Chapter 3

Identifying Splice Sites in Human Genomic DNA Sequences

Summary

The presence of conserved sequences at splice sites has been well documented over the last 25 years. However, these sequences are not sufficiently informative to permit unambiguous identification of gene structure. Gene prediction programs, such as GENSCAN (Burge and Karlin, 1997), combine splice site predictions with other information to predict complete gene structures. This chapter describes two novel models for the identification of canonical splice sites (sections 3.1 and 3.2) and one model, which applies standard methodology to identify the most frequent non-canonical splice site (section 3.3). The chapter concludes with a discussion of a human splice site predictor, Stratasplice, whch incorporates the best of these models and should prove useful for genome annotation. This analysis led to the observation that splice sites in GC-rich regions of the genome are slightly different from, and harder to predict than, splice sites in GC-poor regons.

3.1 A block dependence model for donor site identification

Introduction

Probabilistic signal recognition relies on the detection of differences between a training set of confirmed signals and a control set. Simple models, whch detect, for instance, only the order of individual nucleotides, require relatively small training sets, while more complex models, whch may consider overlapping pairs or groups of nucleotides, necessitate much larger sets of training data. Traditionally, a major stumbling block in the development of splice site detectors has been the shortage of reliable training data. However, the recent publication of SpliceDB (Burset *et al.,* 2001), whch contains more than 15,000 confirmed human splice site pairs has largely alleviated this concern.

Previous reports have suggested that, in addition to dependencies between neighbouring bases, the donor splice site contains longer range dependencies, perhaps relating to the binding of the U1 snRNA to the donor site (Burge and Karlin, 1997). This analysis attempts to quantify these longer-range interactions and take advantage of the information they provide to improve *ab initio* splice site identification.

Materials & Methods

Test sets

Training and evaluation sets were generated from the 15,263 confirmed canonical human splice site pairs in SpliceDB (Burset *et al.*, 2001). 786 donor sites and 1,295 acceptor sites with poor or incomplete sequence data were removed from this set, yielding a total of 14,477 confirmed 5‘ splice sites and 13,968 confirmed 3’ splice sites. A control set of genomic DNA used to calculate null model frequencies was extracted from the first 10 kb of repeat-masked DNA chosen from 100 randomly selected Ensembl clones (International Human Genome Sequencing Consortium, 2001). Sets of “false” splice sites were generated by extracting sequences around GT, GC or AG dinucleotides in this random set of genomic DNA. (Some small fraction of these sites will in fact be true).

Independence and first-order dependence models

Two classic pattern recognition techniques, independent weight matrices (Staden, 1984) and first-order dependent weight matrices (Zhang and Marr, 1993) were re-implemented for comparative purposes. These two models yield log-likelihood scores for each potential splice site by comparing the frequency of either individual nucleotides (independent model) or dinucleotides (first-order model) at each position in the splice site window with background genomic frequencies. Given a sequence $X = \{\mathrm{x}, ,x_2,...,x_n\}$, scores were derived from each model as follows:

$S(X) = \sum_i \log_2 \frac{f'_{x_i}}{q_{x_i}}$	Independence Model
$S(X) = \sum_i \log_2 \frac{f^i_{x_i\vert x_{i-1}}}{q_{x_i\vert x_{i-1}}}$	First-order dependence Model

where f'_{x_i} is the frequency of base x_i at position i in the training set, $f'_{x_i|x_{i-1}}$ is the frequency of base $\mathbf{x,}$ at position i following base x_i , at position $i-1$ and q_{x_i} and $q_{x_i|x_{i-1}}$ are genomic nucleotide and dinucleotide frequencies, respectively.

Detection rates

Detection rate curves (see Figure 3.3, for example), whch illustrate a model’s

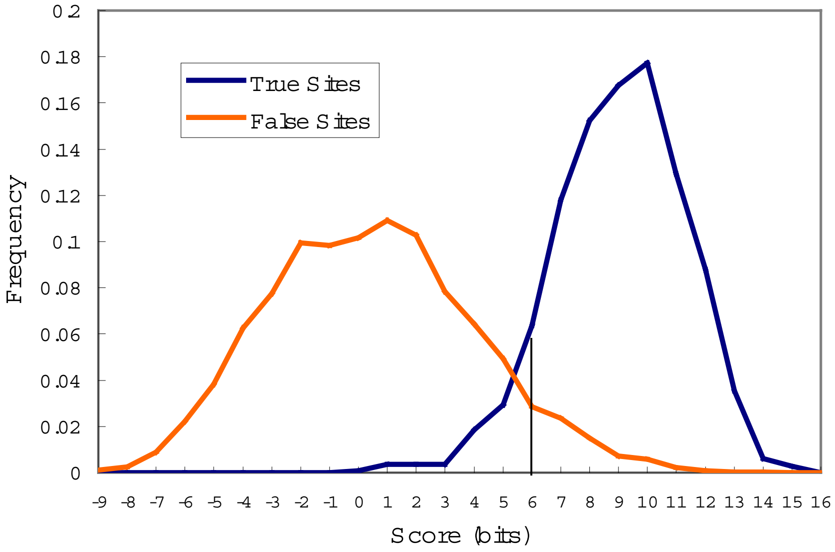


Figure 3.1 - Typical log-odds score distributions for evaluation and control sets. The vertical line, drawn here at an arbitrary threshold of 6 bits, divides the true distribution (blue line) into true positives and false negatives and the false distribution (red line) into true negatives and false positives. Distributions such as this one were used to generate the detection rate curves as described in Materials and Methods.

performance at a variety of threshold values, were used to compare the performance of the various models. The fraction of true sites included and the false positive rate are calculated directly from the evaluation and false sets, respectively. Using 6 bits as the threshold value (illustrated by the vertical bar in Figure 3.1), a point (*x*,*y*) on the detection rate curve would be calculated as follows:

$$x = C(false > 6) * (10{,}000 / G)$$
$$y = C(true > 6) / C(true)$$

where C(condition) represents a simple conditional count and *G* is the amount of genomic DNA from which the relevant set of false sites was extracted.

Mutual information analysis

To identify dependencies between non-neighbouring bases, the mutual information was calculated between all pairwise combinations of bases in donor and acceptor splice sites using the SpliceDB dataset (Burset *et al*, 2000). The mutual information

$$M(i,j) = \sum_{a,b} f(x_i = a, x_j = b) \log_2 \frac{f(x_i = a, x_j = b)}{f(x_i = a) f(x_j = b)}$$

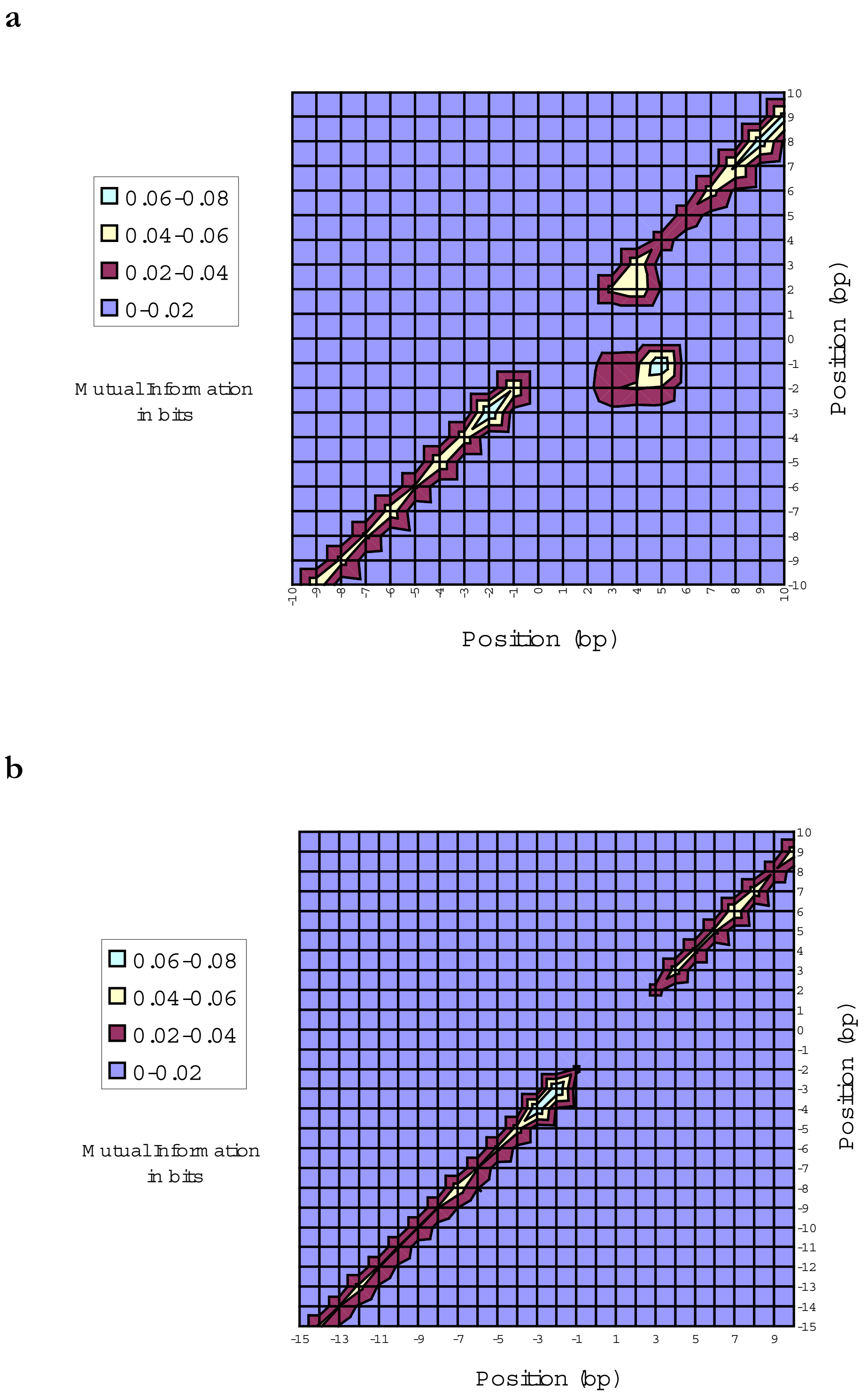
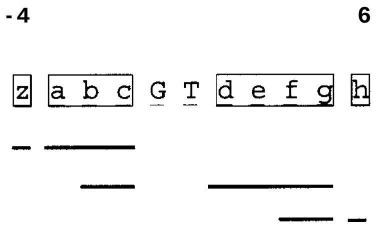


Figure 3.2 - Mutual information around donor (a) and acceptor (b) splice sites. Positions in the donor site window [-10, +10] and acceptor site window [-15, +10] are shown along both axes. The canonical G is at position 0 in (a) and position -1 in (b). The diagonal line present in both (a) and (b) represents the mutual information between neighbouring pairs of bases.

between two positions yields a value in bits indicating the degree of dependence between positions i and j (Durbin *et al.,* 1998).

Score calculations

Based on the mutual information results (see Figure 3.2a) a model was derived whtch divided the region around the donor splice site into blocks as shown below:



Thts model was scored using log-likelihood scoring considering the conditional probabhties of the blocks above (dependencies indicated by the horizontal black lines) and using genomic dinucleotide frequencies for the null model. 'Thus, the score of a sequence X in bits is

$$S(X) = \log_2 \frac{f(abc \mid z) * f(defg \mid bc) * f(h \mid fg)}{q(z)q(a \mid z)...q(c \mid b) * q(d)q(e \mid d)...q(h \mid g)}$$

Frequency values for each possible base combination of each block given its dependencies in the model were calculated by adding pseudocounts based on genomic dinucleotide frequencies to the observed counts. 'Thus, for example,

$$f(abc \mid z) = \frac{C(x_{-4} = z, x_{-3} = a, x_{-2} = b, x_{-1} = c) + 4^3 q(a \mid z)q(b \mid a)q(c \mid b)}{C(x_{-4} = z) + 4^3}$$

Results

Mutual information analysis (see Figure 3.2a) revealed a fair amount of information (> 0.3 bits) between non-neighbouring bases in donor splice sites and a novel block dependence model of donor splice sites was developed to take advantage of this information. 'This model showed moderate improvement over first-order dependence and independent weight matrix models for prediction of canonical donor splice sites (see Figure 3.3).

Mutual information analysis was also performed on the acceptor splice site dataset, but no significant information was found between non-neighbouring bases (see Figure 3.2b).

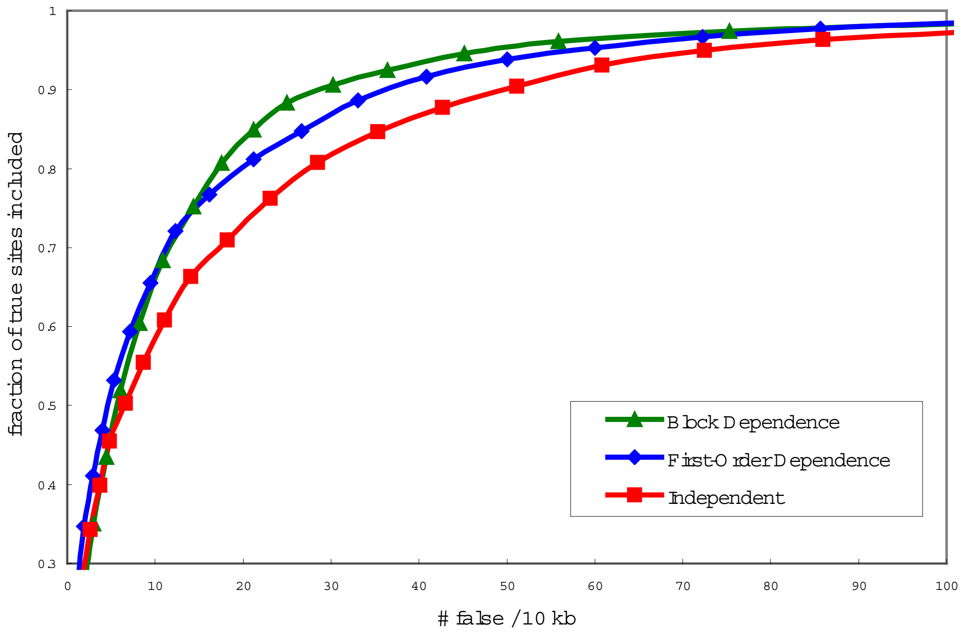


Figure 3.3 - Comparison between independent, first-order dependence and block dependence models for donor splice site identification. The block dependence model predicted fewer false positives at most sensitivity levels than did the other two models.

Discussion

The block dependence model described here shows that splice site signal recognition can be improved by considering higher order and long-range interactions. However, this improvement is quite modest when compared to the first order dependence model. Other splice site identification models including maximal dependence decomposition, the model used by the GENSCAN gene prediction program (Burge and Karlin, 1997), also show modest improvements over a first-order dependence model and the block dependence model presented here is unlikely to represent a substantial improvement over these models.

3.2 Improved splice site identification by considering local GC content

Introduction

Analysis of the draft sequence of the human genome has confirmed the observation that large regions of the genome deviate from the genome-wide average GC level of 41 percent (International Human Genome Sequencing Consortium, 2001). Huge regions (>10 Mb) were observed which differed significantly from the average, while smaller regions (20 kb) showed more variation with GC content levels ranging from 30 to 65 percent. 'These results suggest that traditional probabilistic signal recognition techniques, such as those used to identify RNA splice sites, whch identify differences between a positive model and a genomic null model, are likely to suffer from substandard performance in regions where the actual genome sequence differs greatly from the genome average. Conversely, more accurate modelling of background DNA composition should allow for more accurate discrimination of true splice signals.

The most successful gene prediction programs, such as GENSCAN (Burge and Karlin, 1997) fit different coding models and length distributions for regions of different GC content, but I am not aware of previous stratification of splice site models. I describe here an approach to splice site identification, whch extends the first-order dependence weight matrix technique (Zhang and Marr, 1993) by stratifying the prediction process according to local GC content. This yields significantly improved performance, particularly in GC-rich and, thus, gene-rich areas (Zoubak *et al.,* 1996).

Materials & Methods

Test sets were derived and detection rate curves were generated as described in section 3.1.

GC stratification

Canonical donor (GT) and acceptor (AG) splice sites were stratified by local GC content according to the base composition in the total surrounding sequence (generally 80 bases, excluding 8 bases immediately around the splice junction) included in SpliceDB. The control set was stratified accordmg to the GC content in 300 base chunks. During sequence scans (and during derivation of the false set from the genomic set) potential splice sites were stratified according the base composition in the 75 bases preceding and following an eight-base window centred on the splice site itself.

Splice site windows

I chose splice site windows that included all positions significantly deviating from random background frequencies on the basis of relative entropy calculations (Durbin *et al.,* 1998) and were expanded to convenient sizes.

$$Entropy(i) = f(x_i)\log_2 \frac{f(x_i)}{q(x_i)}$$

This yielded windows from –10 to +10 around GT donor sites (canonical G at position 0) and –25 to +5 around AG acceptor sites (canonical A at position 0).

Score calculations

Log-likelihood scoring was used to generate a log-odds score for each potential splice site (Durbin *et al.,* 1998). Conditional frequency values for each dinucleotide pair at each position in the splice site window $(f^{i}_{a|b})$ were determined by adding pseudocounts to the observed values as follows:

$$f^{i}_{a|b} = \frac{C^{i}_{a|b} + 4q_{a|b}}{\sum_{b} C^{i}_{a|b} + 4}$$

where $C^{i}_{a|b}$ is the observed count of base *a* occurring at position i following base *b* at position i· I and $q_{a|b}$ is the observed conditional frequency for the appropriate control set. Observed conditional frequencies in the appropriate stratified control set were used for the null model. One model was trained (*e.g.* calculation of both *f* and *q* values) for each stratum of each splice signal. Score values in bits for a sequence $X = \{\mathrm{X}_r, x_2,\ldots,x_n\}$ were derived from the appropriate frequency data as follows:

$$S(X) = \sum_{i} \log_2 \frac{f^{i}_{x_i|x_{i-1}}}{q_{x_i|x_{i-1}}} .$$

Prior probability estimation

'The prior probability that a given GT or AG dinucleotide defined a true splice site was calculated using estimates of the total number of G T dinucleotides, AG dinucleotides and introns in the genome. The estimates of the total number of each dinucleotide were generated by counting dinucleotides on one strand of 2 MB of random genomic sequences (10 kb chunks from 200 randomly selected clones) and scaling this value to fit the 3000 MB genome. An estimate of 400,000 introns in the genome was generated by considering a genome consisting of 40,000 genes where each gene had an average of 10 introns. Limiting the analysis to one strand and scaling this number by the overall frequencies of each of the various types of splices sites (e.g. 99.24% GT-AG, 0.69% GC-AG, etc) reported in SpliceDB (Burset *et al.,* 2000) allowed the calculation of "per strand estimates" for each splice site type. Dividing by the corresponding total number of the relevant dinucleotide estimated per strand yielded the final priors $(P(T)$, see 'Table 3.1). As gene densities vary with GC content (Zoubak *et al.*, 1996), the prior probabilities for GT and AG dinucleotides were scaled accordmg to the frequencies of true $(\mathit{ff}(T\mid \mathrm{GC}))$ and false splice sites $(f(F\mid \mathrm{GC}))$ at each GC level as follows:

$P(T\mid \text{GC}) = \frac{f(T\mid GC)}{f(F\mid GC)}P(T)$. The necessary GC-level dependent frequency values were derived from the stratification of the true and false splice site sets (see Figure **3.4).**

	GT Donor Sites		AG Acceptor Sites	
Prior		1.37e-3		9.94e-4
Posterior Threshold	Sensitivity	Specificity	Sensitivity	Specificity
-∞	100	0.1	100	0.1
1e-6	99.8	0.3	99.8	0.3
le-5	99.6	0.4	99.7	0.4
le-4	99.1	0.7	99.0	0.6
lc-3	96.6	1.5	96.2	1.1
1e-2	84.8	3.5	70.4	3.6
5e-2	58.4	7.4	16.3	10.1
le-1	41.5	10.5	0.0	**N/A**

Table 3.1 - Performance of the stratified splice model on genomic sequences. Prior probabilities and sensitivity and specificity values of the stratified splice model at various posterior probability threshold values are indicated. Sensitivity and specificity values are provided as percentages and are calculated assuming an intron density of 67/MB (see Materials and Methods).

Posterior probability calculations

Posterior probability values, which incorporate prior biological information into a statistical framework (Durbin *et al.*, 1998), were used to generate probability values that combined the log-odds scores and the estimated prior probabhties for each potential splice site. Bayes' theorem was used:

$$P(T\mid S(X)=\mathrm{s}) = \frac{P(S(X)=s\mid T)P(T)}{P(S(X)=s\mid T)P(T)+P(S(X)=s\mid F)(1-P(T))}$$

where $P(S(X)=s\mid T)$ reads the probability that the score of sequence X is s, given the knowledge that the sequence is a true splice site and $P(T)$ is the scaled prior probability described above.

The conditional probability values used in the posterior calculation were calculated from the strata-specific distributions of true and false splice sites (see Figure 3.1, for example), assuming that these distributions were Gaussian. In brief, the mean and standard deviation were estimated for each distribution using standard formulas and the conditional frequency values were taken from the hypothetical Gaussian distribution that these two values defined. This approach led to more robust estimation of values in the tails of the distribution than simply using the observed values due to the small number of data points in the tails.

Model evaluation

'The choice of GC-level boundaries for the stratification process was evaluated by a modified version of the equivalence number statistic (EN), whch summarises the selectively and specificity of a given model by comparing the number of false positives and true negatives (Pearson, 1995). In this situation, I use probability distributions rather than raw numbers, and define the equivalence number as the frequency of false positives when the log-odds bit threshold is set to equalise the frequency of false positives and true negatives. **As** my model seeks to minimise both false positives and true negatives, the lower the EN value, the better the model. In order to take into account the effects of stratifying the prediction process, the final EN value was a weighted average of the EN values of each individual model. Weighting was done accordmg to the frequency of true splice sites within each stratum. Given a stratified splice prediction model with *n* strata (e.g. $M = \{m_1, m_2, ...m, \}$)the final EN value would be

$$EN(M) = \sum_{I} EN(m_i) * f(T \mid m_i).$$

In order to maximise use of the available data, I used a jack-knife procedure in whch the avadable data was divided into four sets. Pour training and evaluation cycles were performed holding out each set for evaluation in turn and using the other three sets for training. The results of these four cycles were averaged to produce the final value.

Sensitivity and specificity

Sensitivity and specificity values were determined using the posterior values generated when the model was trained and evaluated using disjoint subsets of the set of all true sites in SpliceDB and on all false sites in the genomic control set. **A** jack-knife procedure identical to the one described above was used and final values are the average of four different training and evaluation cycles. Sensitivity was calculated as the ratio of true positives to all true sites. Specificity calculations depended on an estimate of the density of introns in the genome. 'Two values were used: 67 introns/MB (consistent with the intron density estimates for the prior probability calculations) and 563 introns/MB (the intron density of GENSCAN's evaluation set). Using these estimates I scaled the total number of observed true positives to the expected number in a set the size of the control set and then calculated the specificity as the ratio of true positives to true positives plus false positives.

Results

Previous studies have indicated that GC-rich regions of the human genome are also gene-rich (Zoubak *et al.,* 1996). This is reflected in the distribution of GC content levels near intron splice sites (see Figure 3.4). As expected the GC distribution around GT dinucleotides in a random genomic sample (my control set) was approximately normally distributed with a peak near 40 percent GC. Only 10 percent of background GT dinucleotides are in areas of 60 percent GC or greater. In contrast, 27 percent of true donor splice sites are located in areas of 60 percent GC or greater. Similar results were seen for AG acceptor sites (data not shown).

To determine whether splice site signals had the same composition across the full range of GC content levels, all true splice sites from SpliceDB (Burset *et al.,* 2001) were divided into 3 groups based on the surrounding sequences (excluding 8 bp around the actual splice junction) and simple frequency tables were derived around the splice sites (see 'Tables 3.2a,b). Interestingly the donor site signal is largely conserved across all GC levels except for the thud base in the intron, whch changes from 71 percent **A** and 23 percent C in the low GC content group to 33 percent A and 62 percent G in the high GC content group. **A** similar though less dramatic change involving C and T nucleotides is seen for the thud base of the intron (just before the AG) at acceptor sites as well. The polypyrimidine (C|T) tract found upstream of acceptor splice sites is biased toward C in high GC content regions and toward T in low GC content regions (data not shown).

To explore whether splice site identification could be facilitated by considering local GC content, I developed a splice site identification model based on the first-order dependence weight matrix approach (Zhang and Marr, 1993), whch stratifies both the training data and the null model data according to local GC content. Figures 3.5a and 3.6a use detection rate curves (as described in section 3.1) to compare the performance of two standard weight matrix models and the new stratified model. Strikingly, at GT donor sites, the new stratified model outperformed the non-stratified first-order dependence model as least as dramatically as this model outperformed the independent weight matrix model. Less dramatic, but still useful, improvements were seen for the acceptor site model.

To explore the reasons behind these improvements, I compared the performance of the stratified and the non-stratified first order model on stratified test sets (see Figures 3.5b, 3.6b). These graphs indicate the relative performance of each predictor on splice

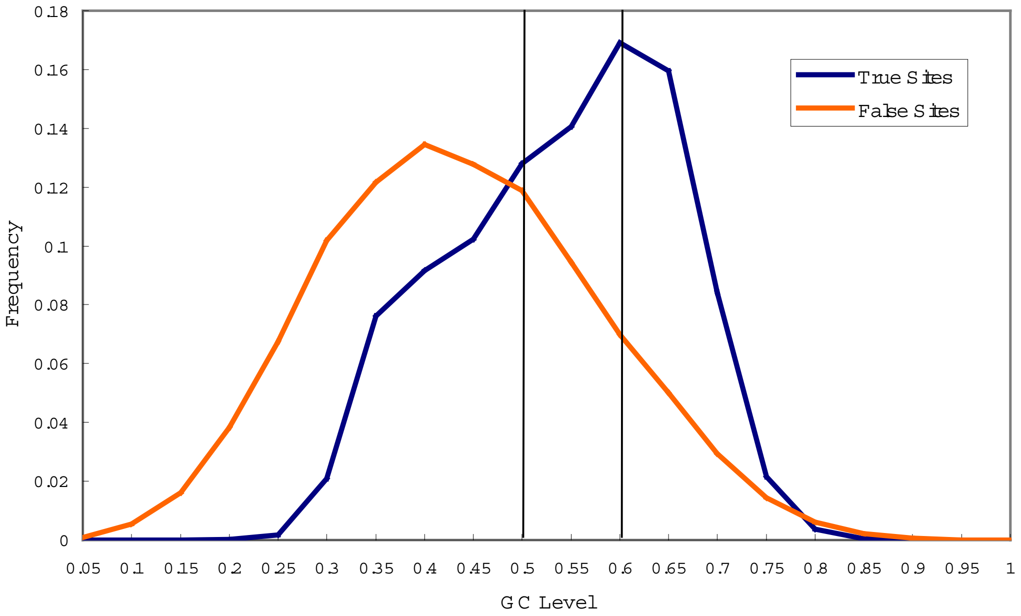


Figure 3.4 - Local GC content near true and false splice sites. The red line indicates the frequency distribution of local GC content near GT sites in the control set, while the blue line indicates this same distribution for true GT donor sites in SpliceDB. The black vertical bars indicate the final stratification boundaries used in the model.

sites in each GC stratum. Interestingly, for the non-stratified donor site predictor, I found significant differences in my ability to accurately identify splice sites according to the stratum. In particular, splice site prediction was easiest in low GC content regions, slightly harder in medium GC content regions, and significantly harder in high GC content regions (Figure 3.5b, compare dotted lines). Using the stratified predictor, I observed only minor improvements in the low and medium strata but found a striking improvement in performance in the high GC stratum.

Breaking down the performance by strata at acceptor sites led to slightly different results (see Figure 3.6b). For the non-stratified model, both the low and medium strata showed similar performance profiles, while the high GC stratum showed significantly worse performance (Figure 3.6b, compare dotted lines). The stratified model yielded improvements across all three strata with the high GC stratum showing the most dramatic improvement. However, improvement for this stratum was not as dramatic as it was for the donor site predictor.

To further quantify the performance of this new splice site identification model, sensitivity and specificity values were calculated at a variety of posterior probability thresholds. Results are shown for two intron density levels. Table 3.1, which uses 67

a

	Low GC Content							
A	0.64	0.11	0.00	0.00	0.71	0.72	0.10	0.20
C	0.10	0.03	0.00	0.00	0.02	0.05	0.05	0.11
G	0.10	0.78	**1.00**	0.00	0.23	0.09	0.76	0.16
T	0.16	0.08	0.00	**1.00**	0.04	0.14	0.09	0.54
	Medium GC Content							
A	0.61	0.08	0.00	0.00	0.44	0.71	0.06	0.15
C	0.13	0.04	0.00	0.00	0.04	0.08	0.06	0.18
G	0.13	0.81	**1.00**	0.00	0.50	0.13	0.84	0.26
T	0.13	0.07	0.00	**1.00**	0.02	0.08	0.04	0.41
	High GC Content							
A	0.55	0.08	0.00	0.00	0.33	0.69	0.03	0.10
C	0.18	0.03	0.00	0.00	0.03	0.11	0.06	0.24
G	0.15	0.83	**1.00**	0.00	0.62	0.15	0.88	0.26
T	0.12	0.06	0.00	**1.00**	0.01	0.05	0.03	0.39
	A	G	**G**	**T**	A|G	A	G	T

b

	Low GC Content						
A	0.10	0.27	0.07	**1.00**	0.00	0.27	0.26
C	0.22	0.22	0.56	0.00	0.00	0.12	0.16
G	0.05	0.17	0.00	0.00	**1.00**	0.50	0.18
T	0.62	0.34	0.37	0.00	0.00	0.11	0.40
	Medium GC Content						
A	0.07	0.22	0.03	**1.00**	0.00	0.22	0.21
C	0.41	0.36	0.78	0.00	0.00	0.15	0.21
G	0.06	0.22	0.00	0.00	**1.00**	0.52	0.22
T	0.46	0.20	0.19	0.00	0.00	0.11	0.35
	High GC Content						
A	0.04	0.17	0.02	**1.00**	0.00	0.21	0.18
C	0.49	0.39	0.87	0.00	0.00	0.14	0.24
G	0.09	0.30	0.00	0.00	**1.00**	0.57	0.28
T	0.38	0.14	0.11	0.00	0.00	0.08	0.30
	C|T		C|T	**A**	**G**	G	

Table 3.2 –Nucleotide frequencies at stratified donor (a) and acceptor (b) splice sites. Splice sites from SpliceDB (Burset *et al.*, 2001) were divided into three groups, less than 50 percent GC, 50-60 percent GC and greater than 60 percent GC, based on local GC content in an 80 bp window around the splice site, and the frequencies of bases at positions surrounding the splice junction are shown. The splice site consensus sequence is shown in the bottom row. Bold text indicates the canonical dinucleotide. Red text indicates bases that show large changes in frequency between strata.

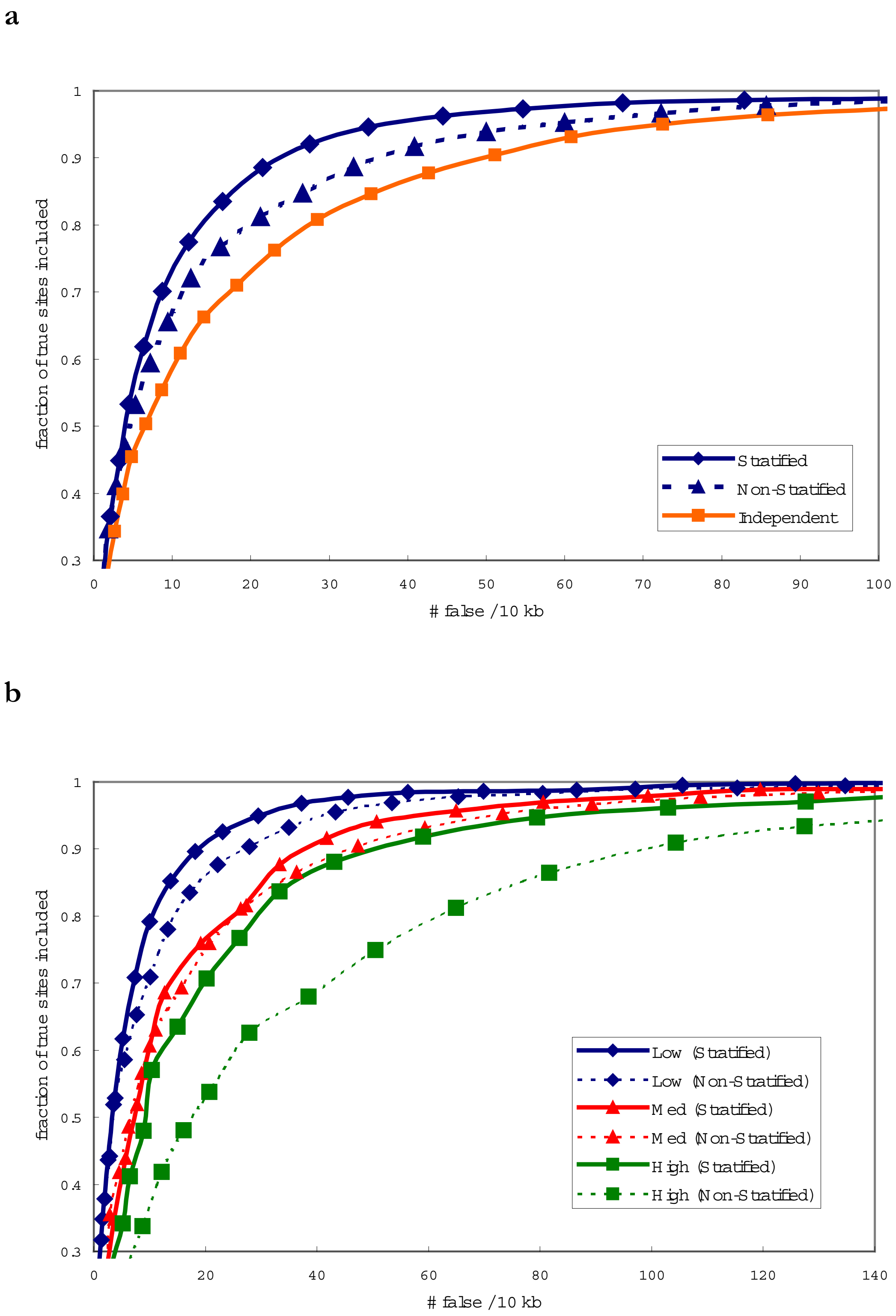


Figure 3.5 - Performance of the stratified splice predictor at GT donor sites. (a) Performance comparison with first-order dependence model (non-stratified) and independent model. (b) Performance comparison with first-order dependence model (non-stratified) on stratified test sets.

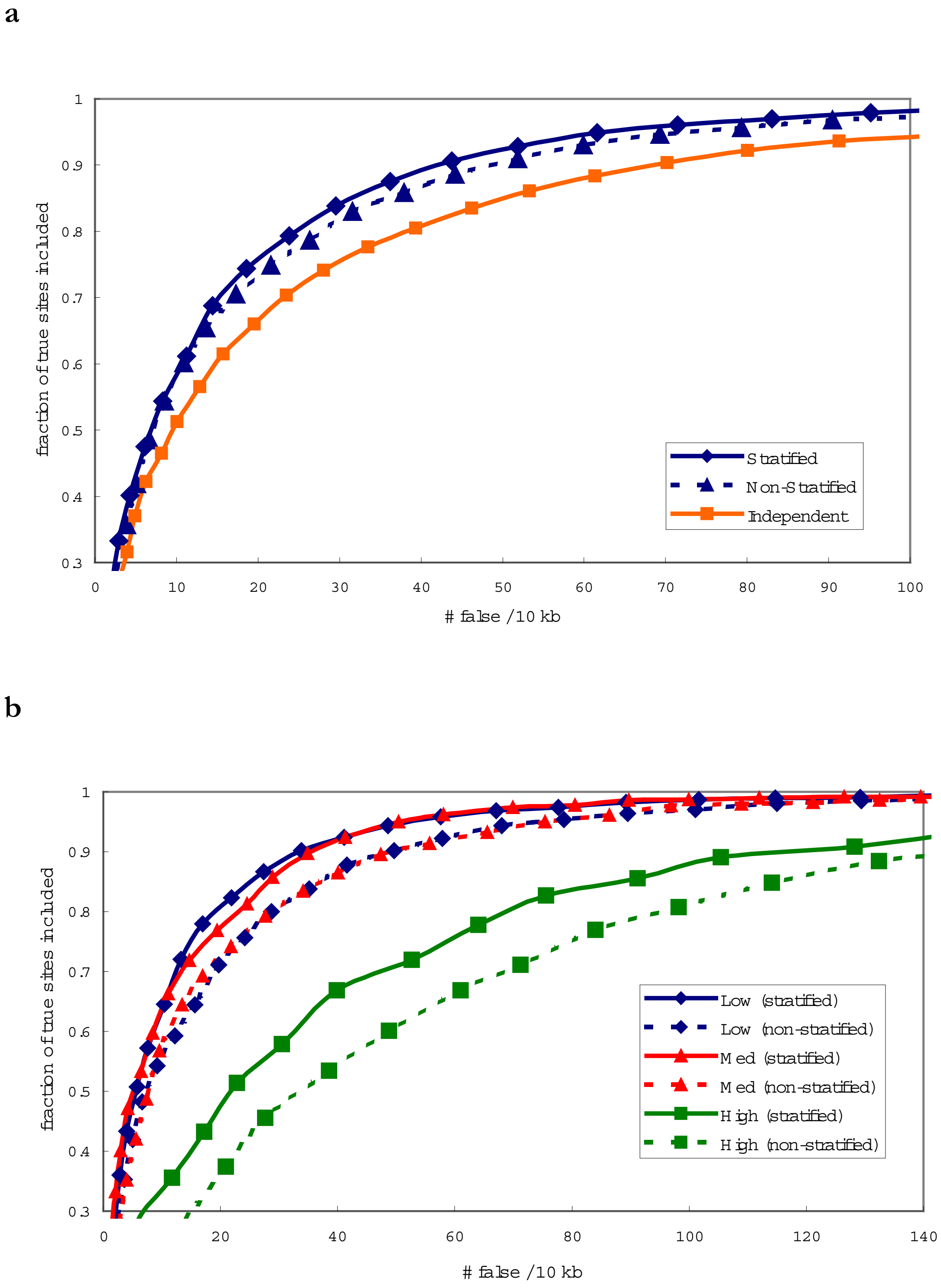


Figure 3.6 – Performance of the stratified splice predictor at AG acceptor sites. (A) Performance comparison with first-order dependence model (non-stratified) and independent model. (B) Performance comparison with first-order dependence model (non-stratified) on stratified test sets.

introns/MB, represents the performance expected on typical genomic DNA sequences whde Table 3.3 uses the much higher density of 563 introns/MB for comparative purposes with the GENSCAN splice site predictors. As Table 3.3 indicates, the stratified splice predictor described here generally gives higher specificity values for a given sensitivity level at both donor and acceptor splice sites than the GENSCRN splice site predictors.

	GT Donor Sites		AG Acceptor Sites	
Sensitivity	Genscan	Stratified	Genscan	Stratified
99	N/A	5.9	N/A	4.6
95	8.7	13.1	5.6	9.5
90	13.4	18.4	8.8	13.7
50	**36.0**	**43.4**	33.8	30.4

Table 3.3 – Comparison between GENSCAN's splice site predictors and my new stratified splice model. The table indicates specificity values as percentages at the indicated sensitivity level. GENSCAN values are taken from (Burge, 1998). Stratified splice model values are calculated assuming the same intron density (563 introns / Mb) as GENSCAN's evaluation set (Burge, 1997). Specificity values for GENSCAN at 99 percent sensitivity have not been published.

In theory, the stratification process can divide the data into any number of strata, but in practice limited data means only four or five models can be reliably trained. A variety of different strata boundaries were explored and evaluated using a modified version of the equivalence number metric (described in Materials and Methods), whch indicates the frequency of false positives when a threshold is selected to balance the frequency of false positives and true negatives. Switching from a non-stratified model to a three-stratum (< 50 percent, 50-60 percent, > 60 percent) model decreased the equivalence number from 12.7 percent to 10.3 percent for the GT model and from 12.0 percent to 10.7 percent for the AG model. Similar results were seen for other three-stratum models and for models with four or five strata. Thus, a three-stratum model with the boundaries at 50 and 60 percent GC was selected.

Discussion

'The simple approach presented here of stratifying the data and training a first-order model for each stratum outperforms higher order models, such as those used in GENSCAN and section 3.1 and shows that stratification by local GC content levels is a powerful technique for improving genomic signal recognition. Although some differences were observed among the consensus sequences of splice sites after stratifying by GC content, much of the improvement seems to be due to the improved null model, whch was generated by stratifying the control set. This observation suggests that similar stratification approaches may yield significant improvements in other signal detection problems, such as promoter motif detection.

The observation that splice site identification is more difficult in GC-rich regions of the genome than it is in GC-poor regions is quite intriguing, particularly considering the correlation between GC content and intron length seen in the human genome (International Human Genome Consortium, 2001). Although my results were derived *in silico*, they seem likely to indicate a biological reality; splice site consensus signals provide less information in GC-rich regions of the genome than they do in others. These observations lead to the enticing hypothesis that splice site recognition by the spliceosome may be a significant constraint on intron evolution, particular in GC-rich regions. Short introns are not associated with GC-rich regions in all vertebrate genomes (Hurst *et al.,* 1999), however, and it would be interesting to consider other hgher organisms to see if these support the observed association.

It is worth observing as well that the problem under consideration here, namely the identification of RNA splice sites from genomic sequence is rather more difficult than that whch the cell performs *in vivo.* Whereas I must attempt to identify splices sites from raw genomic sequence, the cell must only accurately identify splice sites on pre-mRNA. If roughly a quarter of the genome is transcribed (Venter *et al.,* 2001), the splicing machinery in the cell has a search space reduced 8 fold in size (the extra factor of *two* comes because mRNA is single stranded). This partially explains the low specificity scores seen for the genomic analysis (see Table 3.1) and emphasises the importance of considering as much evidence as possible when predicting genes. Systems such as GAZE (Howe and Durbin, unpublished) whch can integrate splice site predictions from one source with promoter predictions from another as well as homology and comparative information seem likely to be the way forward in automated gene prediction.

3.3 – Identifying non-canonical GC donor sites

Introduction

Recent analyses have indicated that roughly 0.7 percent of human introns start with the non-canonical dinucleotide GC in place of the much more common GT (International Human Genome Sequencing Consortium, 2001; Burset *et al.,* 2001). However most gene prediction packages do not consider GC as a potential donor site and miss several thousand introns for this reason. Additionally, automated analysis pipelines, such as the Ensembl project, whch are becoming increasingly important gateways to the human genome sequence, have few of these non-canonical, but still relatively common, introns annotated correctly (M. Clamp, personal communication). Thts chapter describes the development and performance of a simple GC donor site model, whtch should prove useful for genome annotation.

Materials & Methods

Test sets

A training set of 122 true GC donor sites was derived from the set of 270 EST-confirmed and verified non-canonical introns included in SpliceDB (Burset *et al.,* 2001). Control and false sets were generated as described in section 3.1.

First-order dependence model

A first order dependence weight matrix splice site predictor (Zhang and Marr, 1993) was implemented as described in section 3.1 and trained using the training set of GC donor sites.

Prior and posterior probabilities

A prior probability for GC dinucleotides was derived as in section 3.2 except that no corrections were made for local GC content. Posterior probabhties were calculated as in section 3.2.

Results

A first-order dependence weight matrix was built from the training set and used to score sets of true and false GC donor sites. Although GC donor sites are roughly 100-fold less common than GT donor sites, and the prior probability is therefore roughly 100-fold less, performance (see Table 3.4) is only marginally worse at GC sites when compared to GT sites. The GC model is difficult to evaluate accurately due to limited data, but specificity of GC donor predlctions tends to be roughly 15-25 fold worse than for GT donor predictions at a gwen sensitivity level. For instance at a threshold that includes roughly 96 percent of all true sites, 150 out of 10,000 predicted GT donor sites will be true whde 7 out of 10,000 predicted GC donor sites would be true.

Although the GT and GC donor consensus sequences are similar, the GC donor consensus is more highly conserved and contains nearly 13 bits of information, as opposed to approximately 8 bits of information in the GT donor consensus. For this

	GT Donor Sites		GC Donor Sites	
Prior		1.37e-3		1.12e-5
Posterior Threshold	Sensitivity	Specificity	Sensitivity	Specificity
-∞	100	0.1	100	0.001
le-6	99.8	0.3	100	0.008
le-5	99.6	0.4	100	0.02
le-4	99.1	0.7	95.9	0.07
le-3	96.6	1.5	76.2	0.3
1e-2	84.8	3.5	48.4	1.7
5e-2	58.4	7.4	27.9	6.9
le-1	41.5	10.5	17.2	76.8

reason, the GC model developed here should be expected to significantly outperform the simple approach of identifying GC donor sites by simply replacing the 'T' with a C in a standard GT donor site model.

Discussion

The model described here for identifying GC donor sites is interesting not because it is novel or complex, but because it is immediately useful. Many of the 2000 or so genes with GC introns may have been incorrectly annotated during the early stages of automated genome analysis. Yet, if the goal of delineating the full collection of human genes is to be achieved, these genes, whch contain non-canonical introns, must be included.

Although the number of false positives is high for GC donor site identification, this result is not unexpected, nor is it a major concern. Many gene prediction systems are tailored to work with a large set of predlctions and can combine a variety of types of evidence to separate true and false signals.

3.4 – Stratasplice: A human splice site predictor

Introduction

Stratasplice is a stand-alone splice site predictor designed for use on human genomic sequences. It utilises the stratified splice site identification model described in section 3.2 to identify canonical GT and AG splice sites and the model described in section *3.3* to identify non-canonical GC donor sites. Stratasplice utilises a Bayesian probabilistic framework and reports both log-odds bit scores and posterior probabhties for all of its predlctions. For easy integration into gene prediction systems such as DOUBLESCAN (Meyer and Durbin, unpublished) or GAZE (Howe and Durbin, unpublished), Stratasplice accepts fasta files as input and outputs its predlctions in GFF format (see http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml). Stratasplice is avadable free of charge from the Sanger Centre website at http://www.sanger.ac.uk/software/analysis/stratasplice/).

Program usage

StrataSplice is a command-line program written in Java 1.2 (available from http://java.sun.com). It has been extensively tested on a variety of Unix platforms but should run on Windows and other environments that support Java as well. Stratasplice is provided as a Java Archive file (.jar) and is run in its default mode as follows:

```
java -fast -jar Stratasplice.jar filename
```

In addition to the filename, whch should be the full path to any valid fasta file containing one or more sequences, a number of parameters (see 'Table 3.5) may be used to customise StrataSplice's performance. 'The order of the parameters is not important as long as each parameter is provided at most one time and all parameters precede the file name.

Flag	Type	Description	Default
s	Numeric	Log-odds bit threshold	Negative infinity
p	Numeric	Posterior probability score threshold	*0*
g	String	Genomic data file	–1 Mb taken in 10 kb chunk from 100 random genomic clones
a	String	AG true file	training file derived from SpliccDB
b	String	AG false file	training file derived from genomic data set
c	Numeric	AG prior probability	9.94e-4
d	String	GT true filc	training file derived from SpliceDB
e	String	GT false file	training file derived from genomic data set
f	Numeric	GT prior probability	1.37c-3
h	String	GC true file	training file derived from SpliceDB
i	String	GC false file	training file derived from genomic data set
j	Numeric	GC prior probabhty	1.12e-5