MODELLING TRANSLATION START AND STOP SITES

5.1 Introduction

As explained in chapter 1, transcription and translation, understood to be coupled together in prokaryotic organisms, have now been found to be interlinked in eukaryotes as well. Non-sense Mediated Decay (NMD) and protein synthetic capability in the nucleus add support for this view (for review, Cook, 1999). NMD is triggered due to the encounter of a pre-termination codon by the translating ribosome machinery (for details refer, Hillman *et al.*, 2004; Iborra *et al.*, 2004). So just like transcription, the translation mechanism is also under the control of regulatory elements on the RNA (transcribed from DNA). Translation start and stop signals are important regulatory signals and so far various methodologies have been used to study them.

With detailed knowledge of translation regulatory elements and machinery, computational detection of translation start and stop codons and their auxiliary sequences has been relatively easy and techniques from simple positional weight matrices to artificial neural networks to support vector machine have been used for this purpose. Almost all the translation start models are based on the important *Kozak* consensus sequence (Kozak, 1987) at the translation initiation site in detecting the start codon. However later algorithms, in an effort to improve prediction coverage and accuracy, used other features as well. Among them, detecting the coding potential of the sequence following the start codon, open reading frame length (the distance between the predicted start and stop codon) and distance of first ATG from the start of the sequence took a serious role. In addition, a few techniques even analysed the density of trimers, tetramers and pentamers in the sequences before the start codon and its property for non-coding potential. These properties, although improving the prediction system, giving it better accuracy and coverage, have limited it to be used successfully in cDNA and EST sequences or incorporated into an *ab initio* gene prediction system. However as standalone programs for prediction on genomic sequences they are likely to make numerous errors. So to address this issue, here I attempted to use the Eponine model trainer to learn translation start and stop signals. Also, the translation models can be used with other Eponine models in devising an *ab initio* gene prediction system like splice site models explained in the previous chapter.

In the remainder of this chapter, I will explain the datasets and parameters used to derive translation models and compare their performance with existing programs.

5.2 Datasets

5.2.1 Translation start model

Deriving a reasonable dataset with better annotation is one of the key factors in deriving any prediction models. Screening for such data from a massive amount of unannotated and incomplete cDNAs and ESTs is a formidable task. With annotated data the issue of upstream ATGs has to be addressed. Translation as explained earlier is known to occur by a cap dependent scanning mechanism or cap independent internal initiation process. The common cap dependent process helps the ribosome to start translation from the first ATG it encounters from the 5' end. However it is not always the case and at times ATGs further downstream can be used. The presence of multiple ATGs at the 5' end may confuse annotators leading to the identification of wrong ATGs as translation initiation sites. One estimate shows about 37% of human and 36% of mouse sequences in the 5' UTR database (Pesole *et al.,* 1996) have upstream ATGs with reference to annotated translation start sites (Rogozin *et al.*, 2001). Thus deriving a dataset with correctly annotated translation start sites is one of the most difficult steps.

As explained earlier, Eponine trainer requires two kinds of dataset – A *positive dataset* having DNA sequences that are likely to have translation start sites and a *negative dataset* with sequences of no such sites. Here I used two sets of positive sequences – one from genomic and another from cDNA for training an EAS model.

(A) Genomic sequences – Using the annotation of coding sequences in human chromosome 22 by (Collins *et al.*), 330 sequences with translation start sites and at least 200 bases of 5' UTR were extracted. Two hundred nucleotides upstream and downstream of the ATG codon formed the positive set, *pos-1*.

Likewise, 200 bases on either side of the start codon from 506 transcripts of 327 genes in chromosome 20 were dumped from VEGA database (Ashurst, 2002). This dataset formed

another positive set, *pos-2*. Only transcripts from the 'known' gene category from the VEGA database were used here.

Combining both *pos-1* and *pos-2* datasets, a set of 836 transcripts with annotated translation start sites was formed *pos-3*.

Random and intergenic sequences of 400 nucleotides from chromosome 20 and chromosome 22 were dumped to form negative datasets for training the EAS model from *pos-1*, *pos-2* and *pos-3*. Equal numbers of positive and negative sequences were used for each training cycle.

(B) cDNA sequences – From the Reference Sequence database (RefSeq, Pruitt and Maglott, 2001), 14038 cDNA sequences were dumped in EMBL format. Out of this 5693 sequences were categorised as 'provisional', 2350 as 'predicted', 2523 as 'curated' and 3472 as 'genome annotation' based on the types of evidences and annotation done on these sequences. Reviewed RefSeq records represent full length cDNA sequences with manual curation of gene features. Hence I took the 2523 sequences and screened for the ones with at least 200 bases of 5' UTR and resulted in a subset of 676 sequences. Out of these 676 sequences, only 563 sequences have annotations in the ENSEMBL database (Birney *et al.*, 2004) and thus were used for training purposes. The positive dataset, *mpos-1* was derived by extracting 200 bases on either side of the ATG codon present in these 563 sequences. An against all BLAST (Altschul *et al.*, 1990) search was carried out on the *mpos-1* set to make sure no identical sequences were present in the positive dataset. The remaining 113 (676-563) sequences were used for testing the models.

As training on cDNA sequences will tend to be biased towards learning coding potential of the sequences downstream of the ATG codon, two types of negative datasets were synthesized to tackle it. Two hundred nucleotides of noncoding or intronic sequence (from intergenic or intron regions of chromosome 22) and 200 nucleotides of coding or exonic sequence (from exon regions of chromosome 22) were concatenated together to form a 400 base pair negative sequence. This way, both the positive and negative set had exonic sequences downstream of ATG and hence the trainer is less likely to model the coding

potential in the positive set. A set of 563 such sequences (equal to the positive set) formed the negative set, *mneg-1*.

In another negative set (563 sequences), *mneg-2*, the intronic (196 bases) and exonic (197 bases) sequences are concatenated together with AXXATGG sandwiched between them. AXXATGG resembles the consensus sequence near translation initiation sites and by incorporating it in the negative set; the trainer is restricted to learn other position constraints that can meaningfully classify the sequences.

With these different positive and negative datasets I trained the EAS translation start model.

5.2.2 Translation stop model

I used the sequences from the 'PolyA site' database (Tabaska and Zhang, 1999) to derive a positive set for training the translation stop model. The database was formed by aligning ESTs of a UniGene cluster (Wheeler *et al.*, 2004) with all of its DNA and non-EST RNA sequences (for details, read Tabaska and Zhang, 1999). A hundred bases upstream and downstream of the stop codon were dumped from 124 sequences from the database to form a positive set. For a negative set of sequences where translation is unlikely to terminate, I extracted 124 sequences of 200 bases each from random regions of chromosome 20. Thus the randomly picked sequences are equal in length to the positive sequences. Out of 248 sequences (124 positive and 124 negative), I kept apart 28 sequences (14 positive and 14 negative) for initial testing of the model. These 28 sequences are randomly picked during different training runs. The remaining 220 sequences were used for training the model. The 113 human RefSeq cDNA sequences (explained earlier) set apart for testing the translation start model were also used for determining the performance of the translation stop model. The first base in the termination codon was used as the anchor point. Models were also trained with a training set derived from 200 bases upstream and downstream of this anchor point.

5.3 Training the translation models

5.3.1 Translation start model

With the datasets available, I initially used *pos-1, pos-2, pos-3* positive datasets and random negative sequences to train the EAS translation start model. The nucleotide A in the first codon, ATG was set as the anchor point. As explained earlier each positive and negative sequence is of 400 bases length, spanning 200 nucleotides upstream and downstream of this anchor point. During training the trainer is likely to fish out informative positional constraints from these sequences to identify positive from negative sequences. However selection of any constraints near to the edges of the sequence is likely to cause the trainer to cross the boundary, as it will be difficult to estimate a Gaussian distribution for such motifs. So to avoid such cases, I have limited the window size for screening for constraints to 160 bases either side of the anchor point. This appears to be sufficient to capture any regulatory motifs that determine the translation start site as training done with increased window sizes did not find any new constraints. However reducing the window size from 160 bases to 50 bases (-50 to +50 bases from anchor point) and 20 bases (-20 to +20 bases from anchor point) produced models with only ATG and *Kozak* motifs and thus had less predictive power than previous models.

The trainer with these datasets and default parameters was allowed to run for a maximum of 6000 cycles to learn a simplistic model that can significantly classify translation start site from other sequences. Typically each training run took nearly 1 hour in a personal computer with 1GHz Pentium CPU and 256 MB RAM.

A typical model learnt from the *pos-1* positive dataset and random sequences as the negative dataset was shown in Figure 52a. Different training cycles showed that positional constraints, especially those present downstream of the ATG codon, are not converging and the trainer tended to learn negative constraints. Likewise, models trained from the *pos-2* dataset also showed similar results (Figure 52b). The models from the *pos-2* dataset are even more complex with more negative constraints. Intergenic sequences as the negative dataset did not improve the model.

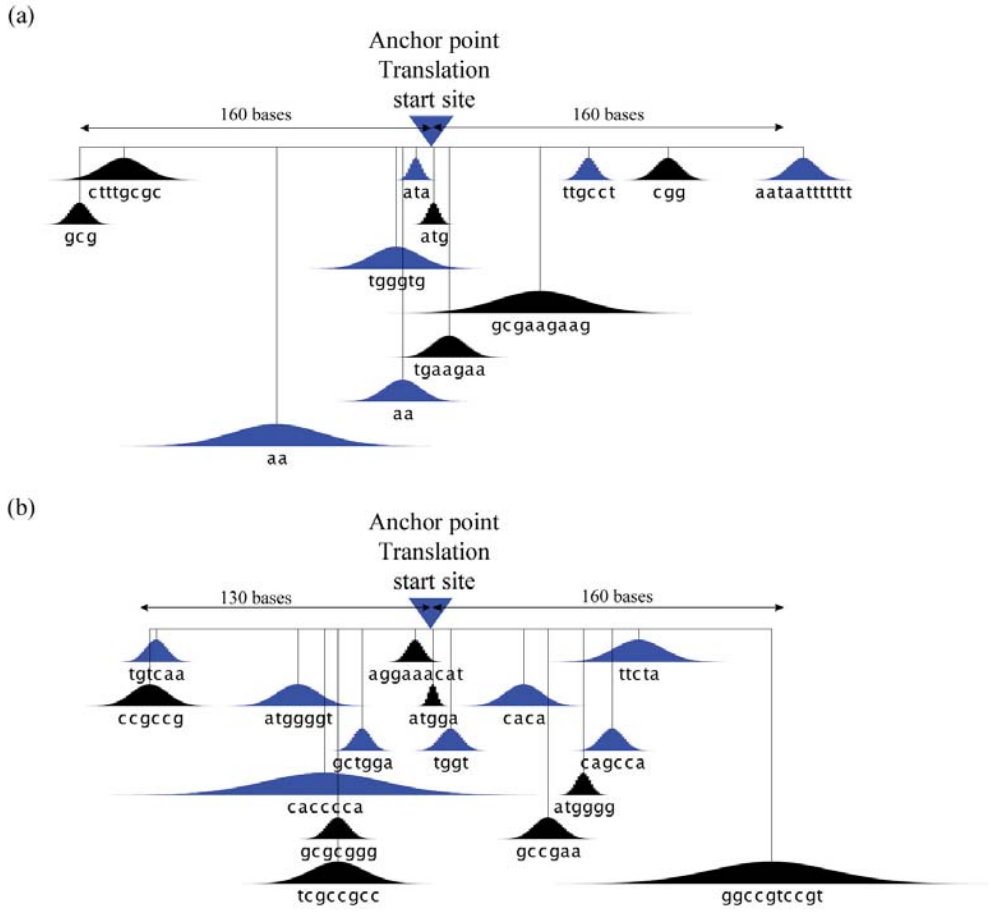

Figure 52. Translation start model trained from (a) chromosome 22 (b) chromosome 20 genomic sequences

Non-convergence of positional constraints might be due to existence of intronic sequences in the positive dataset. While extracting 200 bases downstream of the ATG codon to form the positive set, in cases where sequences followed by the start codon are less than 200 bases, nucleotides from introns are likely to be dumped and added to the positive set. Thus the variation present in the sequences downstream of ATG might be the cause for non-convergence of the model.

Hence to avoid this problem, I switched to training models from cDNA sequences using datasets *mpos-1* and *mneg-1*. A of ATG is again set as the anchor point with the trainer allowed for scanning positional constraints within 160 bases from it. The trainer ran between 5000 and 8000 cycles during various training trials. Examining different trained

models showed constraints that are positioned between 140 bases upstream and 120 bases downstream of the anchor point. This suggests that motifs that can identify translation start sites are closely associated with the start codon. Figure 53 shows a typical model trained from cDNA sequences.

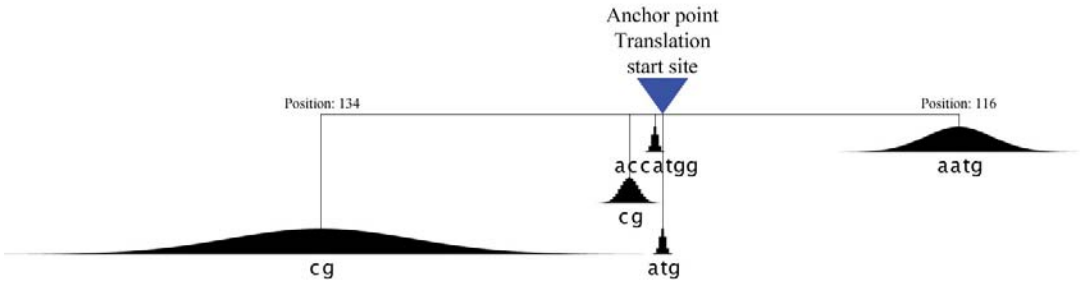

Figure 53. Translation start model trained from RefSeq cDNA sequences

Similar to the results found for genomic sequences, training Eponine models on cDNA datasets with window sizes - -20:+20, -50:+50, -75:+50, -75:+75, -100:+100, -150:+150, -170:+170 and -180:+180 did not yield better models. Models trained from window sizes less than 160 bases produced complex models and those with increased window sizes do not learn any new constraints.

A list of motifs found while training cDNA sequences along with their frequency of distribution is given in Table 8. ATG codon is represented in all the models and in few they are found more than once as indicated by the occupancy score.

The position, constraint weight and the Gaussian distribution width of the motifs represented in the above model is given in Table 9. The model obtained a strong signal for the ATG codon at position 0 with a narrow Gaussian distribution. A distribution width of 1.10 means most of the positional variation of the ATG constraint is within 3 bases from the point given in the table. This signal also has a bigger weight than the other constraints meaning the first codon is the strongest signal to determine the translation start site. Another strong constraint with a narrow Gaussian width is the *Kozak* motif positioned 3 bases upstream of the anchor point. The motif agrees with the previously reported consensus sequence (Kozak, 1987).

Table 8. Occupancy value for motifs detected in the translation start site models.

Number of models considered - 23	
Occupancy value for motifs below -20 bp	
Motifs	**Occupancy Value**
cg	0.83
atg	0.09
ccgcg	0.09
cgcg	0.09
gctggg	0.04
tcttc	0.04
cgcggcgc	0.04
ca	0.04
aaaat	0.04
cgccgcg	0.04
tgcccagct	0.04
cagatc	0.04
cctccc	0.04
ggctaac	0.04
aga	0.04
Occupancy value for motifs between -20 and 20 bp	
atg	1.35
at	0.26
atgg	0.17
gcaatg	0.13
gccatg	0.13
accatg	0.09
ccatg	0.09
gcgc	0.09
agtc	0.09
accatgg	0.09
tgg	0.09
atgatggt	0.04
aatgcc	0.04
aagatg	0.04
ca	0.04
Occupancy value for trts motifs above 20 bp	
aatg	0.26
aa	0.09
gaat	0.09
atgaa	0.09
acga	0.09
aatataatt	0.04
cgct	0.04
ctct	0.04
aacga	0.04
caggcct	0.04
aaaaataa	0.04
tg	0.04
ccgctcg	0.04
ccccgctc	0.04
cca	0.04

The sequence logo of the *Kozak* motif shows the interesting distribution of nucleotides at each position. The 1st and 7th position in the motif has higher distribution for A and G nucleotides respectively and this agrees with the importance previous computational methods have given for those positions in identifying translation start sites (Cavener and Ray, 1991; Hatzigeorgiou, 2002; Zeng *et al.*, 2002). The two other motifs found in the region upstream of the anchor point may capture the CG richness in the sequence between transcription start site and translation initiation site. The CG motif at position 134 bases upstream of the anchor point notably has a broad Gaussian distribution and may represent so called CpG islands, known to be associated with the 5' end of genes. An interesting constraint in the model is the AATG motif centred at 116 bases downstream of the anchor point. This motif has not been reported previously by other machine learning algorithms. It is not clear if this motif can act as another ATG codon and serve as an alternative translation start site. The Gaussian width for this motif is 13.19, meaning that most motifs would occur between 75 and 150 bases downstream of the start codon. It will be interesting to test if this constraint is involved in the leaky scanning mechanism of the translation machinery.

Table 9. Position constraints of translation start model learnt while training RefSeq cDNA sequences

MOTIFS	POSITION	CONSTRAINT WEIGHT	GAUSSIAN WIDTH
	-134	4.86	36.02
	-13	3.14	3.55
	-3	6.43	1.10
	0	15.98	1.10
	116	7.81	13.19

Overall the model given in Figure 53 appears to have captured both previously known regulatory motifs and the additional interesting AATG motif positioned downstream of the anchor point. Unlike other methods the model does not rely on constraints based on the distance between the 5' end of a cDNA sequence and the ATG codon. This means the model can be used effectively on the genomic sequences as a standalone program to predict translation start sites. Also, the model should be less constrained upon the coding potential of the sequence following the start codon.

5.3.2 Translation stop model

The translation stop model was trained for nearly 5000 cycles and it took less than 1 hour in a PIII laptop. A typical model is shown in Figure 54 along with the sequence logo of motifs, position, constraint weight and Gaussian width in Table 10. Like the translation start model, the stop model is also sparse and informative. The model learnt the stop codon along with a few other sequence motifs. Two position constraints with positive weights were found upstream of the anchor point. The signal positioned at 57 bases upstream of the stop codon has relatively higher constraint weight (20.97) indicating the importance of the signal in determining the stop codon. The role of these signals and others shown with their occupancy score in Table 11 in determining the translation stop mechanism is not known.

Interestingly in this model, there is a position constraint with negative weight (-2.74) just upstream of the stop codon. This constraint – 'TTT' motif represented in blue, simply means, the motif is expected not to be present near the stop codon. However the stretch of U residues is likely to behave as positive signal downstream of the stop codon in the 3' UTR region. This can be inferred from the CCTTT motif positioned 63 bases downstream of the anchor point. Thus poly U residues are less likely to be seen in the upstream than in the downstream of a functional stop codon.

The genetic code table has 3 stop codons – UAG, UGA and UAA. However not all the 3 stop codons are used equally and most organisms have a preference for one of them. In humans, UAA stop codon is the most commonly used. The sequence logo of the TAACC motifs shows, the model has learnt all the three stop codons with preference for UAA. The A and G nucleotide in UAG and UGA stop codons are also modelled separately with a distribution of AG motif at the anchor point. The nucleotide distribution of two bases

following the stop codon is almost equal and hence warrants no emphasis. However the biological implications of the two bases are not known.

Models trained with 200 bases upstream and downstream of the anchor point did not show any improvement over this model. This emphasizes the constraints near to the stop codon are more informative, making the model compact.

Table 10. Position constraints of translation stop model learnt while training RefSeq cDNA sequences

MOTIFS	POSITION	CONSTRAINT WEIGHT	GAUSSIAN WIDTH
	-84	5.15	4.01
	-57	20.97	4.21
	-17	-2.74	5.09
	-2	21.42	2.91
	0	6.95	2.91
	63	12.74	7.66

Table 11. Occupancy value for motifs detected in the translation stop site models.

Number of models considered - 35	
Occupancy value for motifs below -10 bp	
Motifs	**Occupancy Value**
ttt	0.11
cg	0.06
tttta	0.06
ta	0.06
ctcttccacctcaagc	0.03
tctt	0.03
agtt	0.03
cgtgg	0.03
tttttg	0.03
ggtg	0.03
caacg	0.03
tttt	0.03
taatttt	0.03
tataat	0.03
ctacc	0.03
Occupancy value for motifs between -10 and 10 bp	
taa	0.71
taag	0.09
tag	0.09
ta	0.06
ag	0.06
tga	0.06
tgaag	0.06
tgact	0.03
tc	0.03
agtaac	0.03
gcgt	0.03
ggggc	0.03
ga	0.03
cac	0.03
acg	0.03
Occupancy value for motifs above 10 bp	
at	0.09
ccctttt	0.06
ag	0.06
aa	0.06
tccct	0.03
ctgcccc	0.03
gtataa	0.03
ccttt	0.03
accctc	0.03
agggt	0.03
tgaattcat	0.03
gctgccttctgccttccg	0.03
ca	0.03
ctcttt	0.03
ataatg	0.03

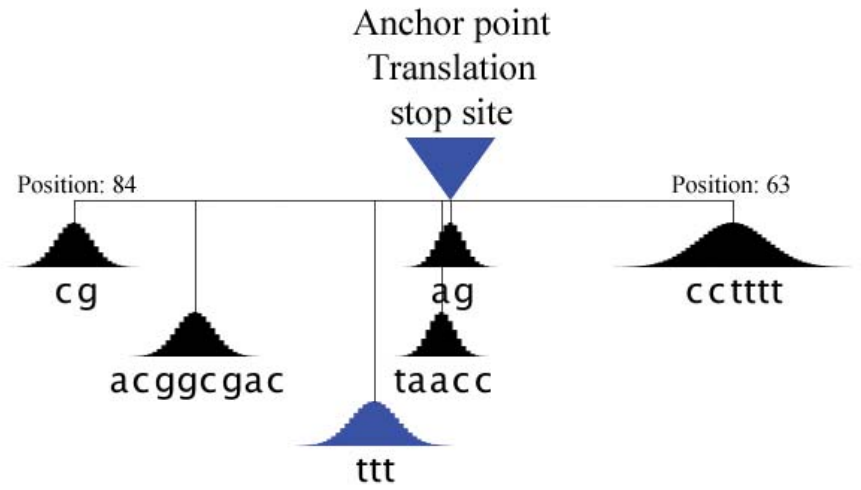

Figure 54. Translation stop model trained from chromosome 22 cDNA sequences

5.4 Validating and testing the models

I tested the performance of both the translation start and stop models in different datasets as explained below –

For quantification purposes, I defined a prediction as accurate if it is positioned within 200 bases upstream or downstream of the annotated start codon. Accuracy is calculated as the number of annotated start sites predicted over total number of predictions. Whereas coverage is number of annotated start sites predicted over total number of annotated start sites. Initial testing of the models was done on the set of sequences set apart while training. As explained before during each run, 226 (113 positive + 113 negative) sequences were kept apart from the trainer to be unseen while training. These 226 sequences were randomly picked up from the positive and negative set and thus vary for each run. So these test sets are fairly representative of the sequences available in the database and the model was tested on it. The performance of the model on this set was found to have good coverage and accuracy. However in this case, testing was done by scanning only the few bases around the anchor point to determine whether the sequence is a positive or negative. Hence, I used three independent test sets to analyse the performance of the models by allowing them to scan the whole cDNA sequence.

I took human reviewed RefSeq human and mouse and Riken mouse cDNA with at least 200 bases upstream to test the models. RefSeq database has cDNAs with different levels of annotation and thus they are not of equal degree. Among the different levels, manually reviewed cDNA are of high quality and I limited my test sets to these sequences alone. The 113 sequences used here were human cDNAs of this quality with at least 200 bases upstream. Predictions are made by scanning the sequence moving from left to right and evaluating the probability of the fit of the sequence motifs in the model in the cDNA sequence. The model found 3169 predictions (same strand) covering 87% (99 start codons out of 113, at threshold of 0.99) of annotated translation initiation sites. Figure 55 shows the ROC curve for the translation start model for predictions in the same (prediction in the same direction as of the gene), opposite (prediction in the reverse direction compared to annotation) and both (strand details are ignored) strands. Although the model predicted the strand of the annotated site correctly in most cases, few predictions are found in the opposite strand compared to the start codon. Combining predictions in both strands shows better coverage and accuracy than strand specific predictions.

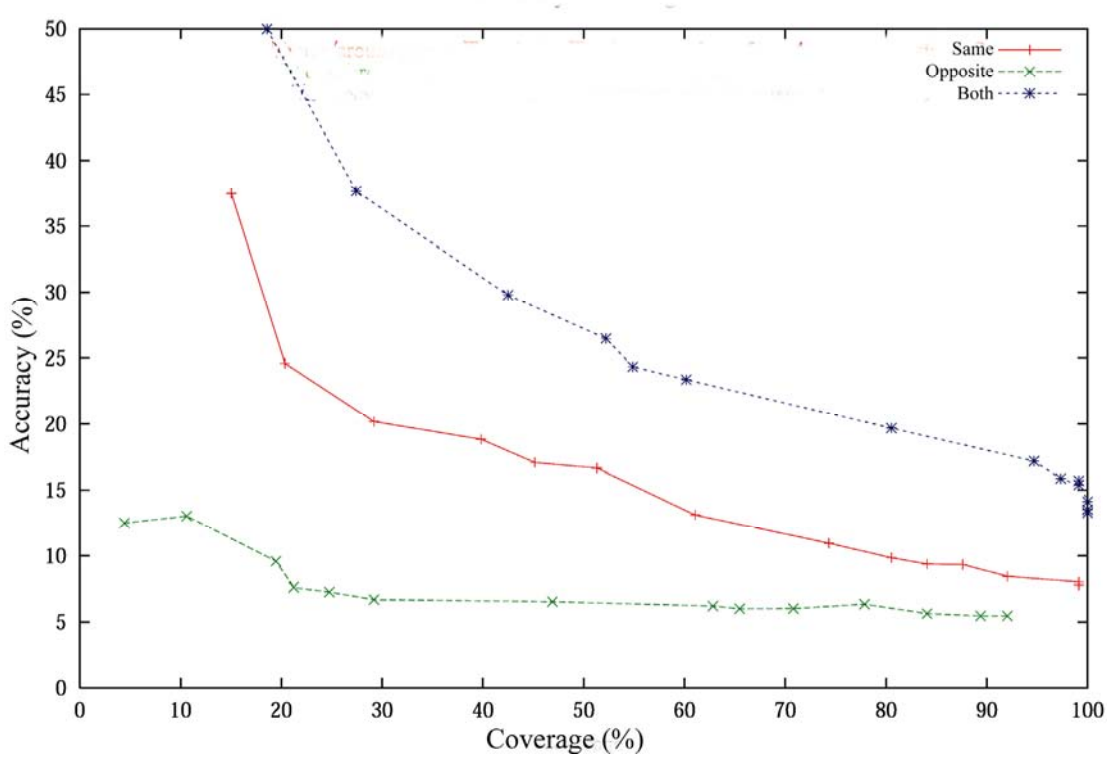

Figure 55. ROC curve on human RefSeq cDNA dataset for Eponine translation start site predictions in same, opposite and both strands

A set of similar quality entries from RefSeq was extracted for mouse cDNAs. This set resulted only in 37 sequences. The translation start model predicted 65% of the annotated start codons with 1537 predictions in total at a relatively less stringent threshold score of 0.95. The low coverage may be due to the requirement of 200 bases upstream of the translation start site by the model for scanning and only few sequences in the dataset met this criterion. Also some of the predictions might be true start sites and annotations are not available at present to validate them. In cases where the annotated sites are identified correctly, the positions of the predictions are limited to -2 to +2 bases from the annotated site. An ROC plot of the performance of model for this dataset is given in Figure 56a and Figure 56b for same and both strand predictions respectively.

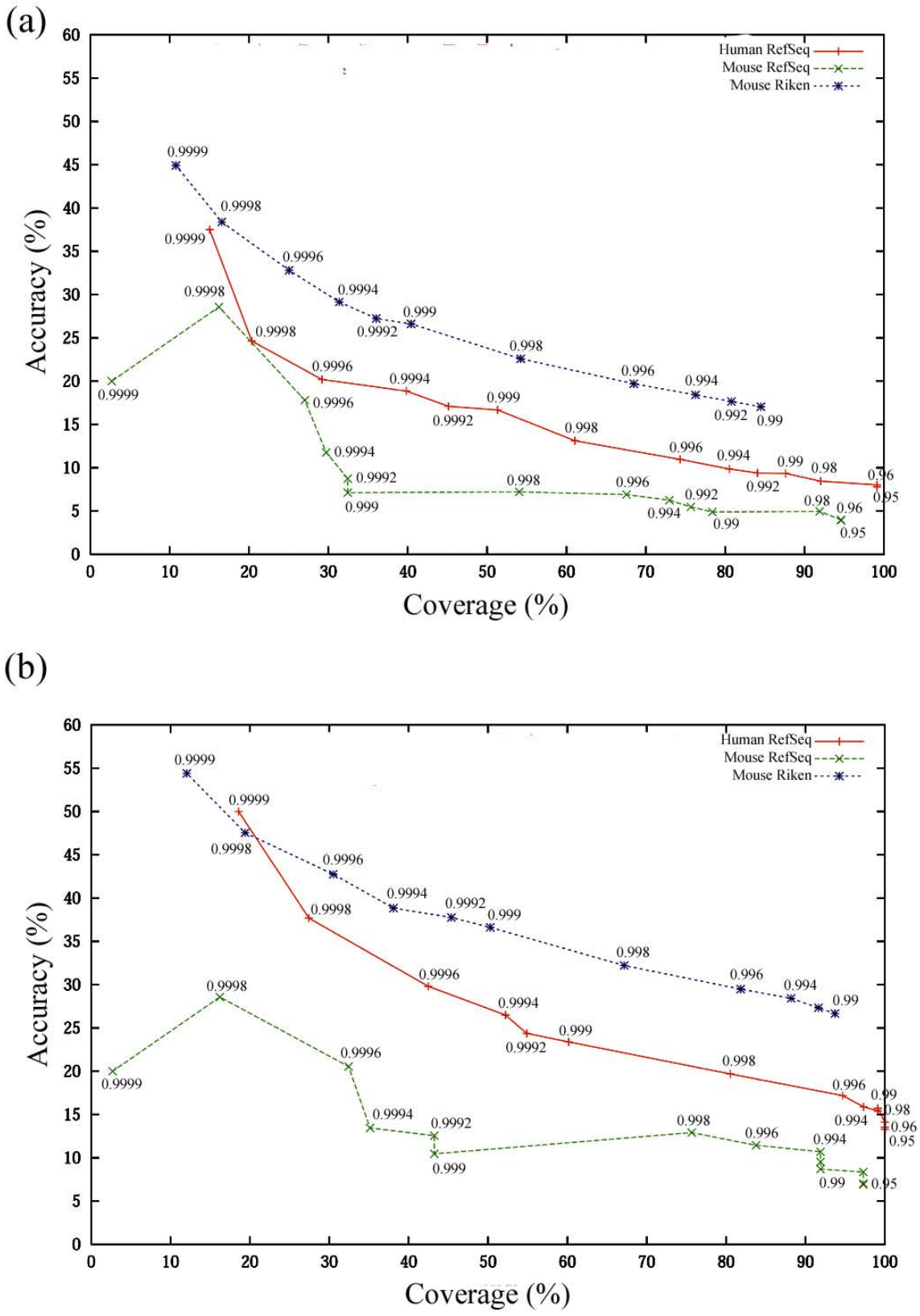

Figure 56. ROC curve for translation start model on human RefSeq, mouse RefSeq and mouse Riken cDNA datasets (a) Predictions in the same strand (b) Predictions in both strands

I extracted another set of mouse cDNA sequences of comparable quality from the RIKEN database. This set had 1593 sequences with each sequence having at least 200 bases upstream of the annotated start codon. The scanning of all these sequences using the model shown in Figure 53 gave 20400 predictions with a threshold value of 0.992. The predictions (same strand) covered 1287 translation initiation sites (80% coverage). ROC plots calculated from predictions in the same and both strands are given in Figure 56a and Figure 56b respectively.

Like the start model, the translation stop model was first tested on the test sequences set apart while training. As the test sequences are randomly selected from positive and negative sequences and they are different with each run of training, the trainer has less chance to 'overfit' the positive dataset.

Accuracy and coverage was calculated as explained above with 200 base tolerance in the prediction position relative to the annotated stop codon. Figure 57 shows the ROC curve for translation stop sites in human RefSeq dataset (113 sequences). The performance of the stop model is worse compared to the start model.

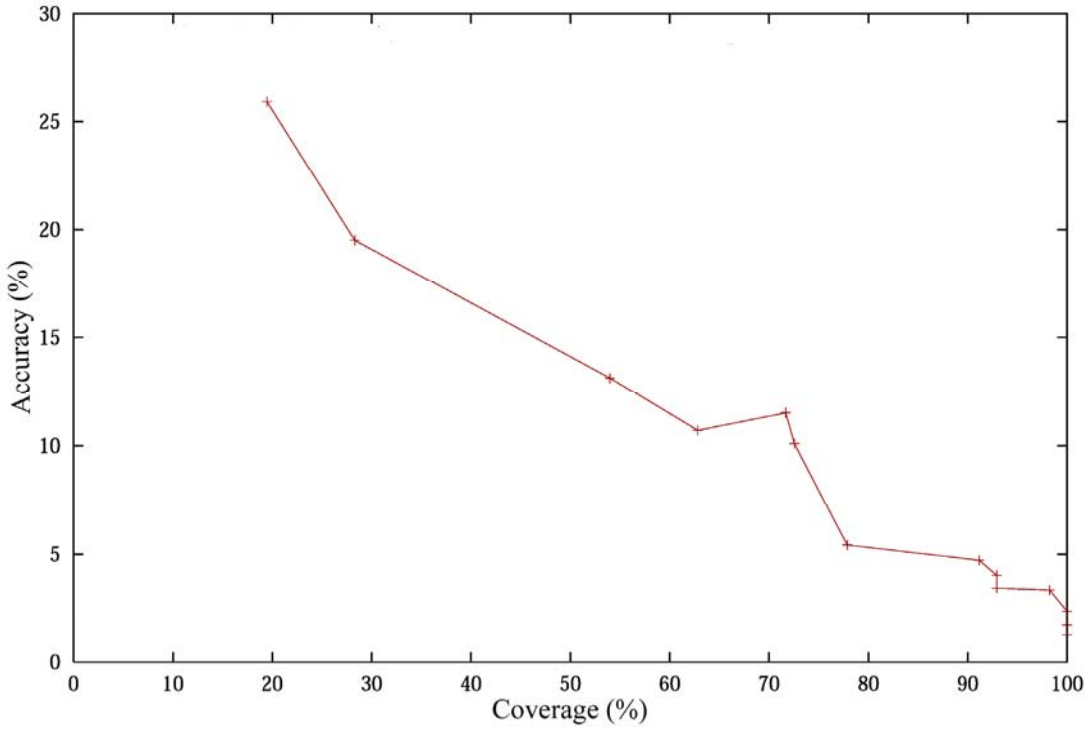

Figure 57. ROC curve on translation stop sites in human RefSeq cDNAs for Eponine model

5.5 Position accuracy of the models

5.5.1 Translation start model

As well as predicting most of the annotated start codon, the translation start model showed reasonable performance in determining the exact position of the start codon. The model is anchored on the A in ATG codon and any point in the sequence predicted by the model correlates with this nucleotide. I calculated the density of the predictions relative to the start codon and plotted the histogram for human RefSeq and mouse RIKEN datasets (Figure 58).

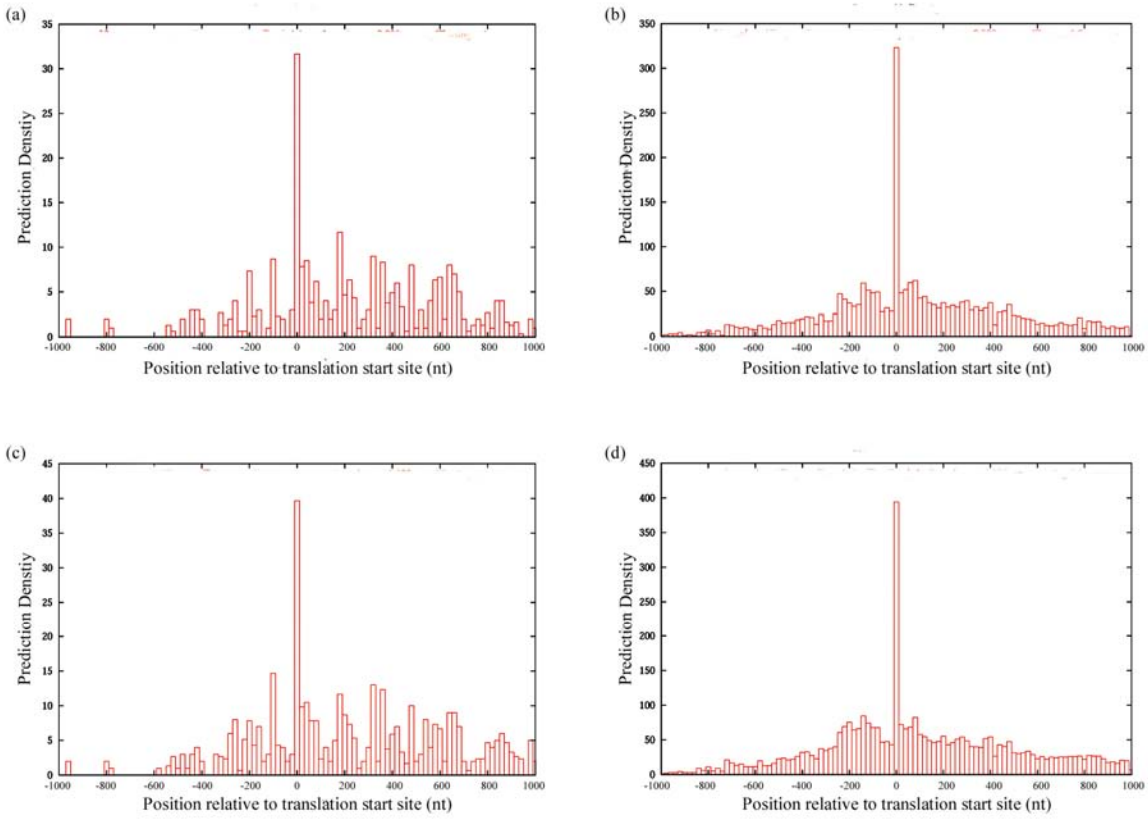

Figure 58. Prediction densities for translation start site model relative to annotated initiation sites (a), (b) Density of predictions in the same strand for human RefSeq and mouse RIKEN data respectively (c), (d) Density of predictions in both strands for human RefSeq and mouse RIKEN data respectively

The results show a clear peak, with many of the predictions centred within 20 bases from the annotated start codon. In some cases, the prediction positions are highly accurate and anchored at +1/-1 bases relative to the initiation site. However the majority of predictions are between -5 and +7 nucleotides relative to the anchor point. Figure 58 shows most of the

predictions are in the same strand as of the annotation although few predictions lie in the opposite direction. Even in cases, where the predictions are in the opposite strand, the predictions are concentrated within 20 bases from the start codon.

5.5.2 Translation stop model

The density of predictions made by this model on the human RefSeq set of sequences is shown in Figure 59. This model cannot predict the position of the stop codon correctly and there is no significant peak near annotated sites. I was surprised to note this result, as the model seems sparse and informative and captured known consensus signals. Further analysis might yield better results in predicting translation stop sites. Nevertheless, the results here simply indicate that the end of translation is determined solely by the stop codon itself and not by any motifs in the surrounding sequence.

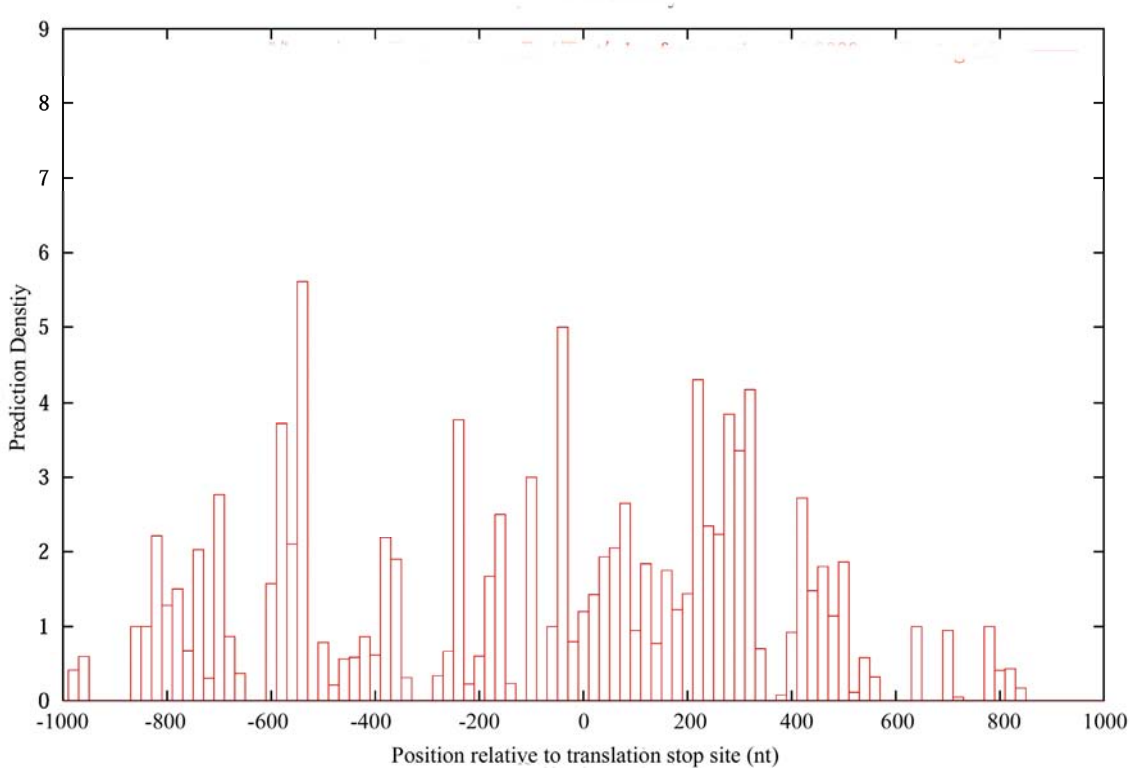

Figure 59. Prediction density for translation stop model relative to annotated stop sites

5.6 Comparison with other models

I compared the performance of Eponine predictions with two other translation initiation prediction programs, NetStart and ATGpr available in the public domain.

NetStart (Pedersen and Nielsen, 1997b) – This program was created as an improvement over using weight matrices to determine translation initiation sites. The program uses Artificial Neural Network (ANN) and was trained on 100 bases upstream and downstream of the start codon. The information surrounding the AUG codon was used primarily for prediction.

ATGpr (Salamov *et al.,* 1998a) – This program along with sequence context used six other characteristics to identify putative start sites. These characteristics are –

(a) Positional weight matrix around an ATG.
(b) Hexanucleotide difference between sequences upstream and downstream of the ATG codon.
(c) Preference for longer reading frames downstream of ATG
(d) Signal peptide characteristic
(e) Presence of another upstream in-frame ATG
(f) Upstream cytosine nucleotide characteristic

Linear discriminate analysis was used to generate a single score from the combination of these properties.

As neither of the programs is available for download, I used their web interfaces to scan human RefSeq test sequences. The NetStart (Pedersen and Nielsen, 1997a) web interface has a restrictions on the number of sequences submitted to the server (at most 50 sequences) and hence I split the dataset (113 sequences) into 3 sets and scanned them separately. The results from the three sets are then combined together for comparison. I used default parameters for vertebrate sequence given in the web-based predictor.

The ATGpr web interface (Salamov *et al.,* 1998b) has even more severe restrictions and it cannot take any sequence longer than 1300 bases and hence I split each sequence into 1150

base chunks with overlapping window size of 10 bases. These chunks were submitted to the server and predictions for a cDNA sequence were obtained by merging the predictions of each chunk. Default parameters for human sequence were used for prediction.

Figure 60 shows the performance of Eponine model compared to these two programs. The ROC curve shows, Eponine performs better than NetStart although less well than ATGpr. This was expected, as ATGpr uses additional information apart from regulatory elements to screen out false positives. NetStart which uses only sequence elements performs less well than Eponine.

5.7 Concluding remarks

Machine learning techniques assume the ribosomes operate in a linear fashion. NetStart developed by Pedersen and Nielsen (Pedersen and Nielsen, 1997a, b) based on a ANN was trained on a 203 nucleotide window centred on the AUG codon. The same dataset was used to train a Support Vector Machine model by Zien *et al.* and an improvement was obtained by using a kernel function to detect the codon bias in the downstream sequence of AUG. Likewise, Salzberg used a conditional positional probability kernel function to improve the ANN model using SVMs (Salzberg, 1997). More recently, Hatzigeorgiou reported a prediction program called DIANA-TIS based on a ANN trained on human sequences. This program combined a consensus ANN with a coding ANN together with the ribosome scanning model. Zeng *et al.*. used similar techniques by combining various informative features generated by different machine learning techniques. They found the following features useful: -3 and -1 position in the sequence relative to AUG; upstream k-grams for k = 3, 4 and 5; stop-codon frequency; downstream in-frame 3-gram; and the distance of AUG to the beginning of the sequence. The k-grams count the frequency of occurrence of a particular pattern in a window of length k that slides upstream and downstream of the AUG codon. Downstream in-frame 3 gram gives measure of the coding potential of the downstream sequences.

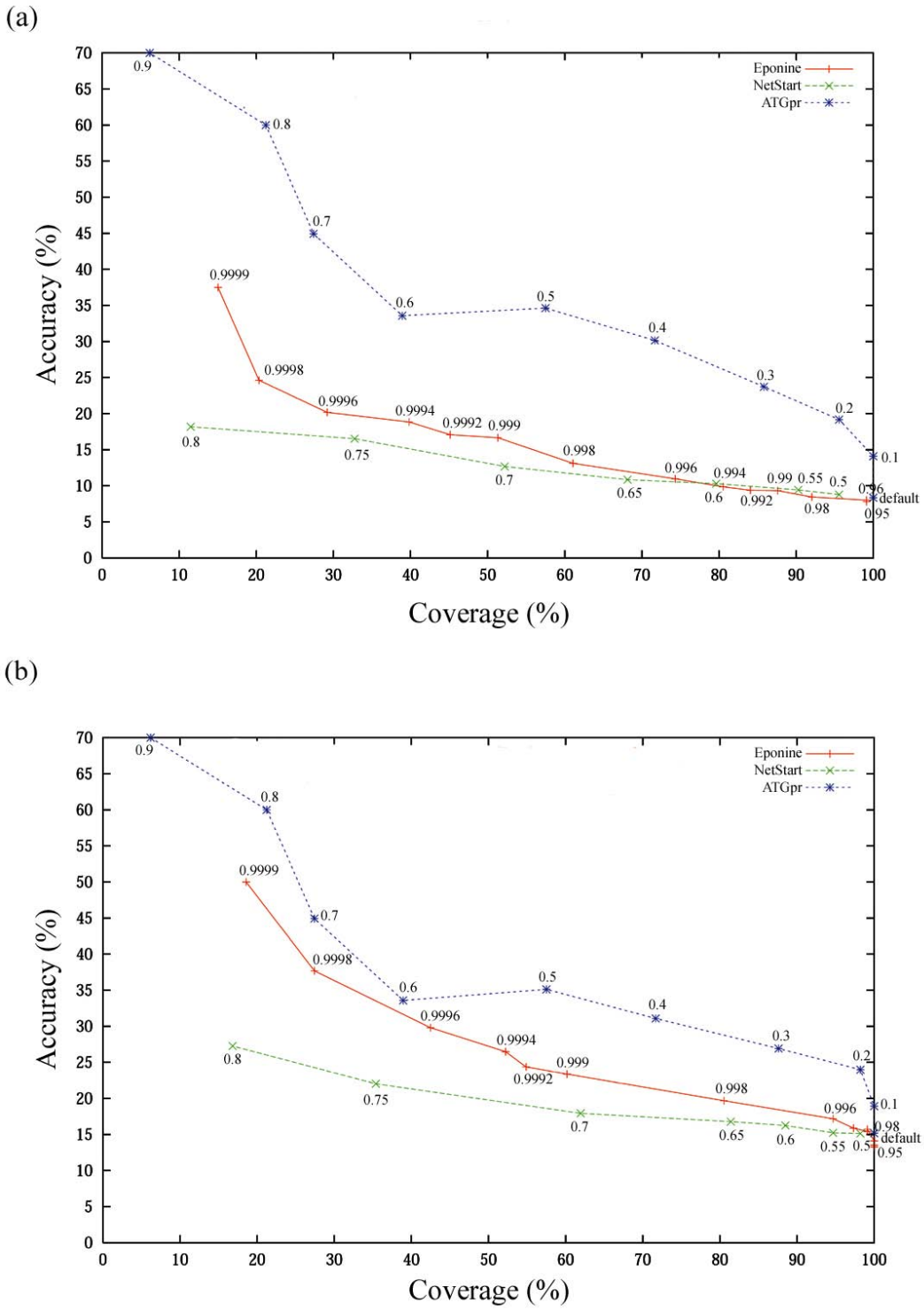

Figure 60. ROC curves for Eponine, NetStart and ATGpr on human translation start sites in RefSeq cDNA sequences without (a) and with (b) strand information

Thus these programs use a significant amount of ‘content’ information from the cDNA sequences to predict translation start codons. Here I attempted to make a translation stop and stop model that can scan genomic sequences and predict start and stop codons respectively purely based on regulatory signals. Despite using only signal information, the Eponine translation start model performed better than NetStart. The positional accuracy of the start model in cDNA sequences is good and few of the predictions are in the opposite stand relative to the annotated site.

With transcription, splicing and translation models learnt so far I show the advantage of making an *ab initio* gene prediction program combining them using GAZE in the next chapter.