MODELLING DONOR AND ACCEPTOR SITES

4.1 Introduction

Recent *in vivo* experiments show splicing or RNA processing events are found to be temporally and spatially related to the transcription process and that the CTD of the RNA polymerase II itself initiates such interactions (refer chapter 1, Cramer *et al.,* 2001; Manley, 2002). Hence while studying transcription termination, it is natural to consider splice sites and investigate whether modelling them is required to create an *ab initio* gene prediction system based on regulatory signals. Combining transcription termination models with splice site models may remove some of the internal predictions by the termination model and thereby help in predicting the correct gene structure.

In existing *ab initio* gene prediction systems splice site determination has formed a major role as they define the exons and introns of a gene. The first exon differs from internal exons as it lacks an acceptor site as the 5' end of the mRNA is capped with 7-methyl guanosine. Likewise, the last exon is different from other internal exons as it lacks a donor site as the mRNA is terminated with a poly(A) tail attached to the cleavage site. Each internal exon has an acceptor and donor site at its 5' and 3' ends.

These sites along with other regulatory elements recruit an array of protein and RNA factors depending on the splicing signals and remove the intervening sequences or introns from the nascent RNA and stitch the exons together. This process is referred as *Splicing* and details of this process are explained in chapter 1. Recognition of donor sites is relatively easy as the donor signals are more conserved than acceptor signals (for details refer reviews, Black, 2003; Jurica and Moore, 2003; Reed, 2000).

Several programs are available in the public domain that can detect donor and acceptor sites in the mRNA sequences. They function either as a stand-alone splice site finder or as part of gene prediction programs. The performance of most of the gene finding systems is greatly influenced by their accuracy at determining splice sites. In theory, a program that could identify all splice sites would do a nearly perfect job of *ab initio* gene finding as it would determine all protein coding regions correctly given the transcription start site (Brendel and

Kleffe, 1998; Burge and Karlin, 1997; Solovyev and Salamov, 1997). However identifying all the potential splice sites and gene structures is difficult, in particular because each gene may be spliced in a number of different ways. Recent experimental data suggests that in human at least one-third of genes are alternatively spliced (Ashurst and Collins, 2003). This increases the complexity of predicting donor and acceptor splice sites in the RNA sequences and the identification of gene structure by *ab initio* programs.

Thus in order to meet the objective of developing an *ab initio* gene prediction program based on regulatory signals, here I attempt to create a donor and acceptor site model using the EAS system explained in chapter 2.

4.2 Datasets

For the purpose of training and testing the model, I used annotated splice sites from human genomic sequences from the database, *SpliceDB* (Burset *et al.*, 2001). From this database, 28468 canonical and non-canonical human splice pairs were extracted. After removing splice site sequences with undetermined base pairs (denoted as 'N'), 24808 donor and 24894 acceptor splice sites were dumped to derive a positive dataset. From the 24808 donor sites, 500 sequences of 82 bases (40 bases on either direction of the consensus site + the 2 consensus bases at the donor site) each were extracted randomly to form a training set for donor sites. Likewise, 500 sequences of 82 bases (40 bases on either direction of the consensus site + the 2 consensus bases at the acceptor site) each picked randomly from 24894 acceptor sites formed the training set. GT and AG of donor and acceptor site respectively formed the anchor point and both the sets of sequences are collectively referred as *positive datasets*.

Eponine classifier requires another set of sequences where donor and acceptor signals are unlikely to occur. As this is the critical step in deriving the model, I tried different set of negative sequences – random, exonic and intronic sequences. A set of 946 random sequences of 82 bases each were dumped from chromosome 20 to form a negative set. Using BLASTN, an all-against-all search was done and sequences were removed such that none had more than 90% sequence identity with any of the others. This left a set of 561 sequences, which was referred as *randneg* negative dataset.

A list of 781 exons of at least 500 bases was dumped from chromosome 20 and 22 to derive the exonic negative dataset. After removing 52 redundant exons (with same identifiers), 82 base sequences from the centre of each exon were extracted. Then I did an all against all search on these 729 (781-52) sequences using BLASTN and again sequences were removed such that none had more than 90% sequence identity to any other. This formed a set of 500 sequences, *exonneg*. By extracting sequences from the middle of the exon, any sequence elements near exon-intron and intron-exon boundaries are excluded from the negative dataset.

Likewise, 1000 introns from chromosome 20 of at least 500 bases were used to form the intronic negative dataset. Redundant introns with same identifiers were removed leaving 891 introns. From this set, 82 base sequences from the centre of each intron were extracted. After removing any sequence with more than 90% sequence identity to any other as detected using BLASTN, a set of 507 sequences were dumped to form the *intronneg* dataset.

Another set of 500 sequences were created from randneg, exonneg and intronneg datasets by picking random sequences. This set was referred to as *combneg*.

4.3 Training the splice site models

With the availability of positive and negative datasets for training and testing, I used 900 sequences (450 positive + 450 negative) for training and the remaining 100 sequences (50 positive + 50 negative) for testing the models. The test sequences are unseen while training and used only for initial testing of the models. I trained an Eponine donor site model by allowing the trainer to run for 6000 cycles. Each cycle took a few seconds in a 256MB RAM PIII Pentium laptop. The anchor point for the training is fixed at the first base of the donor consensus sequence - *i.e.* G in GT consensus signal. The window size for training the model was restricted to 35 bases on either side of the anchor point as any constraints selected near to the edges of the sequence are likely to cause the trainer to trip, leading to problems in determining the Gaussian distribution for the constraint. As we are interested in capturing only the donor consensus signals rather than all regulatory elements conserved in exon or intron, the window size of 35 bases was found sufficient. During the training, models were dumped at various checkpoints to analyse the performance of the trainer and

determine the convergence of the model. Figure 36 shows a typical donor site model learnt by the EAS system. The model seems to be complex with positive and negative overlapping constraints. Different training cycles with modified parameters and negative datasets showed similar results and the model did not converge even at varied numbers of training cycles.

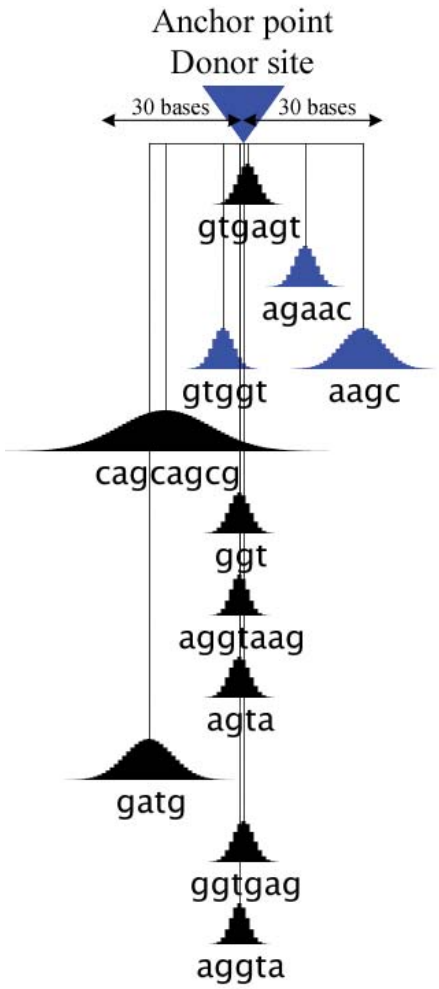

Figure 36. Donor site model trained from SpliceDB sequences

Likewise, the acceptor splice site model was also found to be complex with more negative constraints (Figure 37). Here, A of AG in the consensus acceptor site was used as the anchor point with a window size of 35 bases.

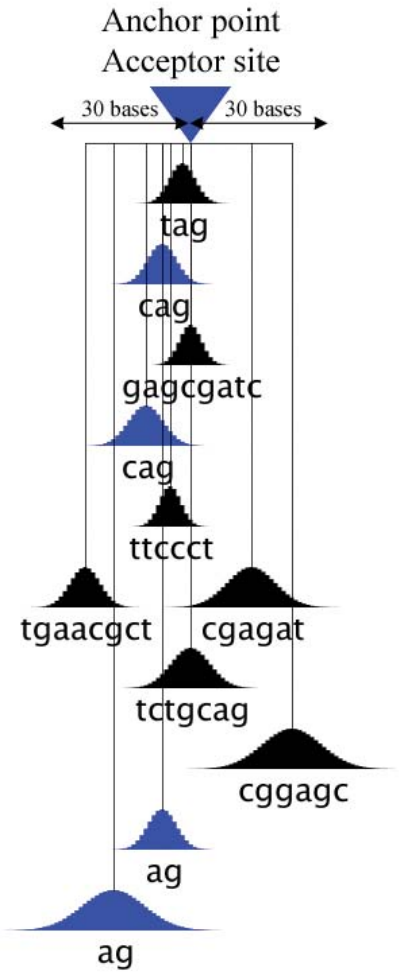

Figure 37. Acceptor site model trained from SpliceDB sequences

4.4 Refining the models

So to refine the models explained earlier, I adopted two strategies –

Eponine models are created by linking positional weight matrices scanned while training from the positive sequences. As explained in chapter 2, an initial set of constraints or weight matrices are sourced from the training set and on training, informative constraints are kept, removing the uninformative ones. This process continues until the datasets can be modelled using a set of sequence motifs. Thus in the EAS model, the complex natural data are simplified to a sparse model by projecting the data into feature space. However in some instances, selecting those few constraints that can effectively define the feature space of the

datasets might be difficult and hence the models are unlikely to converge. Here donor and acceptor site models have reached this point and hence to facilitate the learning process, I used weight matrices calculated from donor and acceptor sites of chromosome 22 as an input along with the DNA sequences. This reduced the difficulty in learning an appropriate set of constraints that can optimally classify positive from negative sequences. From chromosome 22, experimentally annotated 2348 donor and acceptor sites were dumped from coding genes and weight matrices showing the probability distribution of the nucleotides at each position of the sequence was constructed. Figure 38 shows the probability distribution for 30 nucleotides around donor and acceptor sites. The donor weight matrix has captured the canonical consensus sequences reported earlier. Likewise, the acceptor site matrix has captured the consensus sequence along with the polypyrimidine motif preceding the signal. These weight matrices are used as input along with DNA sequences to learn a sparse EAS model by including the following lines in the parameter file.

(a) Donor site weight matrix (b) Acceptor site weight matrix

Figure 38. Nucleotide Distribution at (a) Donor and (b) Acceptor site from chromosome 22 sequences

```
<child jclass="eponine.model.WMFileBasisSource">
    <string name="fileName" value="./donorORacceptorWMfile.xml" />
    <int name="position" value="-25" />
     <double name="minDistWidth" value="2.5" />
    <double name="maxDistWidth" value="50.0" />
    <boolean name="reversible" value="false" />
</child>
```

Positional distribution of the constraints learnt by Eponine is usually captured as a Gaussian distribution. However, the system allows for various other distributions to be used depending on the conditions of the problem. Here as the consensus sequence of the donor

and acceptor sites in the training sequences are less likely to vary in their position, a Delta distribution (instead of Gaussian distribution) is appropriate in modelling the positional variations. Thus for learning splice site models I implemented a Delta distribution along with the Gaussian for capturing the offset values of the constraints relative to the anchor point. I added the following lines to the parameter file –

```
<child jclass="eponine.model.MakeDeltaBasisSource" />
<child jclass="eponine.model.BreakDeltaBasisSource" />
```

With these changes I retrained the model with a new set of data derived from chromosome 22. From chromosome 22, 550 donor site sequences of 350 bases each (50 bases upstream and 300 downstream of GT signal. 300 bases downstream of GT is used in the aim to find out any unknown signals in this region) were dumped to form the positive dataset. Likewise, 550 acceptor site sequences of 350 bases each (300 bases upstream and 50 bases downstream of AG signal. 300 bases upstream of AG signal will include the well known Branch Point region) formed the acceptor site positive dataset. An equal number of random sequences from chromosome 22 formed the negative set. From the positive and negative datasets, 1000 sequences (500 positive + 500 negative) were used for training and the remaining 100 sequences (50 positive + 50 negative) for initial testing of the model. G of GT and A of AG in donor and acceptor signals respectively were used as anchor points. For donor site mode, the window size limits are set to 42 bases upstream and 290 bases downstream of the anchor point whereas for the acceptor model, the limits are 290 bases upstream and 40 bases downstream of the anchor point. Although the window size spans to 290 bases, the positional constraints learnt by both models are within 30 bases from anchor points, emphasising the fact that regulatory elements determining splice sites are closely linked with donor and acceptor signals.

Figure 39 and Figure 40 show the refined models with new parameters and datasets for donor and acceptor sites respectively. Figure 41 shows the position, constraint weight and Gaussian width of each constraint learnt by the donor model. The consensus signal at the donor sites are captured by 3 positive constraints. One of the constraints had a Delta distribution meaning no positional variation in the motif. All the constraints emphasize the importance of the GT bases in the consensus motif. From the intronic sequences, a

constraint rich in G nucleotides was learnt and positioned at 28 bases downstream of the anchor point. The biological importance of the motif is not known. Table 6 shows the occupancy value of the top 15 motifs represented in various donor site models.

Table 6. Occupancy value for motifs detected in the donor site models.

Number of models considered - 27	
Occupancy value for motifs below -10 bp	
Motifs	**Occupancy Value**
cgac	0.07
gccgc	0.07
ccg	0.07
gttaa	0.04
ttag	0.04
accg	0.04
taagtt	0.04
cgg	0.04
tgggt	0.04
taag	0.04
acga	0.04
ggctaccgc	0.04
tgaaact	0.04
gggt	0.04
cg	0.04
Occupancy value for motifs between -10 and 10 bp	
gt	0.41
ggt	0.33
aggt	0.26
gta	0.22
ggtaag	0.22
gtacg	0.19
ggtgagt	0.15
aggtaag	0.15
gtaagt	0.15
aggta	0.15
ggta	0.11
gtaag	0.11
gat	0.07
gtaagtc	0.07
ga	0.07
Occupancy value for motifs above 10 bp	
ataa	0.11
ggggtggg	0.07
aagc	0.04
gca	0.04
tggtagt	0.04
gggggg	0.04
gcgg	0.04
ctatatcaca	0.04
cgg	0.04
taa	0.04
ttgtgggt	0.04
tttg	0.04
gcg	0.04
accaa	0.04
tatacgg	0.04

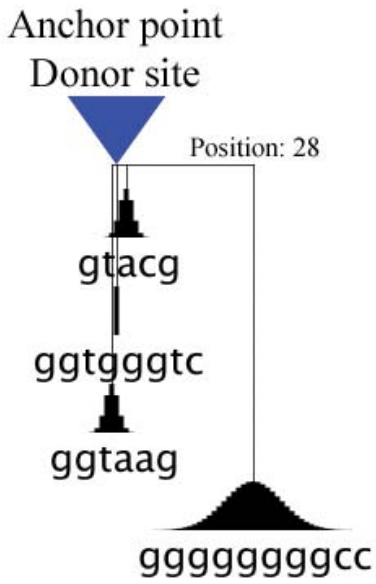

Figure 39. Donor site model trained from chromosome 22 sequences and donor site weight matrix

MOTIFS	POSITION	CONSTRAINT WEIGHT	GAUSSIAN WIDTH
	-1	9.54	1.33
	0	4.66	-
	2	2.81	1.33
	28	5.32	5.83

Figure 40. Position constraints of donor site model learnt while training chromosome 22 sequences

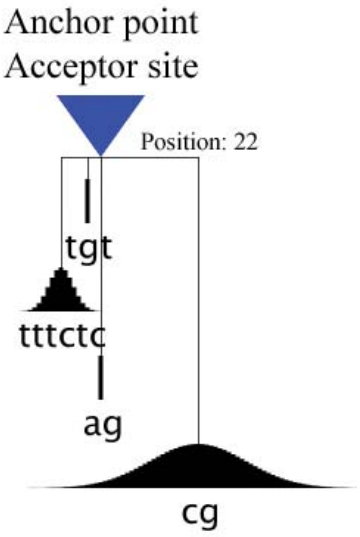

Figure 41. Acceptor site model trained from chromosome 22 sequences and acceptor site weight matrix

Figure 42 shows the properties of positional constraints learnt by the acceptor model. The consensus AG signal is well captured along with the polypyrimidine motif known earlier. A CG rich motif was found 22 bases downstream of the acceptor site (in exonic sequence) with a Gaussian distribution width of 10.96 and constraint weight of 9.42. The values emphasise the signal is important but the role of the motif is not known. The frequency of distribution of these motifs and other top 15 motifs in different regions of the model are shown in Table 7. A value of more than 1 indicates the motif is represented more than once in few models.

Thus, this second training approach, seeding the training with a position distribution and making use of a Delta function, was able to generate simple models containing only positive weights (Figure 39 and Figure 41) compared to the earlier training approach (Figure 36 and Figure 37).

Table 7. Occupancy value for motifs detected in the acceptor site models.

Number of models considered - 34	
Occupancy value for motifs below -10 bp	
Motifs	**Occupancy Value**
ag	0.44
agg	0.12
ctga	0.09
cag	0.09
ttttccttttttttttttccttccagg	0.09
tttccttttttttttttccttccaggt	0.09
ttagg	0.06
ttttccttttttttttttccttccaggt	0.06
gt	0.06
agctccttttttttttttccttccagg	0.06
ctgac	0.06
ccttttttttttttttttcaggtccaggt	0.06
agc	0.06
ttttttttttttttttctttccgggcag	0.06
gtggc	0.03
Occupancy value for motifs between -10 and 10 bp	
ag	1.03
cag	0.68
tag	0.26
agg	0.21
gg	0.18
tagg	0.15
ta	0.09
tgt	0.09
cagg	0.09
gt	0.06
aggcgggt	0.06
agaactc	0.06
cagct	0.06
ca	0.06
gcag	0.06
Occupancy value for motifs above 10 bp	
cg	0.15
cga	0.06
gca	0.06
gacgacc	0.03
cgcggaga	0.03
atgatga	0.03
tgctgc	0.03
ggta	0.03
cgggga	0.03
cggct	0.03
gaagttctgcagg	0.03
gcggaggagttc	0.03
gaacgcggaggagttc	0.03
gtta	0.03
gtctta	0.03

MOTIFS	POSITION	CONSTRAINT WEIGHT	GAUSSIAN WIDTH
	-3	3.34	-
	-9	7.99	2.71
	0	4.65	-
	22	9.42	10.96

Figure 42. Position constraints of acceptor site model learnt while training chromosome 22 sequences

4.5 Validating and testing the models

I tested the performance of these refined models of donor and acceptor sites using datasets derived from chromosome 20. From the VEGA annotation of chromosome 20 (on build NCBI 33) (Ashurst, 2002), I extracted 614 genes. These genes are defined as 'Known' or 'Novel_CDS' and 'Novel_transcript' in the database. 'Putative' and 'Pseudogene' categories are not considered. Constitutive and alternative exons known in these 614 genes totalled to 8771 and they were dumped to extract donor and acceptor sites. Out of 614, 42 genes are single exon genes and are thus omitted in this study. From the remaining 572 genes, 8037 and 8141 constitutive and alternative donor and acceptor sites are extracted respectively. After removing the redundancy present in the set of donor and acceptor sites, 5835 and 6166 unique donor and acceptor sites are found and I used this set to test the performance of the models.

Coverage is defined as the set of genes with at least one prediction within it. The value is calculated from the number of the genes with a prediction over the total number of genes (572). Accuracy is defined as the set of predictions that matches the annotated sites over total predictions within the gene. Any predictions (in the same strand matching the

annotation, prediction on the opposite strand are considered as false positives) present within 10 bases of the centre point of annotated sites are considered as true predictions. Any other predictions, within transcription unit or intragenic region, are considered as false positives.

Coverage was calculated at exon level as well and in this case, exons with predictions are counted as predicted.

Figure 43 and Figure 44 show the coverage and accuracy of the donor and acceptor site model respectively as a ROC curve. The donor site model registers higher accuracy at low coverage rate. Comparatively the performance of the acceptor site model is less than the donor site model. This is expected as the acceptor sites are less conserved and they are relatively more difficult to identify than donor sites.

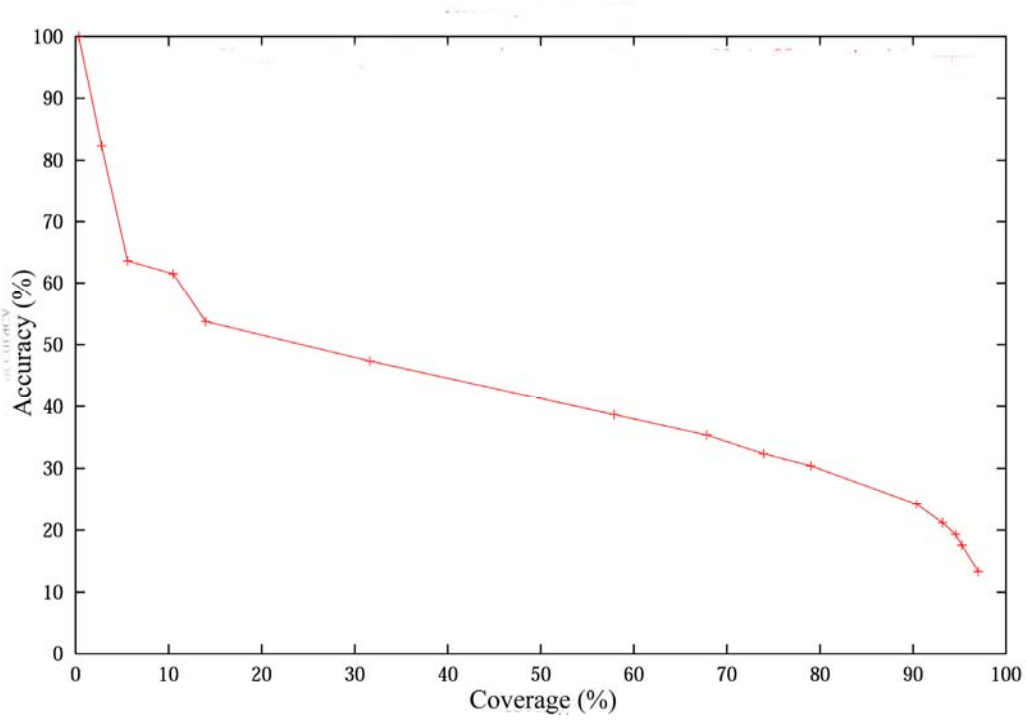

Figure 43. ROC curve for Eponine donor site model on chromosome 20 dataset

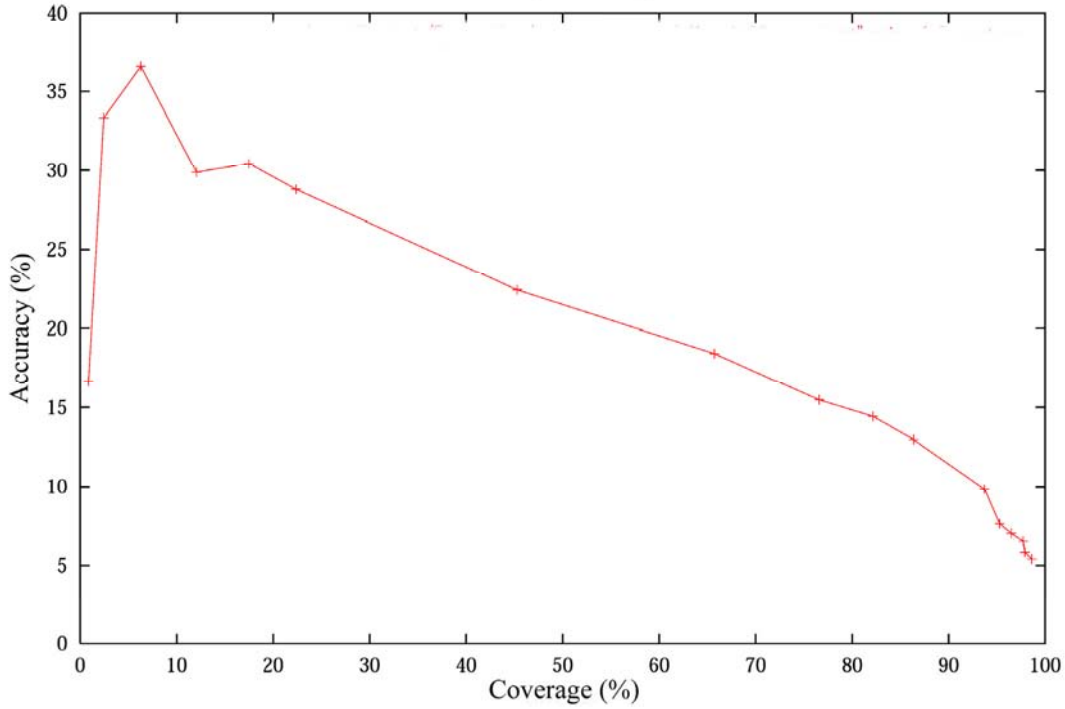

Figure 44. ROC curve for Eponine acceptor site model on chromosome 20 dataset

4.6 Position accuracy of the models

I calculated the densities of predictions of the donor and acceptor site model in relation to the annotated donor and acceptor sites in chromosome 20. The histograms of the densities calculated are shown in Figure 45 and Figure 46. The X-axis represents the position of the sequence relative to the annotated sites and the Y-axis represents the density of predictions at each position. In both figures the densities are drawn for 100 bases upstream and downstream of the annotated site.

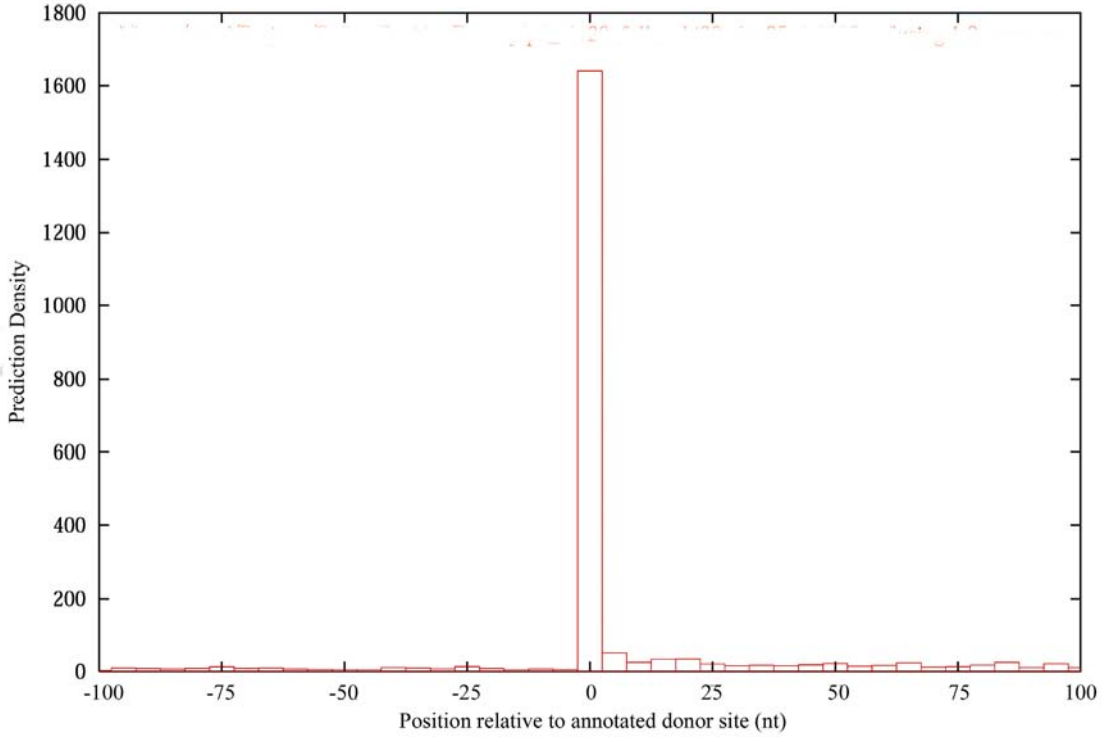

Figure 45. Prediction density for donor site model relative to annotated sites

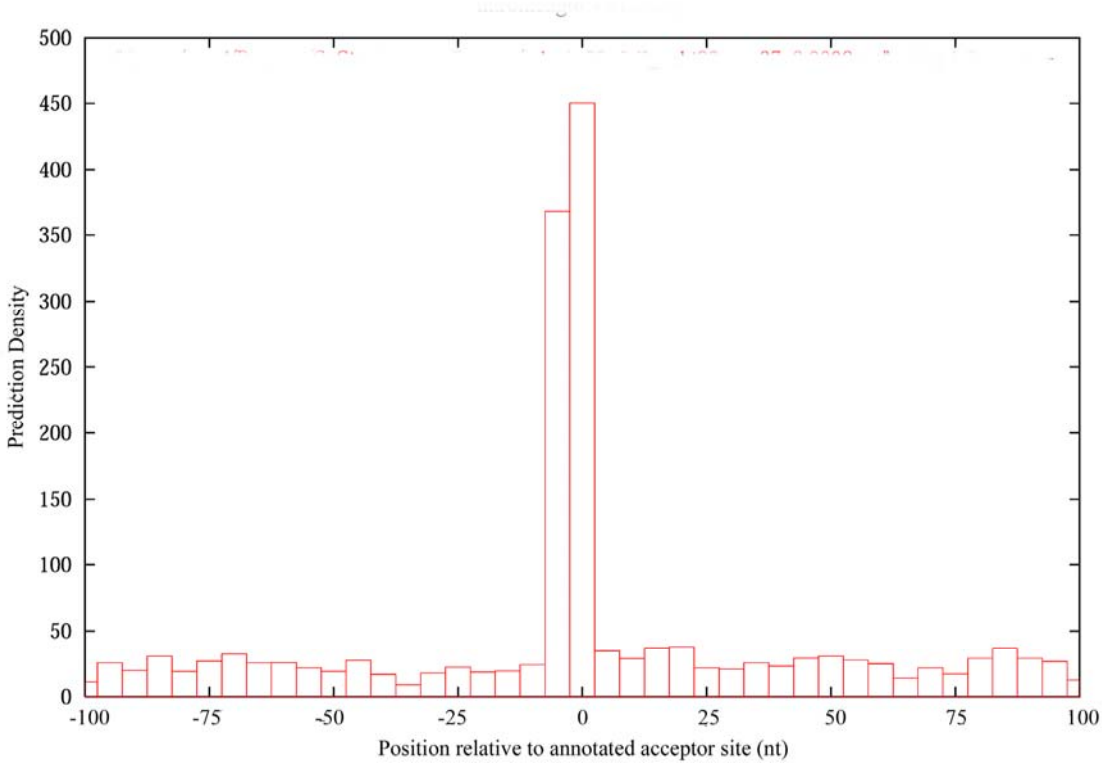

Figure 46. Prediction density for acceptor site model relative to annotated sites

The results show a clear peak exactly on the site of annotated donor and acceptor sites. The accuracy of the predictions by the donor site model corresponds to within 5 bases from the annotated site whereas acceptor site model predictions are within 10 bases. Figure 46 shows few predictions on either side of the peak: this might correspond to the false positives of the model but some of them might be due to alternative acceptor sites in the sequence.

Both the models are good at detecting the directionality of the sites and the figures shown here are for predictions in the same strand as the annotation. Density histograms of the predictions on the reverse strand show no peak at the annotated site.

4.7 Comparison with other models

I compared the Eponine splice site models with two other splice site programs available in the public domain. They are –

StrataSplice (Levine, 2001a) – This program uses a new splice site prediction model that combines the local GC content (80 bases upstream and downstream of the splice site) with a standard probabilistic pattern recognition technique. The method predicts both canonical (GT-AG) and minor variant (GC-AG) splice sites and is designed to integrate easily into a variety of gene prediction and annotation systems. The performance of the model is better in gene-rich high GC regions.

GeneSplicer (Pertea *et al.*, 2001) – This program uses a decision tree method called maximal dependence decomposition, first developed by Burge and Karlin (Burge and Karlin, 1997), enhanced with markov models that capture additional dependencies (16 bases around donor site and 29 bases around acceptor sites) surrounding the splice sites. This method considers only a small window around the splice junctions, which contains most of the information recognised by the spliceosome. It also takes into account the coding and non-coding sequence switch at the splice junctions and the local score optimality feature developed by Brendel and Kleffe. (Brendel and Kleffe, 1998).

I used the test set described earlier – 5835 donor and 6166 acceptor splice sites from 572 genes from chromosome 20 to compare the performance of the methods. As both programs

are available in the public domain (Levine, 2001b; Pertea, 2001), I downloaded them and scanned chromosome 20 sequence locally in a 1GB Compaq Tru64 UNIX machine. Donor sites are predicted with higher accuracy by StrataSplice than acceptor sites. I collected the StrataSplice predictions for donor sites at posterior probabilities: 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.12, 0.14, 0.18, 0.20, 0.24, 0.26, 0.30, 0.34, 0.38, 0.42, 0.44, 0.48, 0.50, 0.54 and 0.58 while for acceptor sites at 0.02, 0.03, 0.04, 0.05 and 0.06. No acceptor site predictions are made for probabilities above 0.06. Likewise donor and acceptor predictions of GeneSplicer are extracted at score thresholds: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22. Again donor site predictions by GeneSplicer had higher score values than acceptor site predictions.

Figure 47 and Figure 48 shows the performance of Eponine, StrataSplice and GeneSplicer on chromosome 20. Predictions within 10 bases from the annotated donor and acceptor site predictions are considered as true predictions. Coverage and accuracy are measured as described above. Figure 47 shows the performance of Eponine is comparable with StrataSplice although less than GeneSplicer. However coverage and accuracy of Eponine acceptor site model and StrataSplice are significantly less than GeneSplicer (Figure 48). To analyse the coverage of exons by the three programs, I did a ROC curve taking only exons having a prediction within 10 bases from donor or acceptor sites as true predictions while calculating coverage. 5835 exons from 572 transcripts are used for this analysis. Figure 49 and Figure 50 shows the exon coverage and accuracy for donor and acceptor sites respectively. GeneSplicer again performs better in detecting the exons boundaries better than Eponine and StrataSplice.

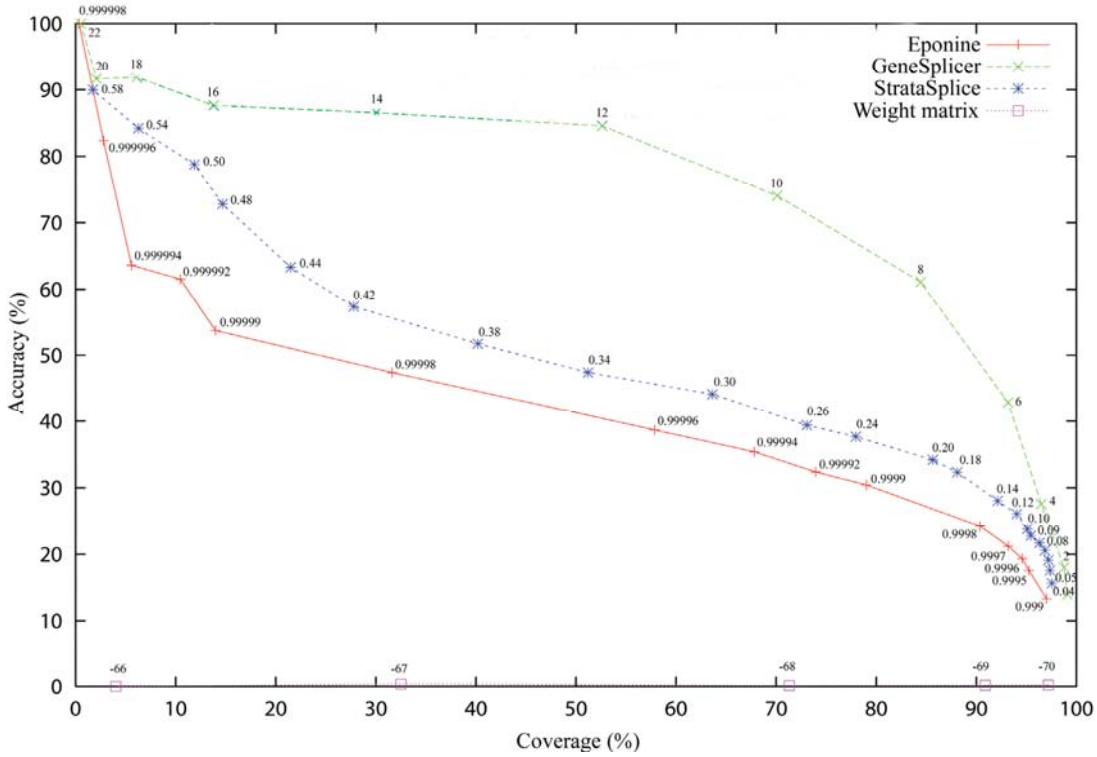

Figure 47. ROC curves on donor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

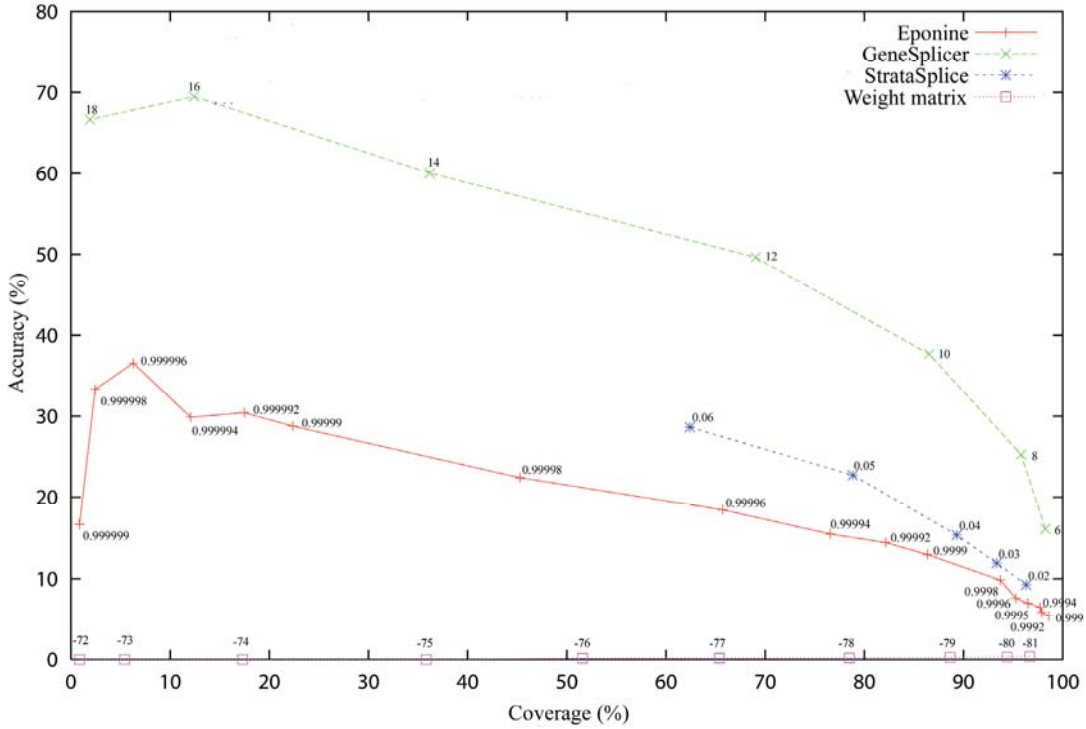

Figure 48. ROC curves on acceptor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

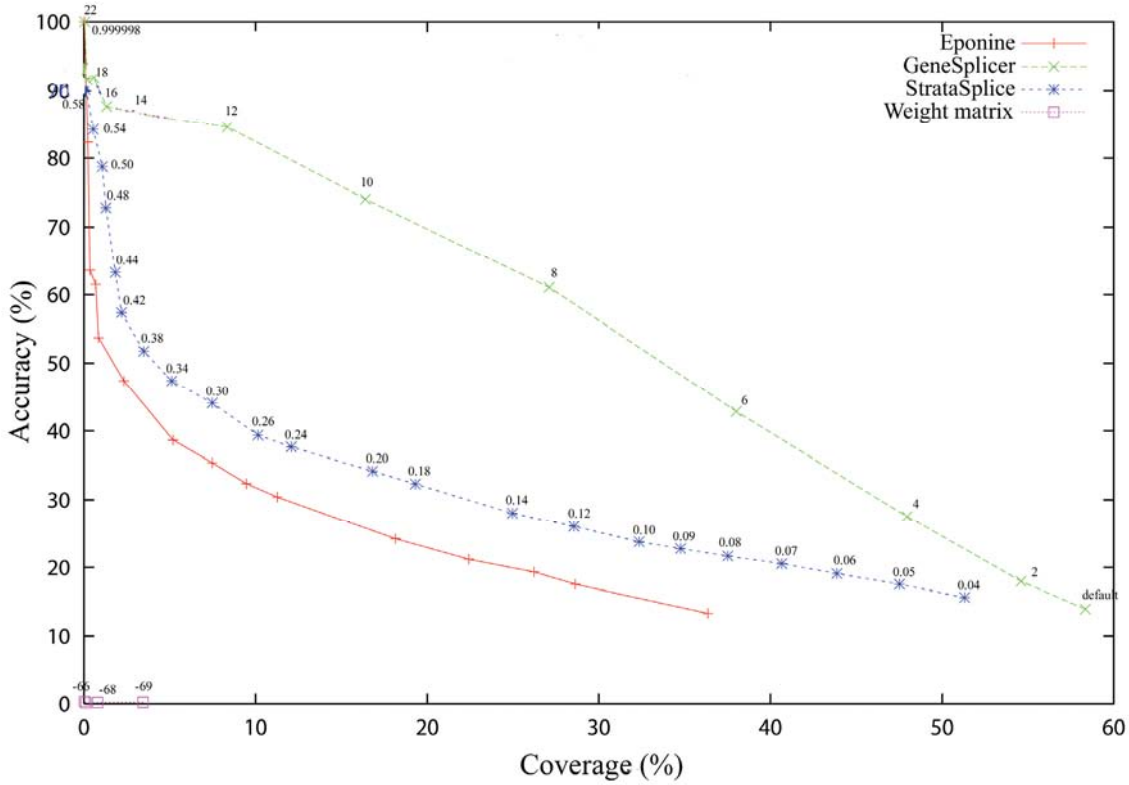

Figure 49. Exon coverage and accuracy on donor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

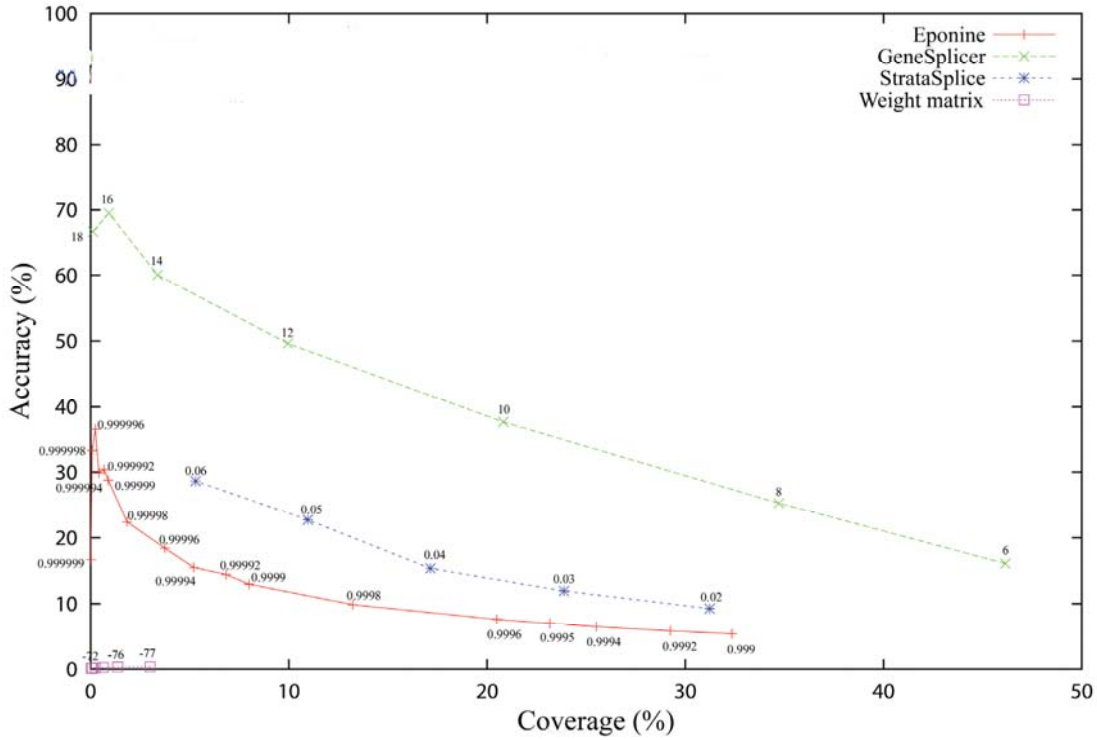

Figure 50. Exon coverage and accuracy on acceptor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

The results are not particularly surprising, as Eponine model is based only on positional weight matrices, whereas, StrataSplice uses local variation in GC content to differentiate true and false signals. The performance of StrataSplice has been shown to be higher at gene-rich regions of chromosomes (Levine, 2001a). GeneSplicer, apart from modelling sequence elements present near splice sites, uses coding/non-coding potential present in exons/introns near splice sites. A significant number of false positives by GeneSplicer are removed by choosing a splice site in a favourable sequence context. The context includes first, the availability of an appropriately spaced complementary splice site such that this pair of sites defines a potential intron and second, absence of nearby sites of the same type with higher score which could favourably compete with the given site for splicing factors (for details refer, Brendel and Kleffe, 1998). Employing both the favourable sequence context strategy and the coding potential is against the objective of developing an *ab initio* gene prediction system purely based on gene regulatory signals. Hence, although informative these strategies are not included in the Eponine models.

However, Eponine splice site models are shown to have significantly better performance than using donor and acceptor site weight matrices only (Figure 47 and Figure 48). These matrices are derived from chromosome 22 splice sites and are described in Figure 38. I scanned chromosome 20 sequence with the donor and acceptor site weight matrices and compared the predictions with the test set of 5835 donor and 6166 acceptor sites described earlier. Predictions extracted at different thresholds in both donor and acceptor site scan indicates weight matrices alone are not informative in predicting the splice sites. Similar results are found in predicting exon boundaries as well (Figure 49 and Figure 50). Thus Eponine splice site models although perform less well than GeneSplicer, they are found better than weight matrices.

Figure 51 shows the positional accuracy of Eponine models (Figure 45 and Figure 46) are equivalent to the predictions of StrataSplice and GeneSplicer.

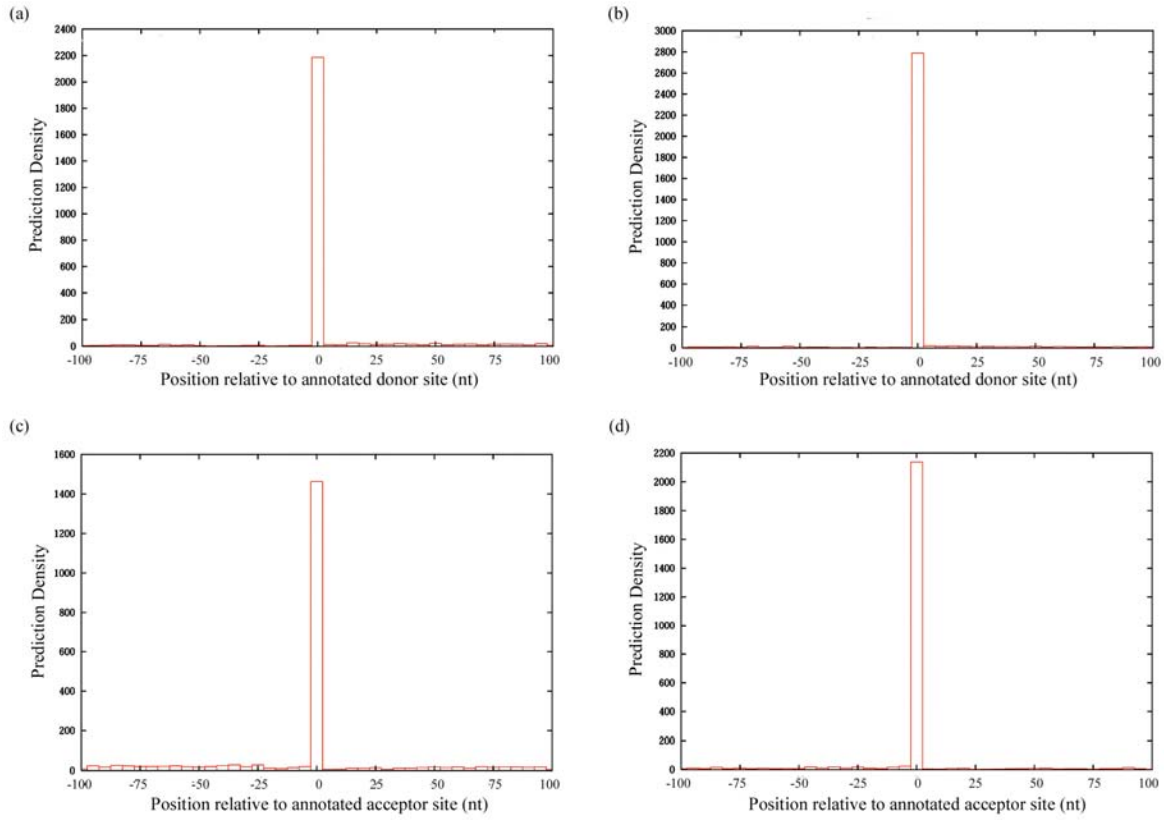

Figure 51. Prediction densities for StrataSplice and GeneSplicer donor and acceptor site predictions (a), (b) Densities for StrataSplice and GeneSplicer donor site predictions relative to the annotated site respectively (c), (d) Densities for StrataSplice and GeneSplicer acceptor site predictions relative to the annotated site respectively

Thus the comparison shows that the performance of Eponine is comparable with StrataSplice and GeneSplicer given the amount of information used to process DNA sequences in identifying splice sites. Also, it reveals the scope for improvement of Eponine predictions using local GC content in the gene rich regions.

4.8 Concluding remarks

With the requirement of splice site models to meet the objective of developing an *ab initio* model based on regulatory elements, I created donor and acceptor site models. Initial training cycles did not yield a sparse model and hence I used weight matrices derived from chromosome 22 as an input to the EAS system along with DNA sequences. A Delta distribution was used to capture the positional variation as it suits the problem better than a

Gaussian distribution. The models learnt the known consensus signals and they are used to predict splice sites in chromosome 20. On comparison, the performance of the Eponine model was found better than using weight matrices alone. Availability of these models facilitates construction of a gene prediction program and to determine the structure of the predicted genes.