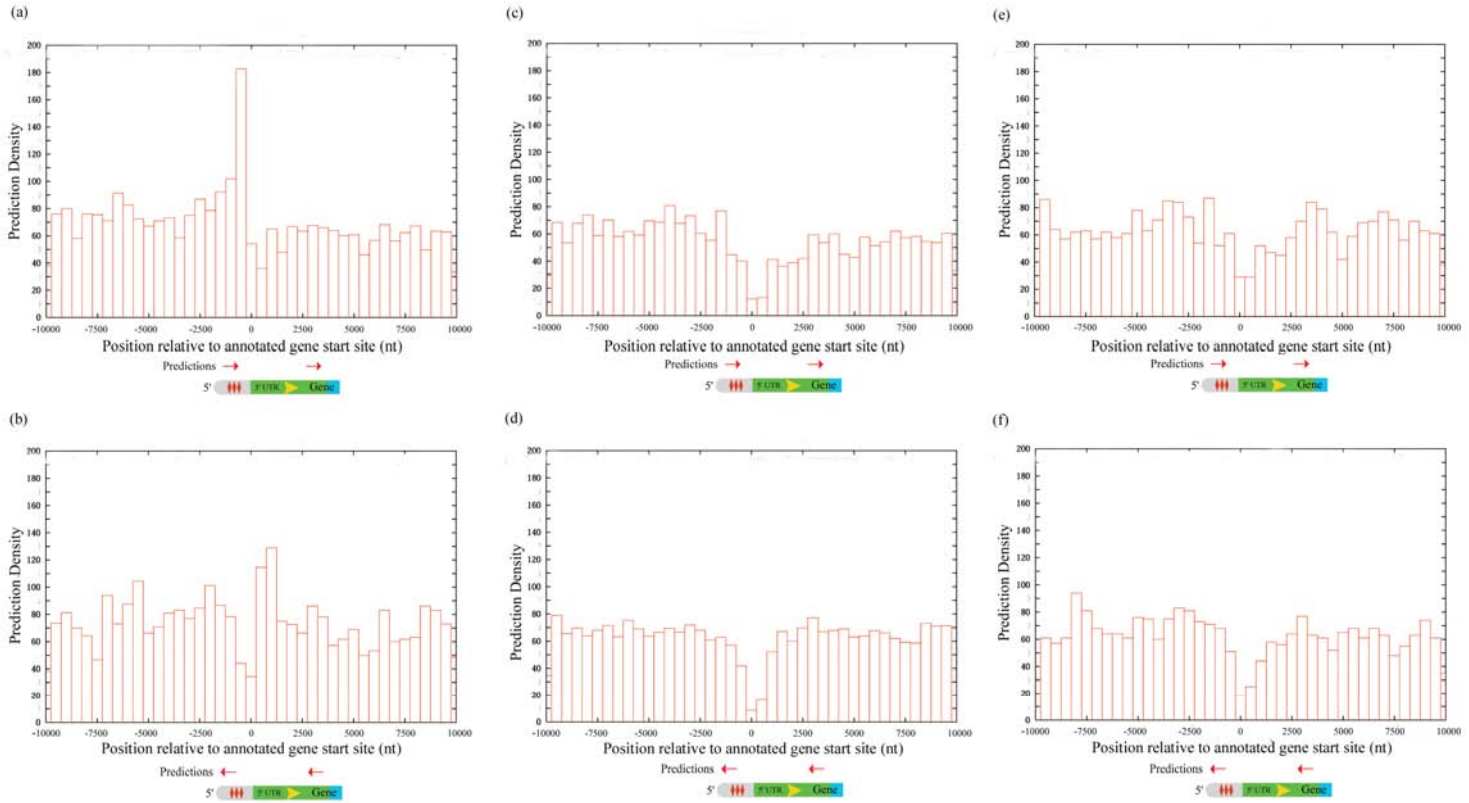

Figure 35. Prediction densities near chromosome 20 and 6 annotated gene start sites. (a) Density of predictions in the same strand as of the gene in chromosome 6 predicted by Eponine (b) Density of predictions in the reverse strand as of the gene in chromosome 6 predicted by Eponine (c) Density of predictions in the same strand as of the gene in chromosome 20 predicted by ERPIN (d) Density of predictions in the reverse strand as of the gene in chromosome 20 predicted by ERPIN (e) Density of predictions in the same strand as of the gene in chromosome 20 predicted by Polyadq (f) Density of predictions in the reverse strand as of the gene in chromosome 20 predicted by Polyadq.

Thus the plots show a higher than expected number of predictions just upstream of annotated gene start sites and on the assumption that this effect is real it is interesting to speculate on possible biological functions.

3.11 Hypotheses

Here, I have attempted to explain this unusual set of predictions by proposing a hypothesis assigning biological function based on previous knowledge from experiments.

Hypothesis I: As promoter detection by the RNA polymerase complex depends on a scanning mechanism, a terminator-like sequence positioned just upstream of a gene start site (referred hereafter as FP-TSS) might help the complex to prevent long regions of scanning and recruit the factors to the promoter elements and localize them.

The FP-TSS in the promoter region just 500 bases upstream of the annotated gene start site in the same strand as that of the gene might act as a guiding signal for the transcription factors to bind to its corresponding DNA binding domains. This will reduce the range of nucleotides need for scanning and help the factors to identify its binding site. If any case, if the factors gets recruited far upstream of the start site, then they are likely to get dissociated and the scanning terminated due to the FP-TSS signals. Thus the FP-TSS might help in positioning the initiation complex close to promoters.

Experimental evidences published recently support this view. Joseph Martens (Martens, 2003) reported a novel transcription interference assisted gene regulation process in the 'Mechanisms of Eukaryotic Transcription' conference held at CSHL, New York. In this case, a short transcript initiated from an upstream TATA box overlaps a downstream TATA site that is responsible for transcription of the downstream gene, SER3 in *Saccharomyces cerevisiae*. The upstream TATA site and transcript lies within the promoter elements of the SER3 gene and transcription of this short transcript is dependent on Snf2 chromatin-remodeling complex. The short transcript interferes with SER3 transcription by masking the overlapping activator elements of SER3. The interference was confirmed with the derepression of SER3 transcription when the upstream TATA site was mutated. Interestingly, apart from having a fully functional TATA box, the upstream short transcript

has a poly(A) signal just like a normal gene. The positional occurrence of this poly(A) signal correlates with the prediction identified earlier by the Eponine model near gene start sites. Thus this experiment confirms the poly(A)/terminator-like signals found upstream of the gene start site are likely to have biological functions.

Evidence for the hypothesis also stems from more unexpected discoveries of interaction between RNA polymerase II and processing machineries. As described in chapter 1, the interaction between polyadenylation machinery factors (CPSF, CstF) and CTD of RNA polymerase II has now been well established (Dichtl *et al.*, 2002b; Osheim *et al.*, 2002). Interestingly this interaction begins right from the promoter and apart from CTD, even general transcription factors like TFIID associate with CPSF (Dantonel *et al.*, 1997) making it a component of the transcription initiation complex. This study adds support to the presence of termination like sequences in the upstream region of gene start sites as predicted by the EAS model. Although the role of CPSF in the promoter region is not known, the presence of it in the initiation complex emphasizes that the FP-TSS are likely to have biological functions.

In a recent review, Calvo and Manley have discussed the interaction between 3' processing factors and initiation complex do not stop with CPSF but includes other factors like symplekin (Pta1 in yeast), Ssu72 and PC4 (Calvo and Manley, 2003). There is no definite answer for such growing presence of polyadenylation factors near to the promoter region although the conditions are explained by linking the factors to different roles when present in promoter and termination regions. However such strategies have created only confusing explanations. For example, Ssu72, a phosphatase that interacts with TFIIB and present in elongating polymerase (Sun and Hampsey, 1996) is expected to have anti-terminator activity. However in other experiments it was found Ssu72 is necessary for 3'-end formation and/or termination (Dichtl *et al.*, 2002a; Gavin *et al.*, 2002; He *et al.*, 2003). Similarly PC4, a co-activator protein that binds single and double stranded DNA displays anti-terminator activity (Aranda and Proudfoot, 2001; Ge and Roeder, 1994) but interacts with CPSF and CstF at the 3' processing signals (Calvo and Manley, 2001). Both PC4 and Ssu72, which are unrelated in primary structure, share a common function of helping transcriptional machinery to identify the gene start site by interacting with the general transcription factor, TFIIB (Sun and Hampsey, 1996; Woychik and Hampsey, 2002). Also the factors interact

with symplekin, a component of CPSF, mutually exclusive with PC4 binding at the 5' end of the gene and Ssu72 at the 3'-end (He *et al.,* 2003). However recent studies show Ssu72 possesses protein phosphatase activity and may be required for both ends for the gene (Ganem *et al.,* 2003; Meinhart *et al.*, 2003). Like other factors, symplekin also facilitates interaction of transcription initiation factors with CTD of polymerase at the time of transcription initiation (Rodriguez *et al.*, 2000). Thus the experiments show the functions of the protein molecules are conflicting and have different roles within the transcription machinery.

However the above studies confirm the unexpected discoveries of huge number of polyadenylation factors near promoter elements and that the poly(A) signal near to the promoter might be functional. This forms the basis for my hypothetical model that recruitment of polyadenylation factors at the promoter regions, might in turn, recruit other general transcription factors and thus help in localizing the initiation complex just upstream of gene start sites. Also, the predicted signals may help in avoiding the unnecessary scanning required to find the start site and take part in gene regulation as described earlier (Martens, 2003). A related puzzling question which remains to be tested is, if FP-TSS is functional, are they similar in activity to the 3'-processing signals? This can be verified by cloning the FP-TSS and surrounding sequences to the end of the gene and expect for termination of RNA polymerase II to occur. If polymerase terminates it will unambiguously confirm the upstream predictions are similar to 3'-processing signals found at the end of genes.

The hypothetical model also fits the transcription machinery model (Cook, 1999) wherein transcription factors and RNA polymerase associate together to form a machinery through which the DNA passes when a particular gene needs to be transcribed. Given the model is true, it is not a surprise to find so many polyadenylation factors in the promoter elements as all the factors are likely to participate in the machinery. Presence of FP-TSS at the promoter region in this machinery model will define the start of a gene by terminating any transcription starts initiated far upstream from promoter.

Hypothesis II: The unusual condition of significantly high number of predictions in the first intron compared to other introns is explained in this hypothesis. The polymerase queuing

model (Heinemann and Wagner, 1997; Wagner, 2000) explained in prokaryotes forms the basis for this. Prokaryotic genes are transcribed by multiple polymerases at any given time, and this leads to a trail of polymerase transcribing a single gene. If the rate of transcription is slow when compared to entry of new polymerase complexes at the promoter region, the transcription complexes are likely to queue all along the gene. I contemplate the same mechanism might act in eukaryotic gene transcription as well. In cases of genes with high expression levels and strong promoter motifs, there is a high probability of more transcription initiation complexes getting assembled and initiating transcription. However if the rate of transcription is likely to be less than the rate of assemblage and initiation then the complexes are likely to be queued. Having a terminator like sequences at the first intron might act as a strong pause signal and induce weak or incompetent complexes to terminate initiation and dissociate from the DNA. However if the complex is competent enough then its likely to continue the transcription process past the pause signals in the first intron and complete the whole gene transcription. Presence of predictions at the first intron compared to other internal introns might save energy from unnecessary transcription as the cell can abandon it just after initiation.

This hypothesis depends on the assumption that transcription of a gene is carried out by multiple polymerases at a given time. However there are split views on this. Supporters of the transcription machinery model argue gene transcription is done by a single polymerase at a given time. However there is no consensus so far. Even if the transcription machinery model holds true, the terminator like signals at the first intron might act as check point to evaluate the processivity of the machinery complex in transcribing the gene.

Experimental evidences support this hypothesis and reports premature termination occurs in the 5’ region of many viral and cellular genes (reviewed in Spencer and Groudine, 1990). Such intragenic termination occurs efficiently near gene start sites. This was shown in c-myc gene where terminator like sequence present 310 bases from the start site when moved to 600 bases resulted in more than five fold decrease in termination efficiency. This emphasizes that the evaluation of processivity of complexes occur early in the transcription process and influenced by the distance from start site (Roberts and Bentley, 1992). Similar observation in c-myc, c-myb, c-fos, β-globin, adenosine deaminase and porphobilinogen deaminase genes show that all intragenic terminations occur in the 5’ region of the genes

usually within 1 kb from the start site (Beaumont *et al.,* 1989; Bentley and Groudine, 1986; Chinsky *et al.,* 1989; Lois *et al.,* 1990; Mechti *et al.,* 1991; Watson, 1988). These experimental results show the predictions in the first intron might have biological function and help in classifying the transcription complexes that initiate from promoters into two heterogeneous sets based on their processivity. Thus the predictions might act as an attenuator and thereby allow only read-through complexes to complete transcription.

The read-through of RNA polymerase II can be assisted by various complexes. One such complex extensively studied in phage, called N and Q anti-termination system involves at least six proteins (Das, 1993; Friedman and Court, 1995; Greenblatt *et al.,* 1993). A similar mechanism is present in eukaryotes as well and so far 5 factors were reported. The first factor, S-II was originally discovered in mouse and found to suppress pausing of polymerase and activate reinitiation (Reines, 1994). The second factor, TFIIF (factor 5) is required for initiation and stimulation of elongation rate of polymerase (Wiest *et al.,* 1992). The third factor, TFIIX, identified in HeLa cell extract stimulates elongation of polymerase II (Bengal *et al.,* 1991). The fourth factor, a yeast protein YES stimulates the elongation rate of polymerase II (Chafin *et al.,* 1991). Finally, a factor, P-TEF stimulates elongation by forming part of productive elongation complexes and restores initiation of paused polymerase (Orphanides and Reinberg, 2002). DmS-II and Factor 5 forms part of the late elongation complex while P-TEF plays role in the early elongation complex (Marshall and Price, 1992). P-TEF phosphorylates DSIF and CTD of polymerase and thus increase the processivity of the elongation complexes (Renner *et al.,* 2001).

Thus the predictions near to the gene start site and first intron are likely to have biological functions and experimental evidence evaluating it will add new knowledge to the understanding of the transcription machinery.

3.12 Concluding remarks

With Eponine transcription termination models, I identified few multiplex pause elements present in the sequences downstream of the cleavage site. Occurrence of these signals repetitively indicates they might complement each other in pausing polymerase before release. The signals are similar to the sequences found in yeast ura4, α-globin, C2, factor B and nmt2 genes (Aranda and Proudfoot, 1999; Birse *et al.,* 1997; Yonaha and Proudfoot,

2000). The A-richness found in human β and α2 globin genes are represented as negative positional constraint (TTTT motif) in the model (Dye and Proudfoot, 2001; Enriquez-Harris *et al.,* 1991). Likewise the G-richness found in the pause elements (Table 1) agree with experimental results from human C2 and factor B genes (Ashfield *et al.,* 1991). MAZ and Sp1 transcription factors bind to these G-rich elements and interestingly in a recent experiment by Affymetrix (Cawley *et al.,* 2004), 34% Sp1 TFBS are found internal or proximal to the 3' end of the gene. These TFBS show a possible correlation with internal predictions identified by Eponine model. Detailed analyses of internal predictions indicate they are not randomly distributed and significantly present in longer genes and shorter introns (less than 1000 bp).

Earlier computational analyses by Nussinov (Nussinov, 1987, 1990) indicated the presence of TATAAA, AGGG and GGGC motifs in the sequences upstream of the transcription initiation site. These motifs resemble the AATAAA and pause elements of the model and thus correlate with significant number of predictions found proximal to the gene start sites.

Thus identification of transcription termination signals in the first intron and proximal to gene start sites encourages future mechanistic investigations and discussions concerning the transcriptional machinery and the possible reconsideration of current concepts of gene regulation in the eukaryotic genome.