GENEPRED – AN *AB INITIO* GENE PREDICTOR

6.1 Introduction

With the availability of models for transcription, splice site and translation, here I introduce an *ab initio* gene prediction system, *GenePred*, created using the regulatory signals identified by Eponine. As explained previously, almost all gene prediction programs use 'content' information, such as codon bias and ORF length. The gene prediction system explained in this chapter is different in this respect as the system uses only 'signal' information. Such a gene prediction system has an advantage over existing gene prediction algorithms in that it has the potential to identify non protein coding RNAs as well as coding RNAs. Recent analyses (Cawley *et al.*, 2004; Mattick, 2001) indicate that a huge amount of non coding transcription occurs within the cell and most of these RNAs are regulated in a similar way to protein coding genes. Various functions are attributed to these RNAs such as RNA interference, co-suppression, transgene silencing, imprinting and methylation. Few attempts (di Bernardo *et al.*, 2003; Rivas and Eddy, 2001; Rivas *et al.*, 2001) have been made to identify these RNAs computationally and so far with only limited success. A gene prediction program based on 'signal' information alone, and thus not biased due to 'content' information, should more closely mimic the biological system than existing gene prediction methods, as the *in vivo* transcriptional machinery does not use 'content' information while transcribing a genomic region. Content information has historically been used to assist computational detection of genes since signal based prediction alone has been insufficient (Guigo, 1997).

The gene prediction model explained in this chapter was constructed using a dynamic programming framework called GAZE (Howe *et al.*, 2002), which can combine features identified by predictive models, such as those described in the previous chapters. GAZE allows evidence for individual gene components to be assembled in order to predict entire gene structures. As explained in chapter 2, the method uses a dynamic programming algorithm to obtain (i) the highest scoring gene structure with the supplied features and (ii) posterior probabilities that each input feature is part of a gene.

In this chapter I explain the details of the features and gene models used in deriving various versions of the gene prediction system. Following this, I compare the performance of the system with the well established gene prediction program called GENSCAN (Burge and Karlin, 1997), as this program is assessed to be one of the best *ab initio* programs available in the public domain (Guigo *et al.*, 2000; Parra *et al.*, 2003). Towards the end of this chapter I revisit the performance of the transcription termination model given the context of splice site model predictions.

6.2 GAZE gene structure models

Many gene prediction programs have two common features –

(i) signal and content measures are used to detect components and regions belonging to genes
(ii) assemblage of these components into complete gene structure prediction for the sequence and scored against some measures

For the first of these steps, different measures, say weight matrices, codon bias, pentamer and hexamer frequencies and splice site predictions can be used to distinguish the components of gene structure from the sequence. For the second of these steps, a choice must be made as to the *model* of gene structure over which the assembly is to be performed. One of the advantages of GAZE is that it decouples these two steps of assembly of signal and content data into gene structure predictions from the generation of the data itself. The inputs for both these steps are provided externally and GAZE does not work directly with genomic DNA. In this project, for the first step, I used Eponine predictions as signal features, which I explain in the next section. For the second step, I used the following models (Figure 61) to validate the assembled components of the gene signal features.

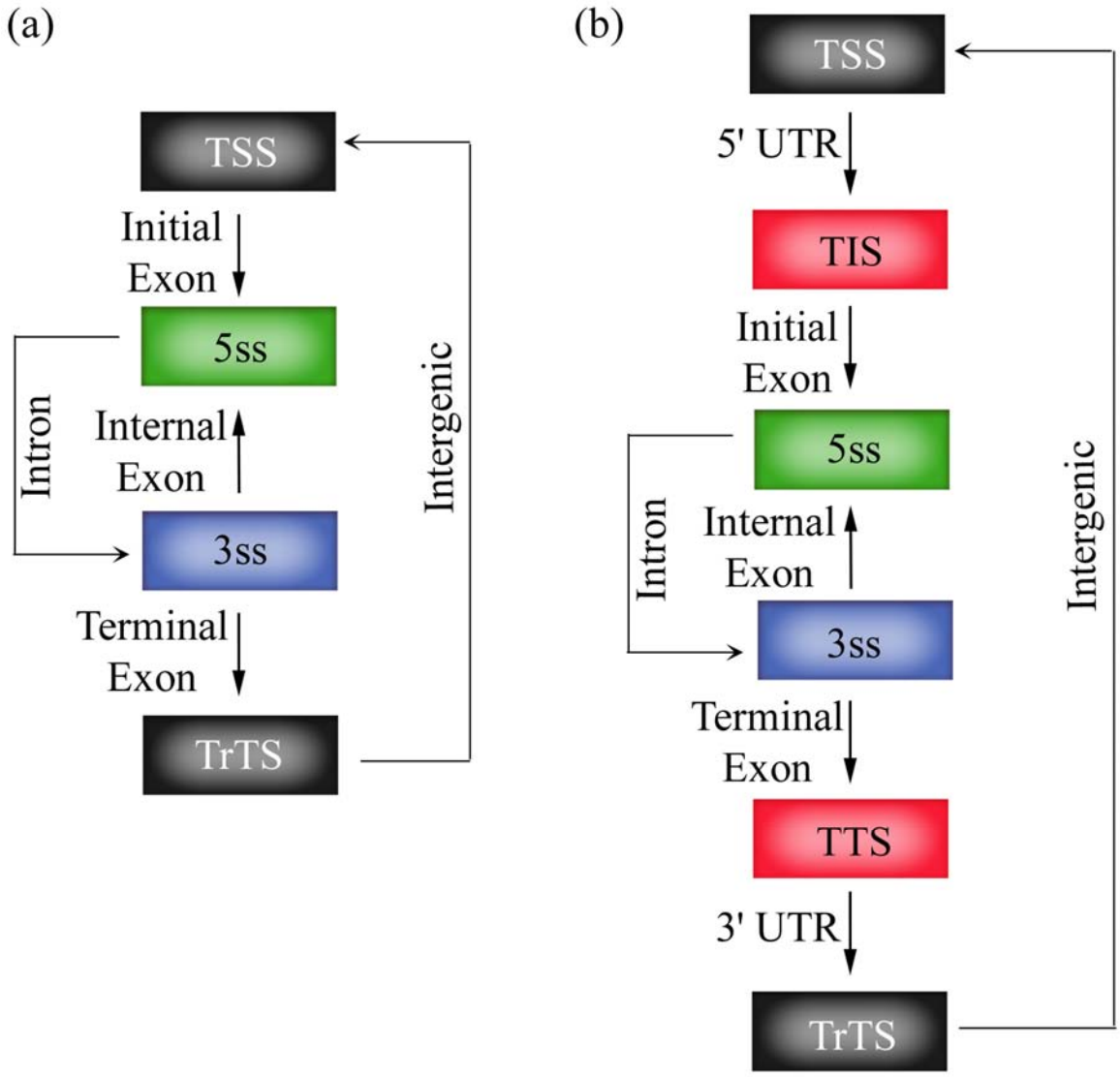

Figure 61. Schematic representation of the gene models used for predicting genes from features in the forward strand. Reverse complementation of the forward strand rules are used for reverse strand gene predictions. (a) Simple gene model without translation models and thus no protein information. (b) Gene model with translation features. Any introns within 5' UTR region are not modeled. Based on these gene structures, candidate genes are predicted on both strands at the same time.

In Appendix C, I have given the configuration files where the gene structure models used are presented in GAZE-XML format. A pictorial representation of these gene structures is given in Figure 61. The configuration file has five sections –

(i) *declarations* – declares the Eponine features that GAZE is going to work with

(ii) *gff2gaze* – dictates how the input files are used to obtain a list of features

(iii) *dna2gaze* – allows for the creation of features from simple sequence motifs observed in the input DNA sequence

(iv) *model* – contains the gene structure rules

(v) *lengthfunctions* – this section describes the length penalties used in defining exons, introns and intergenic regions in the model

The gene structure rule I used here (Figure 61a) is simple and it starts with a transcription start site followed by the donor site. The region between these two predictions defines the *initial exon* segment. The *introns* that interrupt the coding region of the gene are modelled by allowing a transition from donor to acceptor site. Introns might occur between two codons or in the middle of a codon, either between first and second position or between second and third positions. However, since the aim is not to consider any coding information in constructing the gene model, the phase associated with intron interruption is not considered. The donor and acceptor site features are represented as *5ss* and *3ss*. The sequences between a *3ss* and a *5ss* feature forms the *internal exon* of the gene structure. The *terminal exon* is defined as that part of the sequences between an acceptor site and a transcription termination site. Transcription termination site defines the end of the candidate gene. Thus a gene structure is defined with the features from transcription and splice site signals. To form the next gene another list of features are sampled and analysed to fit the rules explained above. That part of the sequence between two genes defined between transcription termination and start features is referred to as *intergenic*. To predict genes in the reverse strand, reverse complementation of the above rules are employed. Single exon genes are not modeled in this case. This is due to the fact that a simple single exon gene model will use only transcription start and termination site without any splice site model predictions. Allowing this simple single exon gene transition will bias the gene structures to terminate just after the start site because of the unusual presence of termination signals near transcription initiation site (refer to chapter 3).

Thus, this gene model without translation components more realistically mimics the biological transcriptome and spliceosome machinery that transcribes the DNA and processes the newly synthesized RNA respectively.

All the features are derived from Eponine models for predicting genes and I did not use the *dna2gaze* section to create a set of features from the DNA sequences. Similarly no constraints on the maximum length of exons and introns are placed in the gene model and thus no length penalty functions are used.

Figure 61b shows a pictorial representation of a different gene model used in predicting genes. In this structure, I used Eponine translation start and stop models as well. The transcription start site is now allowed to transit to the translation start codon, thus defining a new segment called the *5' UTR*. The region between the translation start codon and donor site now defines the *initial exon* segment. Similarly, the translation stop model is incorporated after the acceptor site of the last exon before transition to the transcription termination site. This change will make the GenePred system emit the *3' UTR* segment. By adding translation models and thus start and stop codon signal information, some protein coding information is attached to the gene prediction system. This is done to analyse the influence of the translation models in the GenePred system.

6.3 Eponine prediction models

As explained earlier, given a candidate set of gene features, GAZE predicts genes by deriving a subset of features that according to the given gene structure is the most likely candidate. The gene structure scoring the highest value with the list of features is predicted as a candidate gene. In order to provide the list of features to GAZE, I used Eponine model predictions. The following models are used along with their respective thresholds (given in brackets) to obtain predictions from signals in the DNA sequences.

(i) Transcription Start Site model (0.99)
(ii) Translation Start model (0.99)
(iii) Donor Site model (0.999)
(iv) Acceptor Site model (0.9998)
(v) Translation Termination model (0.999)
(vi) Transcription Termination model (0.99)

Apart from Eponine models, I also used GeneSplicer predictions while testing the performance of the GenePred system. GeneSplicer was used with default options to predict splice site features from the DNA sequence.

All the predictions were dumped in the General Feature Format (GFF, WTSI), a widely used standard for the exchange of gene prediction information.

Here I used chromosome 20 for scanning features and predicting genes as all the Eponine models discussed in previous chapters are trained from chromosome 22.

6.4 Gene prediction with Eponine features

With the availability of features from chromosome 20, I combined them to create a gene prediction system by inputting the features and the gene model structure (Figure 61a) into GAZE.

Figure 62 and Figure 63 show the genes predicted using GenePred as red tracks (the first red track in Figure 62 and the last red track in Figure 63) for a 1 mega base region (57.35 to 58.35 bases) of chromosome 20. For ease of comparison, in Figure 64, I removed all the annotation tracks and kept only VEGA, ENSEMBL annotations and GENSCAN (Burge and Karlin, 1997) predictions.

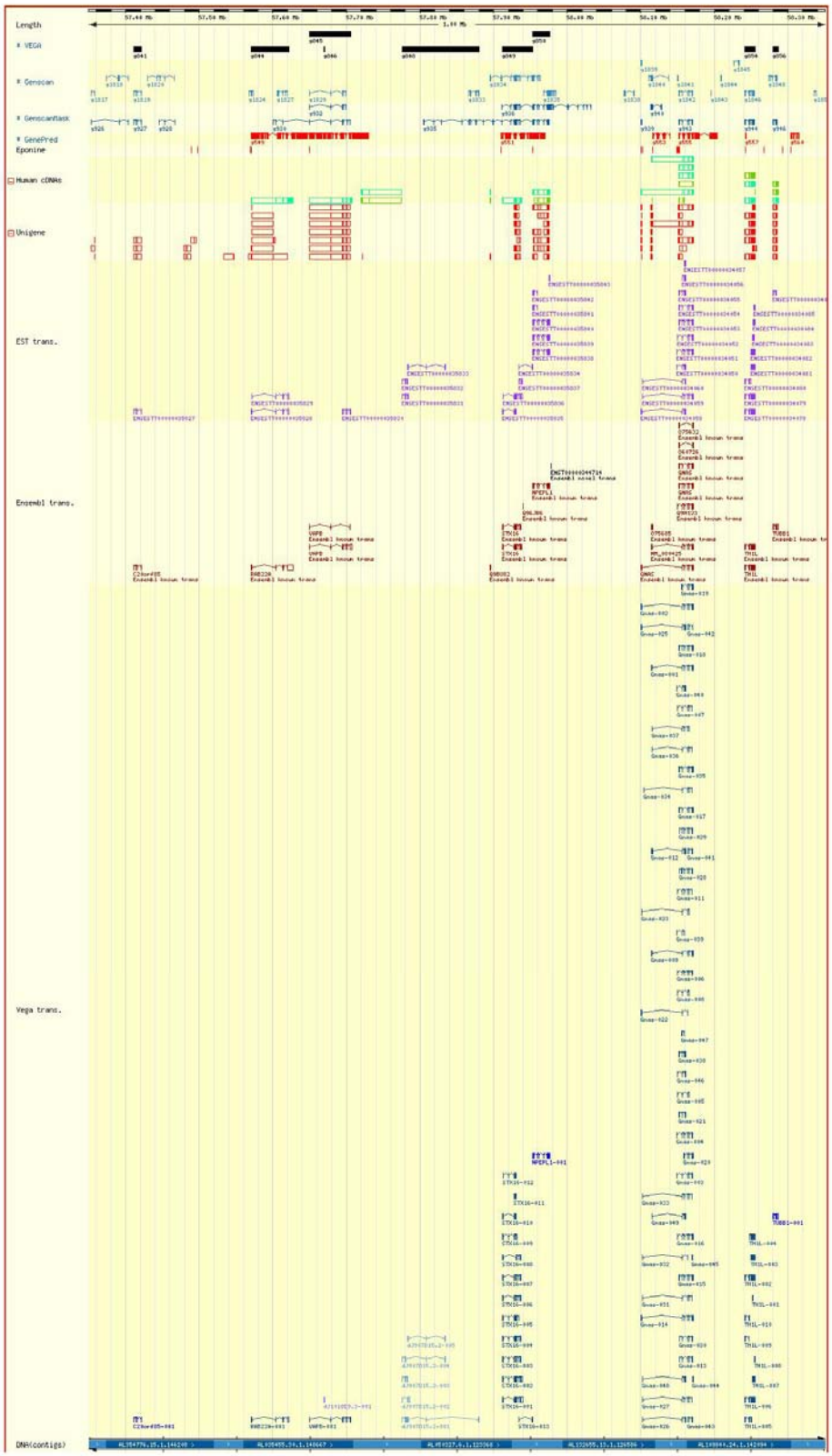

Figure 62. Genes predicted by linking Eponine models using GenePred compared with annotations available in the forward stand. Annotations from VEGA, ENSEMBL, EST transcripts, UNIGENE and Human cDNAs are shown as tracks along with GENSCAN predictions (both on masked and unmasked sequence). The comparison is possible with the ENSEMBL ContigView which can load predictions from external source as DAS tracks.

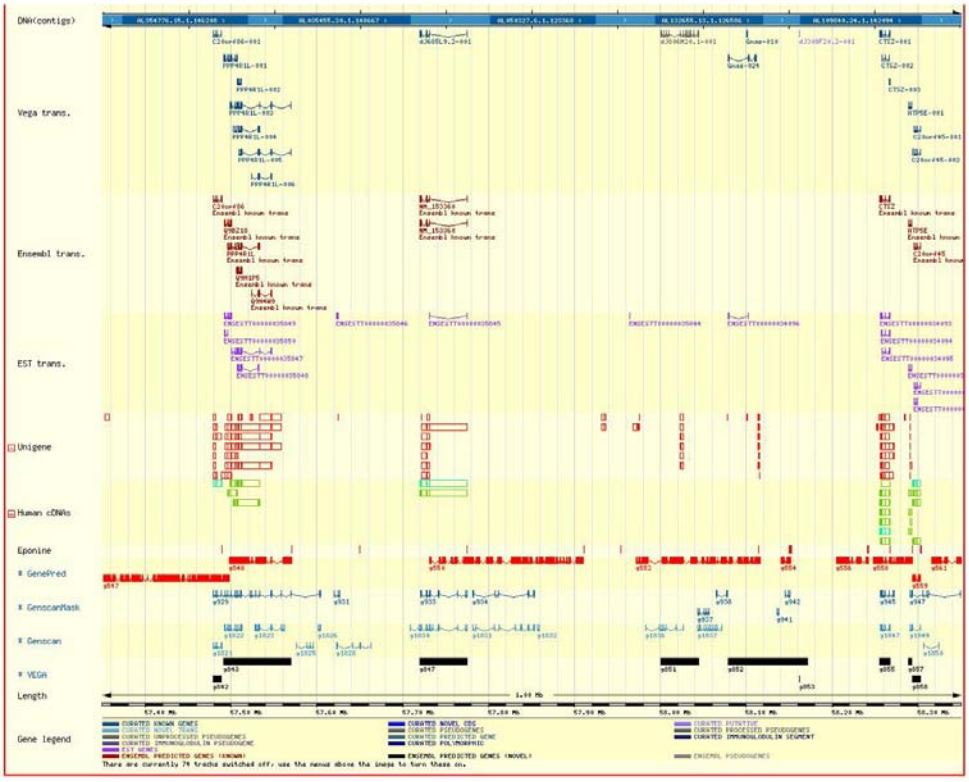

Figure 63. Genes predicted by linking Eponine models using GenePred compared with annotations available in the reverse stand. Annotations from VEGA, ENSEMBL, EST transcripts, UNIGENE and Human cDNAs are shown as tracks along with GENSCAN predictions (both on masked and unmasked sequence). This figure is reproduced from ENSEMBL ContigView viewer.

GENSCAN predictions are derived by scanning repeat masked chromosome 20 sequence. This is done by splitting the chromosome sequence into 200 kb overlapping blocks and GENSCAN predictions on each block are then merged together using a merging algorithm (Hubbard, T., personal communication) to derive the final list of predictions.

I compared the performance of GenePred with that of GENSCAN using the following definition of coverage and accuracy –

(i) Coverage is defined as the number of genes identified over the total number of annotated genes.

(ii) Accuracy is calculated as the number of predictions matching the annotation over the total number of predictions. Predictions that fuse or split the gene are considered as false positives (Figure 65). This included a few predictions matching genes that have an internal gene in the same strand.

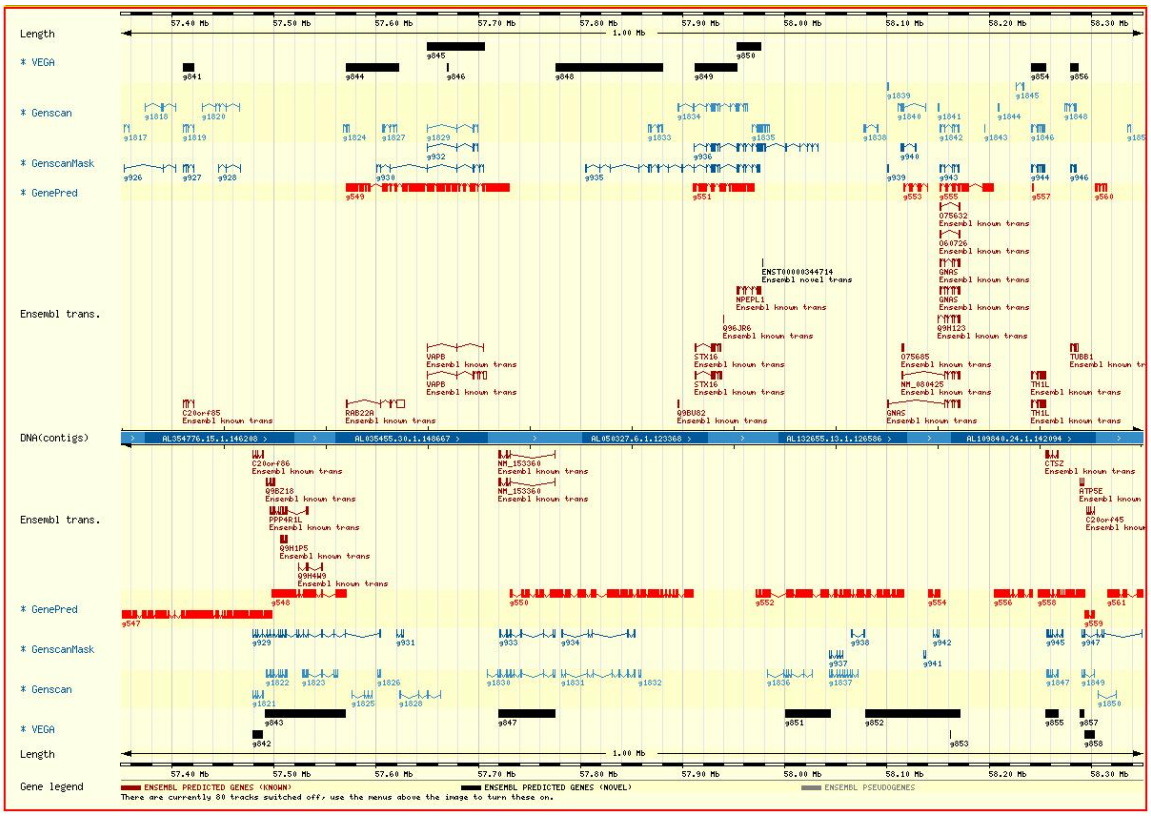

Figure 64. Genes predicted by linking Eponine models using GenePred compared with annotations available in both strands. VEGA annotations are shown as black bars. The region covered by a bar includes all the alternative transcripts of a gene. GenePred predictions are given in red color. The figure also shows GENSCAN predictions and ENSEMBL annotations in different tracks.

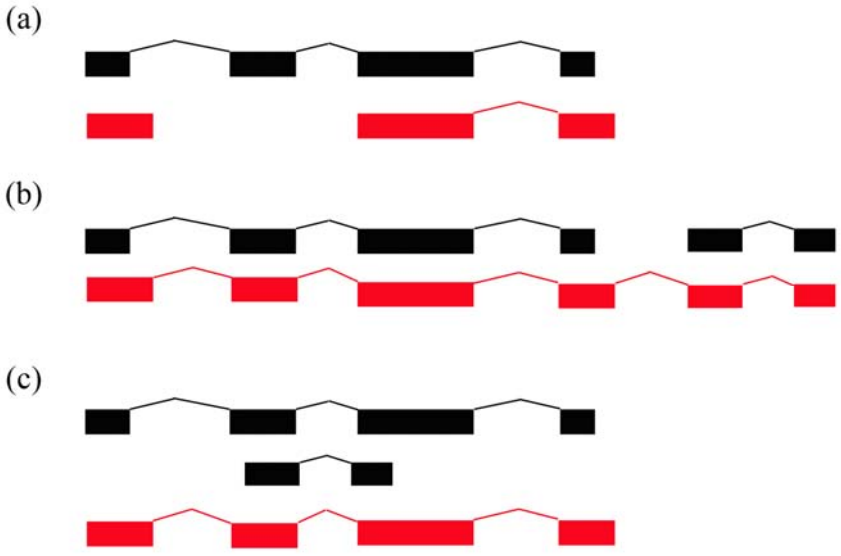

Figure 65. Pictorial representation of (a) split and (b) fused predictions in comparison with annotation. (c) Few annotated genes have internal genes in the same strand. Predictions matching these genes are ignored while calculating accuracy. Annotations are given in black while predictions are drawn in red.

I extracted annotations from the VEGA database (Ashurst, 2002) and found that chromosome 20 had 959 annotated genes (includes, Known, Novel CDS, Novel Transcripts, Pseudogene, Processed pseudogene, Unprocessed pseudogene and Putative categories). GenePred predicted 669 genes while GENSCAN made 1086 predictions after scanning the chromosome 20 sequence (Table 12). GenePred covered 592 genes (61.8%) of the total annotated genes while GENSCAN coverage was roughly 13% higher, identifying 722 genes (75.3%). Accuracy of GenePred and GENSCAN was found to be similar. GenePred made 230 correct predictions (34.4%) while GENSCAN predicted 369 (34.0%). However, GENSCAN made relatively higher number of split predictions (255 predictions). In contrast, GenePred made relatively more fused predictions (198 predictions compared to 149 by GENSCAN) and a smaller number of split predictions (97 predictions). However, GenePred had difficulty in identifying the annotated exon and intron boundaries when compared to GENSCAN (Table 13). Any prediction that overlaps an annotated exon is included in calculating coverage and accuracy. However, split and merge predictions are counted as false positives while deriving accuracy. Out of 6441 exons annotated by VEGA, GenePred predictions overlapped with 2869 (44.5%) while GENSCAN predicted 4132 (64.2%). GENSCAN's coverage is achieved from fewer predictions than GenePred and hence accuracy of GENSCAN (46.3%) is significantly higher than GenePred (12.7%). GenePred is not suited for predicting exact exon-intron boundaries (5512 annotated splice sites) as the donor and acceptor site coverage and accuracy is significantly less than GENSCAN (refer to Table 13). These results are expected as GenePred does not use any 'content' information like other *ab initio* gene prediction systems. Thus, GenePred is good for identifying gene blocks in the DNA sequences, which could be later annotated for exon-intron structure using other algorithms. The high number of fused predictions by GenePred indicates the potential to improve the model by tweaking the parameters and the feature models used to predict genes.

For the above comparison, I used GENSCAN predictions on repeat masked sequence since GENSCAN was known to perform better in masked than unmasked sequence. GENSCAN predictions on unmasked chromosome 20 (2108 predictions) shows significantly less accuracy (19.7%, compared to 34.0% reported earlier) although the coverage remains similar (masked: 75.3%, unmasked: 77.3%). This might be due to the difficulty of GENSCAN in ruling out coding regions in repeat sequences. However, such problems are

not observed with GenePred, as predictions on both masked and unmasked sequence showed similar coverage (masked: 60.8%, unmasked: 61.8%) and accuracy (masked: 36.6%, unmasked: 34.4%).

Table 12. Performance of GenePred and GENSCAN in predicting VEGA annotated genes.

	GenePred	GENSCAN
Total Predictions	669	1086
Fused Predictions	198	149
Split Predictions	97	255
Genes covered	592	722
Coverage	61.8%	75.3%
Accurate Predictions	230	369
Accuracy	34.4%	34.0%

Table 13. Performance of GenePred and GENSCAN in predicting VEGA annotated exons and splice sites.

	GenePred	GenePred *with* GeneSplicer	GENSCAN
Total Predictions	11145	44050	8465
Fused Predictions	853	243	27
Split Predictions	212	842	340
Exons covered	2869	3960	4132
Coverage	44.5%	61.5%	64.2%
Accurate Predictions	1418	3138	3921
Accuracy	12.7%	7.1%	46.3%
Donor site coverage	6.7%	19.0%	57.9%
Donor site accuracy	3.5%	2.4%	43.3%
Acceptor site coverage	4.0%	19.6%	57.8%
Acceptor site accuracy	2.1%	2.5%	43.2%

Out of total predictions from both GenePred and GENSCAN, nearly 40% of predictions (excluding, 34% correct predictions and approximately 26% fused/split predictions) are not correlated with VEGA annotations. A number of these may turn out to represent real transcripts missing from the existing annotation. As GAZE predictions are based on regulatory signals, some of the predictions that do not match the annotation are likely to be non-coding transcripts. Recent experiments by Affymetrix on chromosome 20 and 22 emphasise this fact (Cawley *et al.*, 2004). They found that a significant number of

transcription factor binding sites are correlated with non-coding RNAs and that they are regulated by a mechanism similar to that of protein coding genes. Thus, the excess predictions by GAZE are potential sequence blocks for hunting genes.

6.5 Tweaking GenePred gene prediction system

Having shown that the performance of GenePred is comparable with GENSCAN, I then tweaked Eponine models and configuration files used in making the GenePred prediction system to try to find improvements. I adopted three main approaches, which are explained below.

6.5.1 With Eponine translation models

With the availability of translation start and stop models, I decided to include them in the GenePred system in order to determine if this additional information might help in improving the performance. For this purpose, as explained earlier (Figure 61b), from the transcription start feature the model is allowed to transit to the translation start codon emitting the 5' UTR segment. Likewise, between the acceptor site of the last exon and the transcription termination site, the translation stop signal features are introduced.

I tested this modified gene prediction system with the annotation from chromosome 20 and found that there is no significant change in coverage and accuracy when compared to GenePred without Eponine translation models (Table 14). However, the number of genes predicted by the system increased (886 predictions compared to 669 predictions reported previously) and because of it the coverage increased by a small proportion (64.9%, 622 annotated genes were correctly identified) and accuracy decreased by a small proportion (32.0%, 284 predictions are accurate). As there is a trade-off between coverage and accuracy, the values are comparable with the GenePred system without translation models. However, adding translation models to GenePred created less fused (155 predictions) and more split predictions (170). Thus, Eponine splice sites bias the gene prediction system to extend the gene rather than terminate the extending prediction. This issue is addressed in case (iii) below.

Table 14. Performance of GenePred constructed with translation start and stop features.

Total Predictions	Fused Predictions	Split Predictions	Genes covered	Coverage	Accurate Predictions	Accuracy
886	155	170	622	64.9%	284	32.0%

Thus, adding translation models to the GenePred did not affect the performance in identifying annotated genes from the genomic DNA but modified the number of fused and split predictions.

6.5.2 Eponine Splice site predictions replaced with GeneSplicer predictions

In another attempt, I replaced Eponine splice site model predictions with GeneSplicer predictions while making GenePred. As explained in chapter 4, GeneSplicer performed better than Eponine splice site models by using more information from the DNA sequence. Since splice sites form the essential part in determining the gene structure by any gene predictor, I attempted GeneSplicer predictions with GenePred in predicting genes. Since GeneSplicer predictions are given in bit scores (x), they are first converted to log scores (z) using the expression given below before usage.

$$z = \frac{1}{1 + e^{-x}} \qquad (15)$$

GeneSplicer predictions with log scores are combined with both cases – with all Eponine models and with only Eponine transcription models (without translation models) – to derive a gene prediction system.

With the GeneSplicer features (along with transcription and translation features), the coverage (68.4%) and accuracy (35.6%) improved in comparison with GenePred using Eponine splice site features (Table 15). The increase in accuracy is due to the reduced number of predictions (778 predictions compared to 886) by the model. However, the number of fused predictions increased (196 compared to 155 predictions) when GeneSplicer splice site features are used. The results are similar, except that the number of predictions

increased (709 predictions compared to 669) when translation model predictions were not used along with GeneSplicer. Including GeneSplicer predictions, however significantly improved exon and splice sites coverage by GenePred (Exon: 61.5%, Donor: 19.0%, Acceptor: 19.6%). This improvement in coverage is due to the increase in the number of predictions (44050 predictions compared to 11145, refer to Table 13) and hence the accuracy decreased by a small proportion.

Table 15. Performance of GenePred constructed with and without translation features along with GeneSplicer features instead of Eponine splice sites.

	with translation features	*without* translation features
Total Predictions	778	709
Fused Predictions	196	214
Split Predictions	117	94
Genes covered	655	635
Coverage	68.4%	66.3%
Accurate Predictions	277	244
Accuracy	35.6%	34.4%

Thus, the increase in coverage using features of GeneSplicer features narrowed the margin between the GenePred and the GENSCAN while keeping the high accuracy of the GenePred system.

6.5.3 Scaled down Eponine feature scores

As noted earlier Eponine donor and acceptor site model features are screened for scores above 0.999 and 0.9998 respectively, for constructing GenePred using GAZE.

On evaluating different gene structures from the DNA sequence based on the given model, GAZE tries to balance between splice sites and transcription termination features in extending or terminating the gene. This might be compared to the *in vivo* competition between transcriptome and spliceosome in transcribing a gene. At least in two cases – IgM heavy chain genes and Calcitonin genes – the competing nature of splicing and transcription is shown experimentally. An internal weak poly(A) signal present within an intron of the IgM heavy chain gene under the low amount of CstF-64 transcription factor, misses the poly(A) signal and hence the transcription continues with the influence of the donor splice site present downstream. In cases where CstF-64 is available in relatively high concentrations, as in plasma cells, the transcriptome has the advantage and terminates the transcription (Takagaki and Manley, 1998; Takagaki *et al.*, 1996). Similarly in Calcitonin gene transcription, a weak internal poly(A) signal is used by the transcriptome, if the SRp 20 protein, a splice regulatory factor, fails to get recruited to the nearby splice sites (Zhao *et al.*, 1999).

A high number of fused gene predictions by GenePred might be due to the higher score of splice sites than transcription termination features predicted by the Eponine models. To test this hypothesis, here I attempt to scale down the values of splice site features. This is done by taking the inverse logit of the Eponine score and multiplying it with a scaling factor and reconverting back to the logit score. Inverse logit of the Eponine score was done using the formula –

$$x = \log\left(\frac{z}{1-z}\right) \qquad (16)$$

The inverse logit score (x) for donor and acceptor sites are scaled down by multiplying the values with 0.67 and 0.54 respectively. These values were found to be optimum after different runs and the scaled down scores are more equivalent to the transcription start and termination model scores (0.99). Likewise, the scores for translation stop model features

(0.999) are also scaled down by multiplying the inverse logit scores with a factor of 0.67. Before incorporating the donor and acceptor and translation stop features into GenePred the scores are converted back to logit values using equation 20 explained above.

Table 16 shows the GenePred system with the scaled down feature scores predicted more split predictions (208 and 151 predictions compared to 170 and 97 by GenePred with no scaled down features) and less fused predictions (144 and 165 predictions compared to 155 and 198 predictions by GenePred without scaled down scores). The scenario is similar for exons as well (657 split predictions compared to 212 predictions without scaled down scores). Overall the number of predictions also increased (997 and 824 predictions). Although there is a small increase in coverage (66.7% and 64.1%), it was compensated with a small decrease in accuracy (30.2% and 32.2%) and hence the coverage and accuracy are not significantly different from the above models. However, this tweak showed that the high number of fused predictions by GenePred is due to the splice site score values fed into the gene prediction system.

Table 16. Performance of GenePred system constructed with and without translation after scaling down splice site and translation stop scores.

	with translation features	*without* translation features
Total Predictions	997	824
Fused Predictions	144	165
Split Predictions	208	151
Genes covered	639	614
Coverage	66.7%	64.1%
Accurate Predictions	301	265
Accuracy	30.2%	32.2%

6.6 Revisiting transcription termination predictions

In chapter 3, I showed that the Eponine model works better than existing programs, ERPIN and Polyadq, in predicting transcription termination sites. However, the model made a huge number of false positive predictions and nearly 10% of them lie within the genes. Ruling out these false positive predictions within the gene will increase the accuracy of the model. This is possible by defining the exon-intron structure of a gene and removing any transcription termination predictions lying within exons or introns. The exon-intron structure can be defined using GenePred and thus might help to pin-point the false transcription termination model predictions.

To achieve this objective, I used the GenePred system developed by omitting Eponine translation models (included Eponine transcription start site, donor, acceptor and transcription termination models only) for this purpose. The system predicted genes by including only appropriate transcription termination sites after defining the exon-intron structure using the splice site features given. Transcription termination sites selected by GenePred are then dumped to find the coverage and accuracy of the model by comparing it with the VEGA annotated gene ends in chromosome 20 (Table 17). Out of 98 predictions matching the 213 annotated genes of chromosome 20, 24 predictions lie within 2500 bases from the annotated gene end showing an accuracy of 24.5% with coverage of 40.4%. For a comparable coverage the earlier analysis (refer to chapter 3) showed only 16.6% accuracy for the transcription termination model.

Table 17. Performance of transcription termination model with the support of GenePred prediction system.

Transcription Termination model	Coverage	Accuracy
with GenePred support	40.4%	24.5%
without GenePred support	40%	16.6%

Thus, by defining the exon-intron structure, some of the internal predictions of transcription termination can be removed giving the model better accuracy with no compromise on coverage.

6.7 Concluding remarks

In this chapter, I tried to build a gene prediction system by taking advantage of the sequence features predicted by Eponine models explained in previous chapters and GAZE, a dynamic programming based gene assembler. Various versions of the gene prediction system, GenePred, showed that the coverage and accuracy are comparable with GENSCAN. This is respectable given no protein information is used by GenePred unlike GENSCAN. However, GenePred should be treated as complementary to GENSCAN rather than a replacement, given the following facts: firstly the coverage of the union of predictions of GENSCAN and GenePred is higher than the coverage by the individual programs (Figure 66) and secondly the very poor performance of GenePred in predicting exon-intron structures compared to GENSCAN. Figure 66 shows that out of 959 VEGA annotated genes, 490 genes are predicted both by GenePred and GENSCAN. Twenty percent (102/592) of GenePred predictions and 32% (232/722) of GENSCAN predictions do not overlap with each other. This indicates that by using GenePred and GENSCAN together a better coverage of the annotation can be attained.

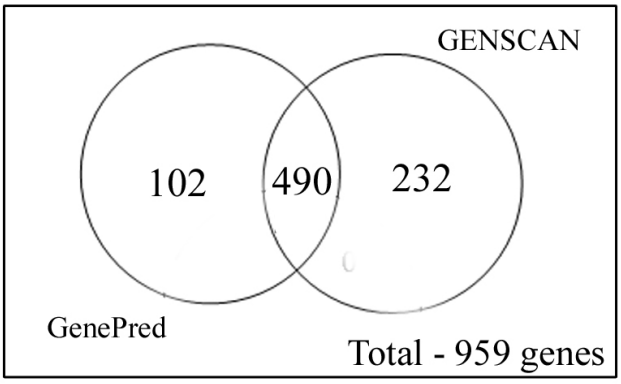

Figure 66. Venn diagram showing the coverage of GenePred and GENSCAN.

The accuracy of GENSCAN can also be improved by supplementing with the predictions of GenePred as indicated below. Table 18 and Table 19 show the accuracy of GENSCAN with

and without GenePred in predicting VEGA annotated genes and exons respectively. A GENSCAN scan on the GenePred predicted regions of chromosome 20 improved its accuracy compared to using it alone on the unmasked chromosome sequence. These results again emphasise that GenePred should be treated as a complement to GENSCAN.

Detailed analysis of the predictions of GenePred as a percentage of nucleotides covered reveal that 97.6% of nucleotides in chromosome 20 are annotated by GenePred (Table 20). This number is very high and significantly higher than the fraction of genome covered by GENSCAN (68.5%, 43654921 bases) or by VEGA annotations (28632433 bases, 44.9%) of chromosome 20.

Table 18. Performance of GENSCAN with and without GenePred in predicting VEGA annotated genes.

	GENSCAN *unmasked*	GenePred + GENSCAN
Total Predictions	2108	1224
Fused Predictions	95	55
Split Predictions	451	345
Genes covered	741	529
Coverage	77.3%	55.2%
Accurate Predictions	417	308
Accuracy	19.7%	25.2%

Table 19. Performance of GENSCAN with and without GenePred in predicting VEGA annotated exons.

	GENSCAN *unmasked*	GenePred + GENSCAN
Total Annotations	6414	6414
Total Predictions	12035	6763
Exons covered	4317	3021
Coverage	67.3%	47.1%
Accurate Predictions	4122	2913
Accuracy	34.2%	43.0%

Table 20. Nucleotide coverage by predictions of GenePred and GENSCAN.

	GenePred (Unmasked)	GENSCAN (Masked)
Total predictions	97.6% (62213556 bases)	68.5% (43654921 bases)
Correct predictions	30.9% (19709650 bases)	19.0% (12138177 bases)
Fused/Split predictions	35.8% (22787447 bases)	29.6% (18879962 bases)

These results indicate that GenePred's prediction accuracy comes mainly by determining the correct strand to transcribe, yet it is performing better than random: Random prediction

accuracy was evaluated by offsetting the predictions of GenePred and GENSCAN by 1, 2 and 3 mega bases (predictions exceeding the length of the chromosome are rotated round to the beginning) and recalculating the coverage and accuracy with respect to VEGA annotation (Table 21). GenePred predictions offset by 3 mega bases shows 42.6% coverage and 16.6% accuracy, which is significantly less than for the original predictions (coverage: 61.8%, accuracy: 34.4%). Similar results are found for GENSCAN predictions as well.

Table 21. Coverage and accuracy of GenePred and GENSCAN for predictions offset by 1, 2 and 3 mega bases.

	GenePred (Unmasked)	GENSCAN (Masked)
Predicitons	61.8% (cov) 34.4% (acc)	75.3% (cov) 34.0% (acc)
1 Mbp offset	49.4% (cov) 20.0% (acc)	49.7% (cov) 17.4% (acc)
2 Mbp offset	44.5% (cov) 17.6% (acc)	46.1% (cov) 14.6% (acc)
3 Mbp offset	42.6% (cov) 16.6% (acc)	45.0% (cov) 14.3% (acc)

Although GENSCAN coverage is better than GenePred overall, it is less likely than GenePred to predict VEGA ‘Novel_transcripts’ and ‘Putative’ genes. This may be partly due to GENSCAN’s reliance on protein information. Novel_transcripts are genes annotated from RNA that have weak evidence for being coding transcripts. Likewise, Putative genes are annotated using EST evidence and these genes also have no clear open reading frame. As the protein information content of this set of transcripts is less than for known genes, this may explain GENSCAN predicting few cases than GenePred (Table 22). For Novel_transcripts, all versions of GenePred discussed above show better coverage percentage with twice the accuracy of GENSCAN. Similarly for Putative genes, GenePred predicted at least 15% more genes with twice the accuracy of GENSCAN or more. On the combined dataset (Novel_transcripts + Putative genes), GenePred’s coverage was at least

10% more than GENSCAN with twice the accuracy or more. Table 22 details the coverage and accuracy of various versions of GenePred (with and without splice site and translation models) compared to GENSCAN. The low accuracy values are a consequence of considering predictions matching only Novel_transcripts and Putative genes as true and the rest as false predictions.

Table 22. Performance of GenePred and GENSCAN in identifying VEGA Novel_transcripts and Putative genes. Coverage and accuracy for each annotation is given for GenePred with and without translation models. Each of these GenePred systems is combined with either Eponine splice site or GeneSplicer features. Numbers in brackets shows the absolute values.

Annotation	GenePred + Eponine splice site		GenePred + GeneSplicer features		GENSCAN
	without translation features	*with* translation features	*without* translation features	*with* translation features	
Novel Transcripts	55.5 (50/90) 4.5 (37/824)	57.7 (52/90) 3.8 (38/997)	57.7 (52/90) 5.1 (36/709)	58.8 (53/90) 5.0 (39/778)	50.0 (47/90) 2.9 (33/1086)
Putative genes	48.4 (76/157) 7.0 (58/824)	50.9 (80/157) 6.2 (62/997)	57.9 (91/157) 8.9 (63/709)	60.5 (95/157) 8.5 (66/778)	35.0 (54/157) 4.3 (45/1086)
Novel Transcripts + Putative	51.0 (126/247) 10.6 (87/824)	53.4 (132/247) 9.0 (90/997)	57.9 (143/247) 12.6 (89/709)	59.9 (148/247) 12.2 (95/778)	40.5 (101/247) 6.5 (75/1086)

Thus, in this project I was able to develop a gene prediction system based purely on gene regulatory signals and show that its performance is encouraging considering that it does not rely on protein coding information. In terms of gene coverage it performs similarly for protein coding genes and better for genes with no coding evidence. At present the major problem is the very poor exon prediction accuracy despite including a splicing model. For easy comparison, in the table below (Table 23), I summarise the results of various versions of GenePred compared to GENSCAN in annotating chromosome 20 VEGA annotated genes.

Table 23. Summary of performance of various versions of GenePred and GENSCAN in identifying VEGA annotated genes in human chromosome 20

	GenePred + Eponine splice site		GenePred + GeneSplicer features		GENSCAN
	without translation features	*with* translation features	*with* translation features	*without* translation features	(Masked)
Total Predictions	669	886	778	709	1086
Fused Predictions	198	155	196	214	149
Split Predictions	97	170	117	94	255
Genes covered	592	622	655	635	722
Coverage	61.8%	64.9%	68.4%	66.3%	75.3%
Accurate Predictions	230	284	277	244	369
Accuracy	34.4%	32.0%	35.6%	34.4%	34.0%