JMB

Prediction of Complete Gene Structures in Human Genomic DNA

Chris Burge* and Samuel Karlin

Department of Mathematics
Stanford University, Stanford
CA, 94305, USA

**Corresponding author*

We introduce a general probabilistic model of the gene structure of human genomic sequences which incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. Distinct sets of model parameters are derived to account for the many substantial differences in gene density and structure observed in distinct C + G compositional regions of the human genome. In addition, new models of the donor and acceptor splice signals are described which capture potentially important dependencies between signal positions. The model is applied to the problem of gene identification in a computer program, GENSCAN, which identifies complete exon/intron structures of genes in genomic DNA. Novel features of the program include the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. GENSCAN is shown to have substantially higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes, with 75 to 80% of exons identified exactly. The program is also capable of indicating fairly accurately the reliability of each predicted exon. Consistently high levels of accuracy are observed for sequences of differing C + G content and for distinct groups of vertebrates.

Keywords: exon prediction; gene identification; coding sequence; probabilistic model; splice signal

Introduction

The problem of identifying genes in genomic DNA sequences by computational methods has attracted considerable research attention in recent years. From one point of view, the problem is closely related to the fundamental biochemical issues of specifying the precise sequence determinants of transcription, translation and RNA splicing. On the other hand, with the recent shift in the emphasis of the Human Genome Project from physical mapping to intensive sequencing, the problem has taken on significant practical importance, and computer software for exon prediction is routinely used by genome sequencing laboratories (in conjunction with other methods) to help identify genes in newly sequenced regions.

Many early approaches to the problem focused on prediction of individual functional elements, e.g. promoters, splice sites, coding regions, in isolation (reviewed by Gelfand, 1995). More recently, a number of approaches have been developed which integrate multiple types of information including splice signal sensors, compositional properties of coding and non-coding DNA and in some cases database homology searching in order to predict entire gene structures (sets of spliceable exons) in genomic sequences. Some examples of such programs include: FGENEH (Solovyev *et al.*, 1994), GENMARK (Borodovsky & McIninch, 1993), GeneID (Guigó *et al.*, 1992), Genie (Kulp *et al.*, 1996), GeneParser (Snyder & Stormo, 1995), and GRAIL II (Xu *et al.*, 1994). Fickett (1996) offers an up-to-date introduction to gene finding by computer and points up some of the strengths and weaknesses of currently available methods. Two important limitations noted are that the majority of current algorithms assume that the input sequence contains

Abbreviations used: Sn, sensitivity; Sp, specificity; CC, correlation coefficient; AC, approximate correlation; ME, missed exons; WE, wrong exons; snRNP, small nuclear ribonucleoprotein particle; snRNA, small nuclear RNA; WMM, weight matrix model; WAM, weight array model; MDD, maximal dependence decomposition.

0022-2836/97/160078–17 $25.00/0/mb970951

exactly one complete gene (so that, when presented with a sequence containing a partial gene or multiple genes, the results generally do not make sense); and that accuracy measured by independent control sets may be considerably lower than was originally thought. The issue of the predictive accuracy of such methods has recently been addressed through an exhaustive comparison of available methods using a large set of vertebrate gene sequences (Burset & Guigó, 1996). The authors conclude that the predictive accuracy of all such programs remains rather low, with less than 50% of exons identified exactly by most programs. Thus, development of new methods (and/or improvement of existing methods) continues to be important.

Here, we introduce a general probabilistic model for the (gene) structure of human genomic sequences and describe the application of this model to the problem of gene prediction in a program called GENSCAN. Our goal in designing the genomic sequence model was to capture the general and specific compositional properties of the distinct functional units of a eukaryotic gene: exon, intron, splice site, promoter, etc. Emphasis was placed on those features which are recognized by the general transcriptional, splicing and translational machinery which process most or all protein coding genes, rather than specialized signals related to transcription or (alternative) splicing of particular genes or gene families. Thus, for example, we include the TATA box and cap site which are present in most eukaryotic promoters, but not specialized or tissue-specific transcription factor binding sites such as those bound by MyoD (e.g. Lassar *et al.*, 1989). Similarly, we use a general three-periodic (inhomogeneous) fifth-order Markov model of coding regions rather than using specialized models of particular protein motifs or data base homology information. As a consequence, predictions made by the program do not depend on presence of a similar gene in the protein sequence databases, but instead provide information which is independent and complementary to that provided by homology-based gene identification methods such as searching the protein databases with BLASTX (Gish & States, 1993). Additionally, the model takes into account many of the often quite substantial differences in gene density and structure (e.g. intron length) that exist between different C + G% compositional regions ("isochores") of the human genome (Bernardi, 1989; Duret *et al.*, 1995).

Our model is similar in its overall architecture to the Generalized Hidden Markov Model approach adopted in the program Genie (Kulp *et al.*, 1996), but differs from most existing programs in several important respects. First, we use an explicitly double-stranded genomic sequence model in which potential genes occuring on both DNA strands are analyzed in simultaneous and integrated fashion. Second, while most existing integrated gene finding programs assume that in each input sequence there is exactly one complete gene, our model treats the general case in which the sequence may contain a partial gene, a complete gene, multiple complete (or partial) genes, or no gene at all. The combination of the double-stranded nature of the model and the capacity to deal with variable numbers of genes should prove particularly useful for analysis of long human genomic contigs, e.g. those of a hundred kilobases or more, which will often contain multiple genes on one or both DNA strands. Third, we introduce a novel method, Maximal Dependence Decomposition, to model functional signals in DNA (or protein) sequences which allows for dependencies between signal positions in a fairly natural and statistically justifiable way. This method is applied to generate a model of the donor splice signal which captures several types of dependencies which may relate to the mechanism of donor splice site recognition in pre-mRNA sequences by U1 small nuclear ribonucleoprotein particle (U1 snRNP) and possible other factors. Finally, we demonstrate that the predictive accuracy of GENSCAN is substantially better than other methods when tested on standardized sets of human and vertebrate genes, and show that the method can be used effectively to predict novel genes in long genomic contigs.

Results

GENSCAN was tested on the Burset/Guigó set of 570 vertebrate multi-exon gene sequences (Burset & Guigó, 1996): the standard measures of predictive accuracy per nucleotide and per exon are shown in Table 1A (see Table legend for details). Comparison of the accuracy data shows that GENSCAN is significantly more accurate at both the nucleotide and the exon level by all measures of accuracy than existing programs which do not use protein sequence homology information (those in the upper portion of Table 1A). At the nucleotide level, substantial improvements are seen in terms of Sensitivity ($Sn = 0.93$ *versus* 0.77 for the next best program, FGENEH), Approximate Correlation ($AC = 0.91$ *versus* 0.78 for FGENEH) and Correlation Coefficient ($CC = 0.92$ *versus* 0.80 for FGENEH). At the exon level, significant improvements are seen across the board, both in terms of Sensitivity ($Sn = 0.78$ *versus* 0.61 for FGENEH) and Specificity ($Sp = 0.81$ *versus* 0.64 for FGENEH), as well as Missed Exons ($ME = 0.09$ *versus* 0.15 for FGENEH) and Wrong Exons ($WE = 0.05$ *versus* 0.11 for GRAIL). Surprisingly, GENSCAN was found to be somewhat more accurate by almost all measures than the two programs, GeneID+ and GeneParser3, which make use of protein sequence homology information (Table 1A). Exon-level sensitivity and specificity values were substantially higher for GENSCAN and Wrong Exons substantially lower; only in the category of Missed Exons did GeneID+ do better (0.07 *versus* 0.09 for GEN-

Table 1. Performance comparison for Burset/Guigó set of 570 vertebrate genes

A *Comparison of GENSCAN with other gene prediction programs*

		Accuracy per nucleotide				Accuracy per exon				
Program	Sequences	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
GENSCAN	570 (8)	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
FGENEH	569 (22)	0.77	0.88	0.78	0.80	0.61	0.64	0.64	0.15	0.12
GeneID	570 (2)	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
Genie	570 (0)	0.76	0.77	0.72	n/a	0.55	0.48	0.51	0.17	0.33
GenLang	570 (30)	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	562 (0)	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL2	570 (23)	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	561 (0)	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	570 (28)	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	478 (1)	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	478 (1)	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09

B *GENSCAN accuracy for sequences grouped by C + G content and by organism*

		Accuracy per nucleotide				Accuracy per exon				
Subset	Sequences	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
C + G <40	86 (3)	0.90	0.95	0.90	0.93	0.78	0.87	0.84	0.14	0.05
C + G 40-50	220 (1)	0.94	0.92	0.91	0.91	0.80	0.82	0.82	0.08	0.05
C + G 50-60	208 (4)	0.93	0.93	0.90	0.92	0.75	0.77	0.77	0.08	0.05
C + G >60	56 (0)	0.97	0.89	0.90	0.90	0.76	0.77	0.76	0.07	0.08
Primates	237 (1)	0.96	0.94	0.93	0.94	0.81	0.82	0.82	0.07	0.05
Rodents	191 (4)	0.90	0.93	0.89	0.91	0.75	0.80	0.78	0.11	0.05
Non-mam. Vert.	72 (2)	0.93	0.93	0.90	0.93	0.81	0.85	0.84	0.11	0.06

A, For each sequence in the test set of 570 vertebrate sequences constructed by Burset & Guigó (1996), the forward-strand exons in the optimal GENSCAN parse of the sequence were compared to the annotated exons (GenBank "CDS" key). The standard measures of predictive accuracy per nucleotide and per exon (described below) were calculated for each sequence and averaged over all sequences for which they were defined. Results for all programs except GENSCAN and Genie are from Table 1 of Burset & Guigó (1996); Genie results are from Kulp *et al.* (1996). Recent versions of Genie have demonstrated substantial improvements in accuracy over that given here (M. G. Reese, personal communication). To calculate accuracy statistics, each nucleotide of a test sequence is classified as predicted positive (*PP*) if it is in a predicted coding region or predicted negative (*PN*) otherwise, and also as actual positive (*AP*) if it is a coding nucleotide according to the annotation, or actual negative (*AN*) otherwise. These assignments are then compared to calculate the number of true positives, $TP = PP \cap AP$ (i.e. the number of nucleotides which are both predicted positives and actual positive); false positives, $FP = PP \cap AN$; true negatives, $TN = PN \cap AN$; and false negatives, $FN = PN \cap AP$. The following measures of accuracy are then calculated: Sensitivity, $Sn = TP/AP$; Specificity, $Sp = TP/PP$; Correlation Coefficient,

$$CC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}};$$

and the Approximate Correlation,

$$AC = \frac{1}{2}\left[\frac{TP}{AP} + \frac{TP}{PP} + \frac{TN}{AN} + \frac{TN}{PN}\right] - 1.$$

The rationale for each of these definitions is discussed by Burset & Guigó (1996). At the exon level, predicted exons (*PP*) are compared to the actual exons (*AP*) from the annotation; true positives (*TP*) is the number of predicted exons which exactly match an actual exon (i.e. both endpoints exactly correct). Exon-level sensitivity (*Sn*) and specificity (*Sp*) are then defined using the same formulas as at the nucleotide level, and the average of *Sn* and *Sp* is calculated as an overall measure of accuracy *in lieu* of a correlation measure. Two additional statistics are calculated at the exon level: Missed Exons (*ME*) is the proportion of true exons not overlapped by any predicted exon, and Wrong Exons (*WE*) is the proportion of predicted exons not overlapped by any real exon. Under the heading Sequences, the number of sequences (out of 570) effectively analyzed by each program is given, followed by the number of sequences for which no gene was predicted, in parentheses. Performance of the programs which make use of amino acid similarity searches, GeneID+ and GeneParser3, are shown separately at the bottom of the Table: these programs were run only on sequences less than 8 kb in length. B, Results of GENSCAN for different subsets of the Burset/Guigó test set, divided either according to the C + G% composition of the GenBank sequence or by the organism of origin. Classification by organism was based on the GenBank "ORGANISM" key. Primate sequences are mostly of human origin; rodent sequences are mostly from mouse and rat; the non-mammalian vertebrate set contains 22 fish, 17 amphibian, 5 reptilian and 28 avian sequences.

SCAN). Use of protein sequence homology information in conjunction with GENSCAN predictions is addressed in the Discussion.

Going beyond exons to the level of whole gene structures, we may define the "gene-level accuracy" (*GA*) for a set of sequences as the proportion of actual genes which are predicted exactly, i.e. all coding exons predicted exactly with no additional predicted exons in the transcription unit (in practice, the annotated GenBank sequence). Gene-level accuracy was 0.43 (243/570) for GENSCAN in the Burset/Guigó set, demonstrating that it is indeed possible to predict complete multi-exon gene structures with a reasonable degree of success by computer. It should be noted that this proportion almost certainly overstates the true gene-level accuracy of GENSCAN because of the substantial bias in the Burset/Guigó set towards small genes (mean: 5.1 kb) with relatively simple intron-exon structure (mean: 4.6 exons per gene). Nevertheless, GENSCAN was able to correctly reconstruct some highly complex genes, the most dramatic example being the human gastric (H + + K+)-ATPase gene (accession no. J05451), containing 22 coding exons. The performance of GENSCAN was found to be relatively insensitive to C + G content (Table 1B), with CC values of 0.93, 0.91, 0.92 and 0.90 observed for sequences of < 40, 40 to 50, 50 to 60, and >60% C + G, respectively, and similarly homogeneous values for the *AC* statistic. Nor did accuracy vary substantially for different subgroups of vertebrate species (Table 1B); CC was 0.91 for the rodent subset, 0.94 for primates and 0.93 for a diverse collection of non-mammalian vertebrate sequences.

A feature which may prove extremely useful in practical applications of GENSCAN is the " forward-backward " probability, *p*, which is calculated for each predicted exon as described in Methods. Specifically, of the 2678 exons predicted in the Burset/Guigó set: 917 had $p > 0.99$ and, of these, 98% were exactly correct; 551 had $p \in [0.95, 0.99]$ (92% correct); 263 had $p \in [0.90, 0.95]$ (88% correct);337 had $p \in [0.75, 0.90]$ (75% correct); 362 had $p \in [0.50, 0.75]$ (54% correct); and 248 had $p \in [0.00, 0.50]$, of which 30% were correct. Thus, the forward-backward probability provides a useful guide to the likelihood that a predicted exon is correct and can be used to pinpoint regions of a prediction which are more certain or less certain. From the data above, about one half of predicted exons have $p > 0.95$, with the practical consequence that any (predicted) gene with four or more exons will likely have two or more predicted exons with $p > 0.95$, from which PCR primers could be designed to screen a cDNA library with very high likelihood of success.

Since for GENSCAN, as for most of the other programs tested, there was a certain degree of overlap between the "learning" set and the Burset/Guigó test set, it was important also to test the method on a truly independent test set. For this purpose, in the construction of the learning set $\mathcal{L}$, we removed all genes more than 25% identical at the amino acid level to the genes of the previously published GeneParser test sets (Snyder & Stormo, 1995), as described in Methods. Accuracy statistics for GENSCAN, GeneID, GeneParser2 and GRAIL3 (GRAIL II+ " assembly" option) on GeneParser test sets I and II are given in Table 2. In this Table, exons correct is the proportion of true exons which were predicted exactly, essentially the same as the

Table 2. Performance comparison for GeneParser Test Sets I, II

Program:	GeneID		GRAIL3		GeneParser2		GENSCAN	
All sequences	I	II	I	II	I	II	I	II
Correlation (CC)	0.69	0.55	0.83	0.75	0.78	0.80	0.93	0.93
Sensitivity	0.69	0.50	0.83	0.68	0.87	0.82	0.98	0.95
Specificity	0.77	0.75	0.87	0.91	0.76	0.86	0.90	0.94
Exons correct	0.42	0.33	0.52	0.31	0.47	0.46	0.79	0.76
Exons overlapped	0.73	0.64	0.81	0.58	0.87	0.76	0.96	0.91
High C + G	I	II	I	II	I	II	I	II
Correlation (CC)	0.65	0.73	0.88	0.80	0.89	0.71	0.94	0.98
Sensitivity	0.72	0.85	0.87	0.80	0.90	0.65	1.00	0.98
Specificity	0.73	0.73	0.95	0.88	0.93	0.87	0.91	0.98
Exons correct	0.38	0.43	0.67	0.50	0.64	0.57	0.76	0.64
Exons overlapped	0.80	0.86	0.89	0.79	0.96	0.79	1.00	0.93
Medium C + G	I	II	I	II	I	II	I	II
Correlation (CC)	0.67	0.52	0.83	0.75	0.75	0.82	0.93	0.94
Sensitivity	0.65	0.47	0.86	0.68	0.86	0.84	0.97	0.95
Specificity	0.77	0.76	0.84	0.91	0.70	0.87	0.90	0.95
Exons correct	0.37	0.29	0.51	0.32	0.41	0.46	0.79	0.79
Exons overlapped	0.67	0.62	0.83	0.38	0.84	0.79	0.96	0.93
Low C + G	I	II	I	II	I	II	I	II
Correlation (CC)	0.81	0.62	0.62	0.62	0.72	0.67	0.92	0.81
Sensitivity	0.82	0.56	0.51	0.45	0.79	0.71	0.93	0.80
Specificity	0.85	0.71	0.87	0.89	0.75	0.67	0.94	0.84
Exons correct	0.80	0.47	0.25	0.16	0.40	0.37	0.85	0.68
Exons overlapped	0.85	0.63	0.55	0.42	0.85	0.58	0.85	0.74

GENSCAN was run on GeneParser test sets I (28 sequences) and II (34 sequences), described in Snyder & Stormo (1995). Accuracy statistics for programs other than GENSCAN are from Table 1 of Snyder & Stormo (1995). For each program, accuracy statistics for test set I are shown in the left column, for test set II in the right column. Nucleotide-level accuracy statistics *Sn*, *Sp* and CC were calculated as described in the legend to Table 1, except that the convention used for averaging the statistics was that of Snyder and Stormo. In this alternative approach, the raw numbers (*PP*, *PN*, *AP*, *AN*, *TP*, etc.) from each sequence are summed and the statistics calculated from these total numbers rather than calculating separate statistics for each sequence and then averaging. (For large sequence sets, these two conventions almost always give similar results.) Exon-level accuracy statistics are also calculated in this fashion. Here, exons correct is the proportion of true exons which were predicted exactly (both endpoints correct), essentially the same as exon-level sensitivity. Exons overlapped is the proportion of true exons which were at least overlapped by predicted exons, a less stringent measure of accuracy not requiring exact prediction of splice sites. Each test set was divided into three subsets according to the C + G content of the GenBank sequence: low C + G (<45%), medium C + G (45 to 60%), and high C + G (>60%).

exon-level sensitivity statistic of Burset & Guigó (1996). Comparison of the GENSCAN accuracy statistics for the two GeneParser test sets (Table 2) with each other and with those for the Burset/Guigó test set (Table 1) show little difference in predictive accuracy. For example, identical correlation coefficient values of 0.93 were observed in both GeneParser test sets *versus* 0.92 in the Burset/Guigó test set. Similarly, the proportion of exons correct was 0.79 and 0.76 in GeneParser test sets I and II, as compared to 0.78 for the corresponding value (exon-level sensitivity) in the Burset/Guigó set. Again, performance of the program is quite robust with respect to differences in C + G content; the somewhat larger fluctuations observed in Table 2 undoubtedly relate to the much smaller size of the GeneParser test sets.

Of course, it might be argued that none of the accuracy results described above are truly indicative of the program's likely performance on long genomic contigs, since all three of the test sets used consist primarily of relatively short sequences containing single genes, whereas contigs currently being generated by genome sequencing laboratories are often tens to hundreds of kilobases in length and may contain several genes on either or both DNA strands. To our knowledge, only one systematic test of a gene prediction program (GRAIL) on long human contigs has so far been reported in the literature (Lopez *et al.*, 1994), and the authors encountered a number of difficulties in carrying out this test, e.g. it was not always clear whether predicted exons not matching the annotation were false positives or might indeed represent real exons which had not been found by the original submitters of the sequence. As a test of the performance of gene prediction programs on a large human contig, we ran GENSCAN and GRAIL II on the recently sequenced CD4 gene region of human chromosome 12p13 (Ansari-Lari *et al.*, 1996), a contig of 117 kb in length in which six genes have been detected and characterized experimentally.

Annotated genes, GENSCAN predicted genes, and GRAIL predicted exons in this sequence are displayed in Figure 1: both programs find most of the known exons in this region, but significant differences between the predictions are observed. Comparison of the GENSCAN predicted genes (GS1 through GS8) with the annotated (known) genes showed that: GS1 corresponds closely to the CD4 gene (the predicted exon at about 1.5 kb is actually a non-coding exon of CD4); GS2 is identical to one of the alternatively spliced forms of Gene A; GS3 contains several exons from both Gene B and GNB3; GS5 is identical to ISOT, except for the addition of one exon at around 74 kb; and GS6 is identical to TPI, except with a different translation start site. This leaves GS4, GS7 and GS8 as potential false positives, which do not correspond to any annotated gene, of which GS7 and GS8 are overlapped by GRAIL predicted exons.

A BLASTP (Altschul *et al.*, 1990) search of the predicted peptides corresponding to GS4, GS7 and GS8 against the non-redundant protein sequence databases revealed that: GS8 is substantially identical (BLAST score 419, $P = 2.6$ E-57) to mouse 60 S ribosomal protein (SwissProt accession no. P47963); GS7 is highly similar (BLAST score 150, $P = 2.8$ E-32) to *Caenorhabditis elegans* predicted protein C26E6.5 (GenBank accession no. 532806); and GS4 is not similar to any known protein (no BLASTP hit with $P < 0.01$). Examination of the sequence around GS8 suggests that this is probably a 60 S ribosomal protein pseudogene. Predicted gene GS7 might be an expressed gene, but we did not detect any hits against the database of expressed sequence tags (dbEST) to confirm this. However, we did find several ESTs substantially identical to the predicted 3'UTR and exons of GS4 (GenBank accession no. AA070439, W92850, AA055898, R82668, AA070534, W93300 and others), strongly implying that this is indeed an expressed human gene which was missed by the submitters of this sequence (probably because GRAIL did not detect it). Aside from the prediction of this novel gene, this example also illustrates the potential of GENSCAN to predict the number of genes in a sequence fairly well: of the eight genes predicted, seven correspond closely to known or putative genes and only one (GS3) corresponds to a fusion of exons from two known genes.

Discussion

As the focus of the human genome project shifts from mapping to large-scale sequencing, the need for efficient methods for identifying genes in anonymous genomic DNA sequences will increase. Experimental approaches will always be required to prove the exact locations, transcriptional activity and splicing patterns of novel genes, but if computational methods can give accurate and reliable indications of exon locations beforehand, the experimental work involved may often be significantly reduced. We have developed a probabilistic model of human genomic sequences which approximates many of the important structural and compositional features of human genes, and have described the implementation of this model in the GENSCAN program to predict exon/gene locations in genomic sequences. Novel features of the method include: (1) use of distinct, explicit, empirically derived sets of model parameters to capture differences in gene structure and composition between distinct C + G compositional regions (isochores) of the human genome; (2) the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occuring on either or both DNA strands; and (3) new statistical models of donor and acceptor splice sites which capture potentially important dependencies between signal positions. Significant improvements in predictive

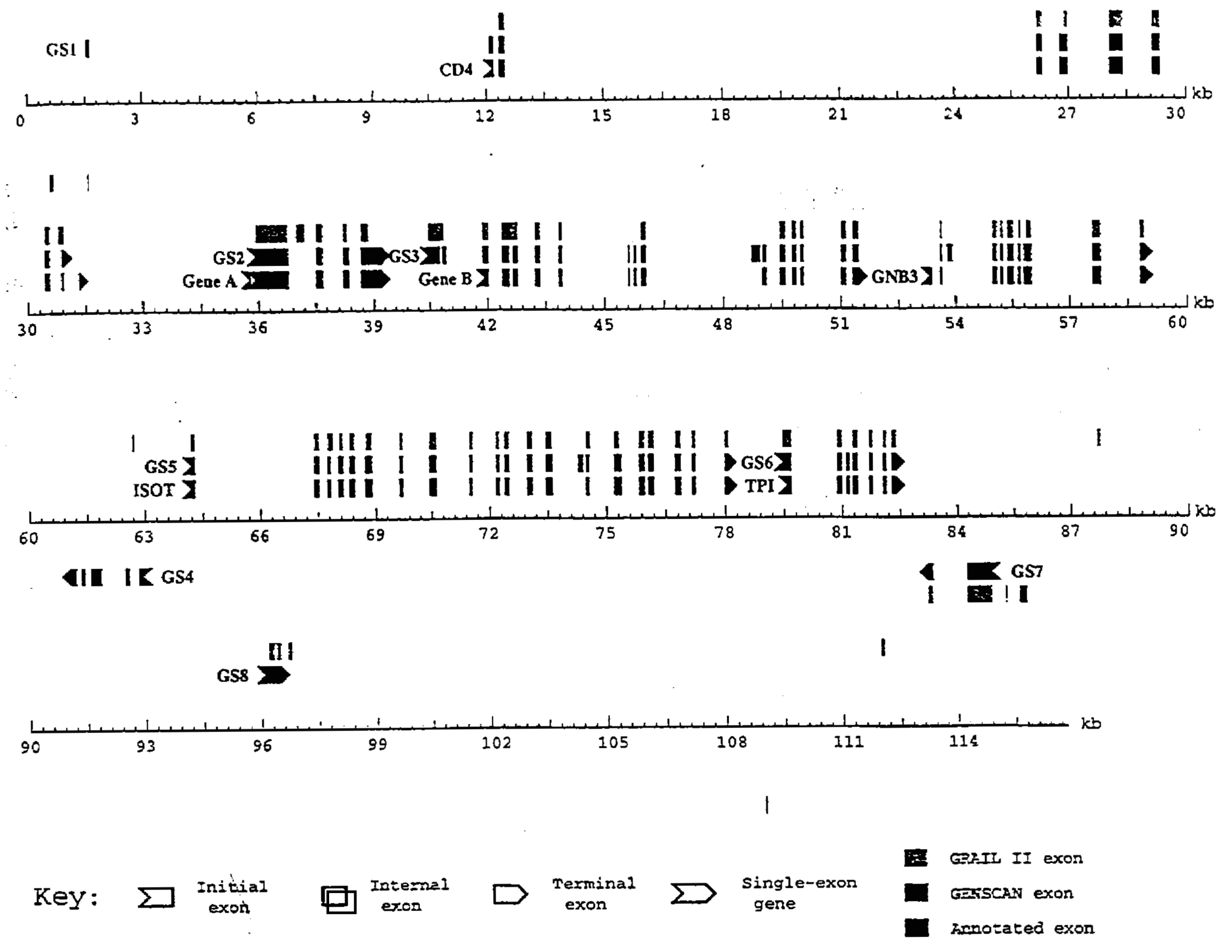


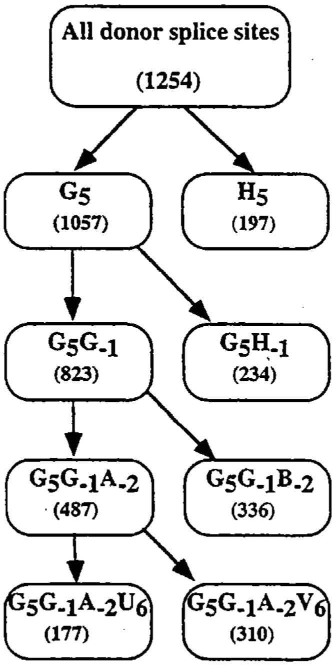
Figure 1. A diagram of GenBank sequence HSU47924 (accession no U47924, length 116,879 bp) is shown with annotated coding exons (from the GenBank CDS features) in black, GENSCAN predicted exons in dark gray, and GRAIL predicted exons in light gray. Exons on the forward strand are shown above the sequence line; on the reverse (complementary) strand, below the sequence line. GRAIL II was run through the email server (grail@ornl.gov): final predicted exons of any quality are shown. Exon sizes and positions are to scale, except for initial, terminal and single-exon genes, which have an added arrow-head or -tail (see key above) which causes them to appear slightly larger than their true size. Since GRAIL does not indicate distinct exon types (initial *versus* internal *versus* terminal exons), all GRAIL exons are shown as internal exons. Gene names for the six annotated genes in this region (CD4, Gene A, Gene B, GNB3, ISOT and TPI) are shown on the annotation line, immediately preceding the first coding exon of the gene. The GENSCAN predicted genes are labeled GS1 to GS8 as they occur along the sequence.

accuracy have been demonstrated for GENSCAN over existing programs, even those which use protein sequence homology information, and we have shown that the program can be used to detect novel genes even in sequences previously subjected to intensive computational and experimental scrutiny.

In practice, several distinct types of computer programs are often used to analyze a newly sequenced genomic region. The sequence may first be screened for repetitive elements with a program like CENSOR (Jurka *et al.*, 1996). Following this, GENSCAN and/or other gene prediction programs could be run, and the predicted peptide sequences searched against the protein sequence databases with BLASTP (Altschul *et al.*, 1990) to detect possible homologs. If a potential homolog is detected, one might perhaps refine the prediction by submitting the genomic region corresponding to the predicted gene together with the potential protein homolog to the program Procrustes (Gelfand *et al.*, 1996), which uses a "spliced alignment" algorithm to match the genomic sequence to the protein. Even in the absence of a protein homolog, it may be possible to confirm the expression and precise 3′ terminus of a predicted gene using the database of Expressed Sequence Tags (Boguski, 1995). Finally, a variety of experimental approaches such as RT-PCR and 3′ RACE are typically used (see, e.g., Ansari-Lari *et al.*, 1996) to pinpoint precise exon/intron boundaries and possible alternatively spliced forms. At this stage, computational approaches may also prove useful, e.g. GENSCAN high probability exons could be used to design PCR primers. The GENSCAN program has been made available through the World Wide Web [http:// gnomic.stanford.edu/GENSCANW.html] and by electronic mail (mail sequence in FastA format to genscan@gnomic.stanford.edu).

It is hoped that studies of the statistical properties of genes may yield clues to the sequence dependence of the basic biochemical processes of transcription, translation and RNA splicing which define genes biologically. As an example of such an application, we close with a discussion of some of the statistical properties of donor splice sites brought out by application of the Maximal Dependence Decomposition (MDD)· approach (see Methods). Overall, the results support the well established hypothesis that base-pairing with U1 snRNA, or with other factors of identical specificity, is of primary importance in donor site recognition (e.g. McKeown, 1993). However, the MDD data of Figure 2 also suggest some fairly subtle properties of the U1:donor interaction, namely: (1) a 5′/3′ compensation effect, in which matches to consensus nucleotides at nearby positions on the same side of the intron/exon junction are positively associated, while poor matching on one side of the junction is almost always compensated by stronger matching on the other side; (2) an adjacent base-pair effect, in which base-pairs at the edge of the donor splice site form only in the presence of adjacent base-pairs; and (3) a G_3 preference effect, in which G is preferred at position +3 only for a subclass of strongly U1-binding donor sites. The evidence for each of these effects is summarized below.

Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19



Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

All sites:					**Position**					
Base	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	
A%	33	**60**	8	0	0	**49**	**71**	6	15	
C%	37	13	4	0	0	3	7	5	19	
G%	18	14	**81**	**100**	0	45	12	**84**	20	
U%	12	13	7	0	**100**	3	9	5	**46**	
U1 snRNA: 3'	**G**	**U**	**C**	**C**	**A**	**U**	**U**	**C**	**A**	**5'**

Figure 2. The subclassification of donor splice sites according to the maximal dependence method is illustrated. Each box represents a subclass of donor splice sites corresponding to a particular pattern of matches and mismatches to the consensus nucleotide(s) at a set of positions in the donor site, e.g. G_5 is the set of donor sites with G at position +5 and G_5G_{-1} is the set of donors with G at both positions +5 and −1. Here, H indicates A, C or U; B indicates C, G or U; and V indicates A, C or G. The number of sites in each subset is given in parentheses. The data set and donor site position conventions are as described in the legend to Table 4. The frequencies (percentages) of each of the four nucleotides at each variable position are indicated for each subclass immediately adjacent to the corresponding box. Data for the entire set of 1254 donor sites are given at the bottom of the Figure: the frequencies of consensus nucleotides are shown in boldface. The sequence near the 5′ end of U1 snRNA which base-pairs with the donor site is shown at the bottom in 3′ to 5′ orientation.

5′/3′ compensation effect

First, G_{-1} is almost completely conserved (97%) in H_5 donor sites (those with a non-G nucleotide at position + 5) *versus* 78% in G_5 sites, suggesting that absence of the G·C base-pair with U1 snRNA at position +5 can be compensated for by a G·C base-pair at position −1, with a virtually absolute requirement for one of these two G·C base-pairs (only five of 1254 donor sites lacked both G_5 and G_{-1}). Second, the H_5 subset exhibits substantially higher consensus matching at position −2 ($A_{-2} = 85\%$ in H_5 *versus* 56% in G_5), while the G_5 subset exhibits stronger matching at positions +4 and +6. Similar compensation is also observed in the G_5G_{-1} *versus* G_5H_{-1} comparison: the G_5H_{-1} subset exhibits substantially higher consensus matching at positions +6 (76% *versus* 42%), +4 (93% *versus* 70%) and +3 (100% R_3 *versus* 93%). Yet another example of compensation is observed in the G_5G_{-1} A_{-1} *versus* G_5G_{-1} B_{-1} comparison, with the $G_5G_{-1}B_{-2}$ subset exhibiting increased consensus matching at positions +4 and +6, but somewhat lower matching at position −3.

Adjacent base-pair effect

H_5 splice sites have nearly random (equal) usage of the four nucleotides at position +6, strongly implying that base-pairing with U1 at position +6 does not occur (or does not aid in donor recognition) in the absence of a base-pair at position +5. The almost random distribution of nucleotides at position −3 of the $G_5G_{-1}B_{-2}$ donor sites also suggests that base-pairing with U1 snRNA at position −3 does not occur or is of little import in the absence of a base-pair at position −2.

G_3 preference effect

Comparison of the relative usage of A *versus* G at position +3 in the various subsets reveals several interesting features. Perhaps surprisingly, G is almost as frequent as A at position +3 (45% *versus* 49%) in the entire set of donor sites, despite the expected increased stability of an A·U *versus* G·U base-pair at position +3. Only in subset H_5 is a dramatic preference for A over G at position +3 observed (81% *versus* 15%), suggesting that only in the absence of the strong G·C base-pair at position +5 does the added binding energy of an A·U *versus* G·U base-pair at position +3 become critical to donor site recognition by U1 snRNA. On the other hand, in the most strongly consensus-matching donor site subset, $G_5G_{-1}A_{-2}U_6$, there is actually a strong preference for G_3 over A_3 (59% *versus* 27%)! Two possible explanations for this observation seem reasonable: either (1) there is selection to actually weaken the U1:donor interaction in these strongly matching sites so that U1 snRNA can more easily dissociate from the donor site to permit subsequent steps in splicing; or (2) G_3 is preferred over A_3 at some step in splicing subsequent to donor site selection.

Methods

Sequence sets

The non-redundant sets of human single- and multi-exon genes constructed by David Kulp and Martin Reese (22 Aug., 1995) were used as a starting point for database construction [ftp://ftp.cse.ucsc.edu/pub/dna/genes]. These sets consist of GenBank files, each containing a single complete gene (at least ATG → stop, but often including 5′ and 3′ untranslated and flanking regions) sequenced at the genomic level, which have been culled of redundant or substantially similar sequences using BLASTP (Altschul *et al.*, 1990). We further cleaned these sets by removing genes with CDS or exons annotated as putative or uncertain (e.g. GenBank files HSALDC, HUMADH6), alternatively spliced genes (HSCALCAC, HSTCRT3D), pseudogenes (e.g. HSAK3PS, HSGKP1), and genes of viral origin (HBNLF1), resulting in a set of 428 sequences. For testing purposes, we further reduced this set by removing all genes more than 25% identical at the amino acid level to those of the GeneParser test sets (Snyder & Stormo, 1995) using the PROSET program (Brendel, 1992) with default parameters. The set of 238 multi-exon genes and 142 single-exon (intronless) genes remaining after this procedure are collectively referred to as the learning set, designated $\mathscr{L}$ (gene list available upon request). The total size of the set is 2,580,965 bp: the multi-exon genes in $\mathscr{L}$ contain a total of 1492 exons and 1254 introns.

All model parameters, e.g. state transition and initial probabilities, splice site models, etc. were derived from this data set as described later in this section, with two notable exceptions: (1) the promoter model, which was based on published sources; and (2) the coding region model, for which this set was supplemented with a set of complete human cDNA sequences derived as follows. All complete human cDNA sequences corresponding to proteins of at least 100 amino acids in length (the length minimum was imposed in order to avoid inclusion of cDNA fragments) were extracted from GenBank Release 83 (June, 1994). This set was then cleaned at the amino acid level using PROSET as above both with respect to itself and with respect to the GeneParser test sets (gene list available upon request). This set was then combined with the coding sequences from $\mathscr{L}$ to form a set $\mathscr{C}$ of 1999 complete coding sequences totaling in excess of 3195 kb.

Model of genomic sequence structure

Figure 3 illustrates a general model of the structure of genomic sequences. In this model, the (hidden) states of the model (represented as circles and diamonds in the Figure) correspond to fundamen-

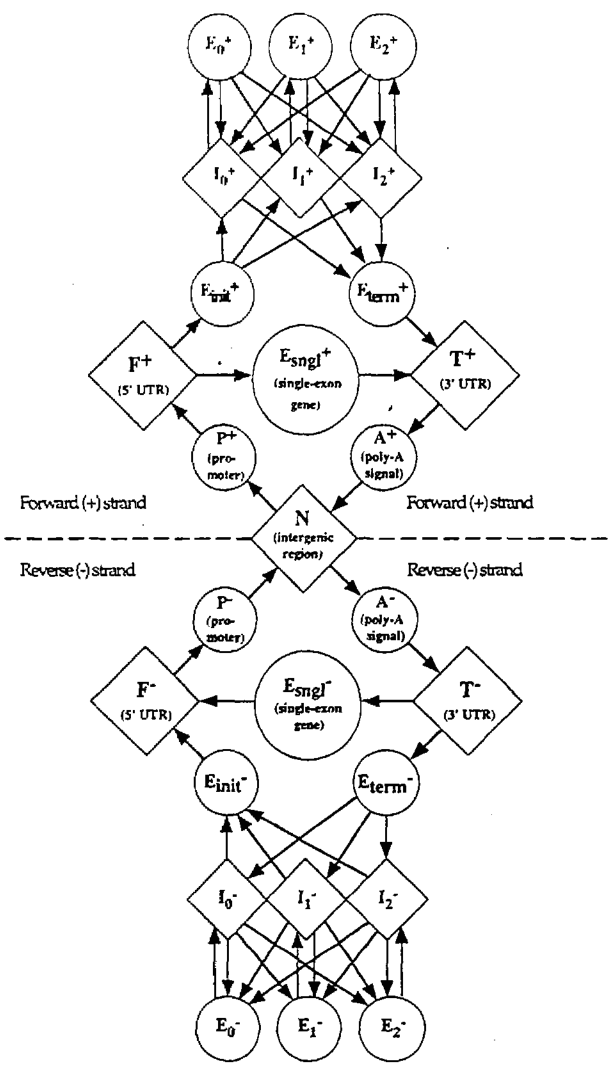


Figure 3. Each circle or diamond represents a functional unit (state) of a gene or genomic region: *N*, intergenic region; *P*, promoter; *F*, 5′ untranslated region (extending from the start of transcription up to the translation initiation signal); E_{sngl}, single-exon (intronless) gene (translation start → stop codon); E_{init}, initial exon (translation start → donor splice site); E_k $(0 \leqslant k \leqslant 2)$, phase *k* internal exon (acceptor splice site → donor splice site); E_{term}, terminal exon (acceptor splice site → stop codon); *T*, 3′ untranslated region (extending from just after the stop codon to the polyadenylation signal); *A*, polyadenylation signal; and I_k $(0 \leqslant k \leqslant 2)$, phase *k* intron (see the text). For convenience, translation initiation/termination signals and splice sites are included as subcomponents of the associated exon state and intron states are considered to extend from just after a donor splice site to just before the branch point/acceptor splice site. The upper half of the Figure corresponds to the states (designated with a superscript +) of a gene on the forward strand, while the lower half (designated with superscript −) corresponds to a gene on the opposite (complementary) strand. For example, proceeding in the 5′ to 3′ direction on the (arbitrarily chosen) forward strand, the components of an E^+_k (forward-strand internal exon) state will be encountered in the order: (1) acceptor site, (2) coding region, (3) donor site, while the components of an E_k^- (reverse-strand internal exon) state will be encountered in the order: (1) inverted complement of donor site, (2) inverted complement of coding region, (3) inverted complement of acceptor site. Only the intergenic state *N* is not divided according to strand.

tal functional units of a eukaryotic gene, e.g. exon, intron, intergenic region, etc. (see Figure legend for details), which may occur in any biologically consistent order. Note that introns and internal exons in our model are divided according to "phase", which is closely related to the reading frame. Thus, an intron which falls between codons is considered phase 0; after the first base of a codon, phase 1; after the second base of a codon, phase 2, denoted I_0, I_1, I_2, respectively. Internal exons are similarly divided according to the phase of the previous intron (which determines the codon position of the first base-pair of the exon, hence the reading frame). For convenience, donor and acceptor splice sites, translation initiation and termination signals are considered as part of the associated exon.

Reverse strand states and forward strand states are dealt with simultaneously in this model, somewhat similar to the treatment of both strands in the GENMARK program (Borodovsky & McIninch, 1993); see the legend to Figure 3. Though somewhat similar to the model described by Kulp *et al.* (1996), our model is substantially more general in that it includes: (1) single as well as multi-exon genes; (2) promoters, polyadenylation signals and intergenic sequences; and (3) genes occuring on either or both DNA strands. In addition, as mentioned previously, partial as well as complete genes are permitted as is the occurrence of multiple genes in the same sequence. Thus, the essential structure of most vertebrate genomic sequences likely to be encountered in genome sequencing projects can be described by this model structure. The most notable limitations are that overlapping transcription units (probably rare) cannot be handled and that alternative splicing is not explicitly addressed.

The model, essentially of semi-Markov type, is conveniently formulated as an explicit state duration Hidden Markov Model (HMM) of the sort described by Rabiner (1989). Briefly, the model is though of as generating a "parse" ϕ, consisting of an ordered set of states, $\bar{q} = \{q_1, q_2 \ldots, q_n\}$, with an associated set of lengths (durations), $\bar{d} = \{d_1, d_2, \ldots, d_n\}$ which, using probabilistic models of each of the state types, generates a DNA sequence *S* of length $L = \Sigma_{i=1}^{n} d_i$. The generation of a parse corresponding to a (pre-defined) sequence length *L* is as follows:

(1) An initial state q_1 is chosen according to an initial distribution on the states, $\bar{\pi}$, i.e. $\pi_i = P\{q_1 = Q^{(i)}\}$, where $Q^{(j)} (j = 1, \ldots, 27)$ is an indexing of the state types (Figure 3).

(2) A length (state duration), d_1, corresponding to the state q_1 is generated conditional on the value of $q_1 = Q^{(i)}$ from the length distribution $f_{Q(i)}$.

(3) A sequence segment s_1 of length d_1 is generated, conditional on d_1 and q_1, according to an appropriate sequence generating model for state type q_1.

(4) The subsequent state q_2 is generated, conditional on the value of q_1, from the (first-order Markov) state transition matrix T, i.e. $T_{i,j} = P\{q_{k+1} = Q^{(j)} | q_k = Q^{(i)}\}$.

This process is repeated until the sum, $\Sigma_{i=1}^{n} d_i$, of the state durations first equals or exceeds the length L, at which point the last state duration d_n is appropriately truncated, the final stretch of sequence is generated, and the process stops: the sequence generated is simply the concatenation of the sequence segments, $S = s_1 s_2 \ldots s_n$. Note that the sequence of states generated is not restricted to correspond to a single gene, but could represent a partial gene, several genes, or no genes at all. The model thus has four main components: a vector of initial probabilities $\bar{\pi}$, a matrix of state transition probabilities T, a set of length distributions f, and a set of sequence generating models P. Assuming for the moment that these four components have been specified, the model can be used for prediction in the following way.

For a fixed sequence length L, consider the space $\Omega = \Phi_L \times \mathscr{S}_L$, where Φ_L is the set of (all possible) parses of length L and $\mathscr{S}_L$ is the set of (all possible) DNA sequences of length L. The model M can then be thought of as a probability measure on this space, i.e. a function which assigns a probability density to each parse/sequence pair. Thus, for a particular sequence $S \in \mathscr{S}_L$, we can calculate the conditional probability of a particular parse $\phi_i \in \Phi_L$ (under the probability measure induced by M) using Bayes' Rule as:

$$P\{\phi_i | S\} = \frac{P\{\phi_i, S\}}{P\{S\}} = \frac{P\{\phi_i, S\}}{\sum_{\phi_j \in \Phi_L} P\{\phi_j, S\}} \tag{1}$$

The essential idea is that a precise probabilistic model of what a gene/genomic sequence looks like is specified in advance and then, given a sequence, one determines which of the vast number of possible gene structures (involving any valid combination of states/lengths) has highest likelihood given the sequence. In addition to the optimal parse, it may also be of interest to study sub-optimal parses and/or sub-optimal exons or introns (to be described elsewhere).

Algorithmic issues

Given a sequence S of length L, the joint probability, $P\{\phi_i, S\}$, of generating the parse ϕ_i and the sequence S is given by:

$$\begin{aligned} P\{\phi_i, S\} = {} & \pi_{q1} f_{q1}(d_1) P\{s_i | q_1, d_1\} \\ & \times \prod_{k=2}^{n} T_{q_{k-1}, q_k}(d_k) P\{s_k | q_k, d_k\} \end{aligned} \tag{2}$$

where the states of ϕ_i are $q_1, q_2, \ldots, q_n$ with associated state lengths $d_1, d_2, \ldots, d_n$, which break the sequence into segments $s_1, s_2, \ldots, s_n$. Here $P\{s_k | q_k, d_k\}$ is the probability of generating the sequence segment s_k under the appropriate sequence generating model for a type-q_k state of length d_k. A recursive algorithm of the sort devised by Viterbi (Viterbi, 1967; Forney, 1973) may then be used to calculate ϕ_{opt}, the parse with maximal joint probability (under M), which gives the predicted gene or set of genes in the sequence. Variations of this algorithm have been described and used on several occasions previously in sequence analysis (e.g. Sankoff, 1992; Gelfand & Roytberg, 1993). Certain modifications must be made to the standard algorithm for the semi-Markov case used here *versus* the simpler Markov case. The specific algorithm used is described by Burge (1997); see also Rabiner (1989, section IV D).

Calculation of $P\{S\}$ may be carried out using the "forward" algorithm; the "backward" algorithm is also implemented in order to calculate certain additional quantities of interest (both algorithms are described by Burge, 1997; see also Rabiner, 1989). Specifically, consider the event $E_{[x,y]}^{(k)}$ that a particular sequence segment $[x, y]$ is an internal exon of phase k. Under M, this event has probability

$$P\{E_{[x,y]}^{(k)} | S\} = \frac{\sum_{\phi_i : E_{[x,y]}^{(k)} \in \phi_i} P\{\phi_i, S\}}{P\{S\}} \tag{3}$$

where the sum is taken over all parses which contain the given exon $E_{[x,y]}^{(k)}$. This sum can be conveniently calculated using the "forward-backward" procedure, which is described in general by Rabiner (1989) and more specifically by Burge (1997); see also Stormo & Haussler (1994) where a similar idea was introduced in the context of exon-intron prediction. This probability has been shown to be a useful guide to the degree of certainty which should be ascribed to exons predicted by the program (see Results). Run time for the GENSCAN program, though at worst quadratic in the number of possible state transitions, in practice grows approximately linearly with sequence length for sequences of several kb or more. Typical run time for a sequence of length X kb on a Sun Sparc10 workstation is about X + 5 seconds.

Initial and transition probabilities

Since we are attempting to model a randomly chosen block of contiguous human genomic DNA as might be generated by a genome sequencing laboratory, the initial probability of each state should be chosen proportionally to its estimated frequency in bulk human (or vertebrate) genomic DNA. However, even this is not trivial since gene density and certain aspects of gene structure are known to vary quite dramatically in regions of differing C + G% content (so-called "isochores") of the human genome (Bernardi, 1989, 1993; Duret *et al.*, 1995), with a much higher gene density in

Table 3. Gene density and structure as a function of C + G composition: derivation of initial and transition probabilities

Group	I	II	III	IV
C + G% range	<43	43-51	51-57	>57
Number of genes	65	115	99	101
Est. proportion single-exon genes	0.16	0.19	0.23	0.16
Codelen: single-exon genes (bp)	1130	1251	1304	1137
Codelen: multi-exon genes (bp)	902	908	1118	1165
Introns per multi-exon gene	5.1	4.9	5.5	5.6
Mean intron length (bp)	2069	1086	801	518
Est. mean transcript length (bp)	10866	6504	5781	4833
Isochore	L1 + L2	H1 + H2	H3	H3
DNA amount in genome (Mb)	2074	1054	102	68
Estimated gene number	22100	24700	9100	9100
Est. mean intergenic length	83000	36000	5400	2600
Initial probabilities:				
Intergenic (N)	0.892	0.867	0.540	0.418
Intron (I_0^+, I_1^+, I_2^+, I_0^-, I_1^-, I_2^-)	0.095	0.103	0.338	0.388
5′ Untranslated region (F^+, F^-)	0.008	0.018	0.077	0.122
3′ Untranslated region (T^+, T^-)	0.005	0.011	0.045	0.072

The top portion of the Table shows data from the learning set of 380 genes, partitioned into four groups according to the C + G% content of the GenBank sequence; the middle portion shows estimates of gene density from Duret *et al.* (1995) for isochore compartments corresponding to the four groups above; the bottom portion shows the initial probabilities used by GENSCAN for sequences of each C + G% compositional group, which are estimated using data from the top and middle portions of the Table. All of the values in the top portion are observed values, except the proportion of single-exon genes. Since single-exon genes are typically much shorter than multi-exon genes at the genomic level (due to the absence of introns) and hence easier to sequence completely, they are probably substantially over-represented in the learning set relative to their true genomic frequency; accordingly, the proportion of single-exon genes in each group was estimated (somewhat arbitrarily) to be one half of the observed fraction. Codelen refers to the total number of coding base-pairs per gene. Data for subsets III and IV are estimated from the Duret *et al.* (1995) data for isochore H3 assuming that one-half of the genes and 60% of the amount of DNA sequence in isochore H3 falls into the 51 to 57% C + G range. Mean transcript lengths were estimated assuming an average of 769 bp of 5′UTR and 457 bp of 3′UTR per gene (these values derived from comparison of the "prim_transcript" and "CDS" features of the GenBank annotation in the genes of the learning set). To simplify the model, the initial probabilities of the exon, polyadenylation signal and promoter states are set to zero. All other initial probabilities are estimated from the data shown above, assuming that all features are equally likely to occur on either DNA strand. The initial probability for all intron states was partitioned among the three intron phases according to the observed fraction of each phase in the learning set. Transition probabilities were estimated analogously.

C + G-rich regions than in A + T-rich regions. Therefore, separate initial and transition probability distributions are estimated for sequences in each of four categories: I (<43% C + G); II (43 − 51); III (51 − 57); and IV (>57), corresponding approximately to isochore compartments L1 + L2, H1 + H2, and two subsets of the H3 isochore, respectively. Details are given in Table 3 and its legend. Note that the differences in estimated initial probabilities are quite dramatic with, for example, the probability of hitting an intergenic region much higher in A + T-rich sequences than for C + G-rich ones.

The (biologically permissible) state transitions are shown as arrows in Figure 3. Certain transitions are obligatory (e.g. $P^+ \rightarrow F^+$) and hence are assigned probability one; all others are assigned (maximum likelihood) values equal to the observed state transition frequency in the learning set $\mathscr{L}$ for the appropriate C + G compositional group. Overall, transition frequencies varied to a lesser degree between groups than did initial probabilities (Table 3). There was a trend (possibly related to biases in the dataset toward genes with shorter genomic length) for A + T-rich genes to have fewer introns, leading to slightly different estimates for the $I_j^+ \rightarrow E_{term}^+$ probabilities.

State length distributions

In general, the states of the model (see Figure 3) correspond to sequence segments of highly variable length. For certain states, most notably the internal exon states $E_{i,j}$, length is probably an important property for proper biological function (i.e. proper splicing and inclusion in the final processed mRNA). For example, it has been shown *in vivo* that internal deletions of constitutively recognized internal exons to sizes below about 50 bp may often lead to exon skipping, i.e. failure to include the exon in the final processed mRNA (Dominski & Kole, 1991), and there is some evidence that steric interference between factors recognizing splice sites may make splicing of small exons more difficult (e.g. Black, 1991). Of course, some very small exons do exist and are efficiently spliced. At the other end, there is some evidence that spliceosomal assembly is inhibited if internal exons are internally expanded beyond about 300 nucleotides (Robberson *et al.*, 1990), but conflicting evidence also exists (Chen & Chasin, 1994), and the lengths of flanking introns may also be important (Sterner *et al.*, 1996). Overall, most results have tended to support the idea that "medium-sized" internal exons (between about 50 and 300 bp in length) may be more easily spliced than excessively long

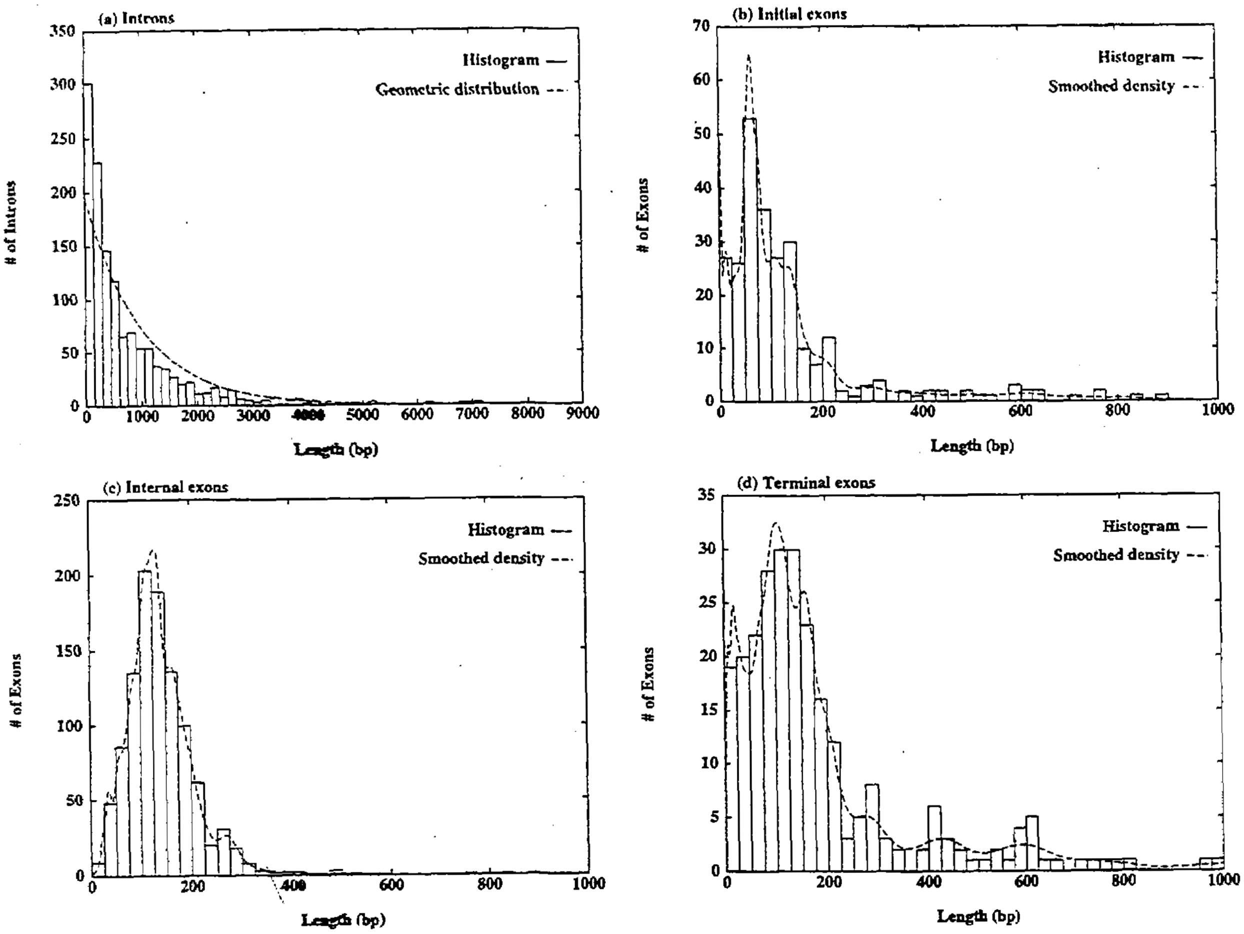


Figure 4. Length distributions are shown for (a) 1254 introns; (b) 238 initial exons; (c) 1151 internal exons; and (d) 238 terminal exons from the 238 multi-exon genes of the learning set $\mathcal{L}$. Histograms (continuous lines) were derived with a bin size of 300 bp in (a), and 25 bp in (b), (c), (d). The broken line in (a) shows a geometric (exponential) distribution with parameters derived from the mean of the intron lengths; broken lines in (b), (c) and (d) are the smoothed empirical distributions of exon lengths used by GENSCAN (details given by Burge, 1997). Note different horizontal and vertical scales are used in (a), (b), (c), (d) and that multimodality in (b) and (d) may, in part, reflect relatively small sample sizes.

or short exons, and this idea is given substantial support by the observed distribution of internal exon lengths (Figure 4(c)), which shows a pronounced peak at around 120 to 150 nucleotides, with few internal exons more than 300 bp or less than 50 bp in length. (See also Hawkins (1988) for an extensive discussion of exon and intron length distributions.) Initial (Figure 4(b)) and terminal (Figure 4(d)) exons also have substantially peaked distributions (possibly multi-modal) but do not exhibit such a steep dropoff in density after 300 bp, suggesting that somewhat different constraints may exist for splicing of exons at or near the ends of the pre-mRNA. Taking these factors into account, we use separate empirically derived length distribution functions for initial, internal, and terminal exons (Figure 4) and for single-exon genes. Substantial differences in exon length distributions were not observed between the C + G compositional groups (data not shown).

In contrast to exons, intron length does not appear to be critical to splicing in most cases, e.g. for rabbit β-globin, intron length was observed to be unimportant for splicing provided that a certain minimum threshold of perhaps 70 to 80 nucleotides was exceeded (Wieringa *et al.*, 1984). The observed distribution of intron lengths (Figure 4(a)) tends to support this idea: no introns less than 65 bp were observed, but above this size the distribution appears to be approximately geometric (exponential), consistent with the absence of significant functional constraints on intron length. Consistent with the results of Duret *et al.* (1995), dramatic differences were observed in intron (and intergenic) lengths between the four C + G compositional groups (Table 3): introns in (A + T-rich) group I genes averaged 2069 bp, almost four times the value of 518 bp observed in very C + G-rich genes (group IV). Thus, intron and intergenic lengths are modeled as geometric distributions with parameter q estimated for each C + G group separately. For the 5′UTR and 3′UTR states, we use geometric distributions with mean values of 769 and 457 bp, respectively, derived from comparison of the "prim_transcript" and "CDS" features of the GenBank files in $\mathscr{L}$. The polyA_signal and promoter model lengths are discussed later. The only other feature of note is that exon lengths must be consistent with the phases of adjacent introns. To account for this, exon lengths are generated in two steps: first, the number of complete codons is generated from the appropriate length distribution; then the appropriate number (0, 1 or 2) of bp is added to each end to account for the phases of the preceding and subsequent states. For example, if the number of complete codons generated for an initial exon is c and the phase of the subsequent intron is i, then the total length of the exon is: $l = 3c + i$.

Signal models

Numerous models of biological signal sequences such as donor and acceptor splice sites, promoters, etc. have been constructed in the past ten years or so. One of the earliest and most influential approaches has been the weight matrix method (WMM) introduced by Staden (1984), in which the frequency $p_j^{(i)}$ of each nucleotide j at each position i of a signal of length n is derived from a collection of aligned signal sequences and the product $P\{X\} = \Pi_{i=1}^{n} p_{xi}^{(i)}$ is used to estimate the probability of generating a particular sequence, $X = x_1, x_2, \ldots, x_n$. A generalization of this method, termed weight array model (WAM), was applied by Zhang & Marr (1993), in which dependencies between adjacent positions are considered. In this model, the probability of generating a particular sequence is: $Pr\{X\} = p_{x_1}^{(1)} \Pi_{i=2}^{n} p_{x_{i-1},x_i}^{i-1,i}$, where $p_{j,k}^{(i-1,i)}$ is the conditional probability of generating nucleotide X_k at position i, given nucleotide X_j at position $i - 1$ (which is estimated from the corresponding conditional frequency in the set of aligned signal sequences). Of course, higher-order WAM models capturing second-order (triplet) or third-order (tetranucleotide) dependencies in signal sequences could be used in principle, but typically there is insufficient data available to estimate the increased number of parameters in such models. Here, WMM models are used for certain types of signals, a modified WAM model is derived for acceptor splice sites, and a new model, termed Maximal Dependence Decomposition (MDD), is introduced to model donor splice sites.

Transcriptional and translational signals

Polyadenylation signals are modeled as a 6 bp WMM (consensus: AATAAA). A 12 bp WMM model, beginning 6 bp prior to the initiation codon, is used for the translation initiation (Kozak) signal. In both cases, the WMM probabilities were estimated using the GenBank annotated "polyA_signal" and "CDS" features from sequences of $\mathscr{L}$. (Similer models of these signals have been used by others, e.g. Guigó *et al.* (1992), Snyder & Stormo (1995).) For the translation termination signal, one of the three stop codons is generated (according to its observed frequency in $\mathscr{L}$) and the next three nucleotides are generated according to a WMM. For promoters, we use a simplified model of what is undoubtedly an extremely complex signal often involving combinatorial regulation. Our primary goal was to construct a model flexible enough so that potential genes would not be missed simply because they lacked a sequence similar to our preconceived notion of what a promoter should look like. Since about 30% of eukaryotic promoters lack an apparent TATA signal, we use a split model in which a TATA-containing promoter is generated with probability 0.7 and a TATA-less promoter with probability 0.3. The TATA-containing promoter is modeled using a 15 bp TATA-box WMM and an 8 bp cap site WMM, both borrowed from Bucher (1990). The length between the WMMs is generated uniformly from the range of 14 to 20 nucleotides, corresponding to a TATA → cap site distance of 30 to 36 bp, from the first T of the TATA-box matrix to the cap site (start of transcription). Intervening bases are generated according to an intergenic-null model, i.e. independently generated from intergenic base frequencies. At present, TATA-less promoters are modeled simply as intergenic-null regions of 40 bp in length. In the future, incorporation of improved promoter models, e.g. perhaps along the lines of Prestridge (1995), will probably lead to more accurate promoter recognition.

Splice signals

The donor and acceptor splice signals are probably the most critical signals for accurate exon prediction since the vast majority of exons are internal exons and therefore begin with an acceptor site and end with a donor site. Most previous probabilistic models of these sites have assumed either independence between positions, e.g. the WMM model of Staden (1984) or dependencies between adjacent positions only, e.g. the WAM model of Zhang & Marr (1993). However, we have observed highly significant dependencies between non-adjacent as well as adjacent positions in the donor splice signal (see below), which are not adequately accounted for by such models and which likely relate to details of donor splice site recognition by U1 snRNP and possibly other factors. The consensus region of the donor splice site comprises the last 3 bp of the exon (positions −3 to −1) and the first 6 bp of the succeeding intron (positions 1 through 6), with the almost invariant GT dinucleotide occuring at positions 1,2: consensus nucleotides are shown in Figure 2. We have focused on the dependencies between the consensus indicator variable, C_i (1 if the nucleotide at position i matches the consensus at i, 0 otherwise) and the

Table 4. Dependence between positions in human donor splice sites: χ^2-statistic for consensus indicator variable C_i *versus* nucleotide indicator X_j

i	Con	j: −3	−2	−1	+3	+4	+5	+6	Sum
−3	c/a	—	61.8*	14.9	5.8	20.2*	11.2	18.0*	131.8*
−2	A	115.6*	—	40.5*	20.3*	57.5*	59.7*	42.9*	336.5*
−1	G	15.4	82.8*	—	13.0	61.5*	41.4*	96.6*	310.8*
+3	a/g	8.6	17.5*	13.1	—	19.3*	1.8	0.1	60.5*
+4	A	21.8*	56.0*	62.1*	64.1*	—	56.8*	0.2	260.9*
+5	G	11.6	60.1*	41.9*	93.6*	146.6*	—	33.6*	387.3*
+6	t	22.2*	40.7*	103.8*	26.5*	17.8*	32.6*	—	243.6*

C_i and X_j are defined in the text. The last three exon bp and first six intron bp were extracted from each of the 1254 donor splice sites in the learning set: positions in this site are labeled −3 through −1, +1 through +6. The invariant positions +1, +2 (always G, T in this set) are omitted. The consensus nucleotide(s) at each position are shown in the second column: nucleotides with frequency greater than 50% are uppercase (see Figure 2). For each pair of distinct positions $\{i, j\}$, a 2 by 4 contingency table was constructed for the indicator variable C_i (1 if the nucleotide at position i matches the consensus, 0 otherwise) *versus* the variable X_j identifying the nucleotide at position j, and the value of the χ^2 statistic for each such table was calculated. Those values exceeding 16.3 (corresponding to $P < 0.001$, 3 df) are indicated with an asterisk. The last column in the Table lists the sum of the values in each row: this value is a measure of the dependence between C_i and the vector $X^{(i)}$ of the nucleotides at the six remaining positions. All values exceeded 42.3 ($P < 0.001$, 18 df) and are therefore indicated with an asterisk.

nucleotide indicator X_j identifying the nucleotide at position j. Table 4 shows the χ^2 statistics for the variable C_i *versus* X_j for all pairs i, j with $i \neq j$ in the set of donor sites from the genes of the learning set (positions +1 and +2 are omitted since they do not exhibit variability in this data set). Strikingly, almost three-quarters (31/42) of the i, j pairs exhibit significant χ^2 values even at the relatively stringent level of $P < 0.001$ indicating a great deal of dependence between positions in the donor splice site. (The stringent P-value cutoff was used to compensate for the effect of multiple comparisons.) It is also noteworthy and perhaps surprising that many non-adjacent pairs of positions as well as most adjacent pairs exhibit significant dependence, e.g. positions −1 and +6, separated by five intervening nucleotides, exhibit the extremely high χ^2 values of 103.8 for C_6 *versus* X_{-1} and 96.6 for C_{-1} *versus* X_6. In order to account for such dependencies in a natural way, we introduce a new model-building procedure, described next.

Maximal Dependence Decomposition (MDD)

The goal of the MDD procedure is to generate, from an aligned set of signal sequences of moderate to large size (i.e. at least several hundred or more sequences), a model which captures the most significant dependencies between positions (allowing for non-adjacent as well as adjacent dependencies), essentially by replacing unconditional WMM probabilities by appropriate conditional probabilities provided that sufficient data is available to do so reliably. Given a data set D consisting of N aligned sequences of length k, the first step is to assign a consensus nucleotide or nucleotides at each position. Then, for each pair of positions, the χ^2 statistic is calculated for C_i *versus* X_j (as defined above) for each i, j pair with $i \neq j$. If no significant dependencies are detected (for an appropriate P-value), then a simple WMM should be sufficient. If significant dependencies are detected, but they are exclusively or predominantly between adjacent positions, then a WAM model may be appropriate. If, however, there are strong dependencies between non-adjacent as well as adjacent positions, then we proceed as follows. (1) Calculate, for each position i, the sum $S_i = \sum_{j \neq i} \chi^2(C_i, X_j)$ (the row sums in Table 4), which is a measure of the amount of dependence between the variable C_i and the nucleotides at the remaining positions of the site. (2) Choose the value i_1 such that S_{i_1} is maximal and partition D into two subsets: D_{i_1} all sequences which have the consensus nucleotide(s) at position i_1; and $D_{\bar{i}_1}$ all sequences which do not. Now repeat steps (1) and (2) on each of the subsets, D_{i_1} and $D_{\bar{i}_1}$ and on subsets thereof, and so on, yielding a binary subdivision "tree" with (at most) $k - 1$ levels (see Figure 2). This process of subdivision is carried out successively on each branch of the tree until one of the following three conditions occurs: (1) the $(k - 1)$th level of the tree is reached (so that no further subdivision is possible); (2) no significant dependencies between positions in a subset are detected (so that further subdivision is not indicated); or (3) the number of sequences remaining in a subset becomes so small that reliable WMM frequencies could not be determined after further subdivision. Finally, separate WMM models are derived for each subset of the tree, and these are combined to form a composite model as described below.

Figure 2 illustrates the MDD procedure applied to the set of 1254 donor splice sites from $\mathcal{L}$. The initial subdivision is made according to the consensus (G) at position 5 of the donor signal (see Table 4), resulting in subsets G_5 and H_5 (H meaning A, C or U) containing 1057 and 197 intron sequences, respectively. We consider the number 175 as a reasonable minimum subset size (corresponding to a parameter estimation error of typically less than 25%, even for base frequencies as low as 10%), so the subset H_5 is not subdivided. The subset G_5 is sufficiently large, and exhibits significant dependence between positions (data not shown), so it is further subdivided according to the consensus (G) at position −1, yielding subsets G_5G_{-1} and G_5H_{-1}, and so on. The composite MDD model for generation of donor splice site sequences is then as follows. (0) The (invariant) nucleotides X_1 and X_2 are generated. (1) X_5 is generated from the original WMM for all donor sites combined. (2a) If $X_5 \neq G$, then the (conditional) WMM model for subset H_5 is used to generate the nucleotides at the remaining positions in the donor site. (2b) If $X_5 = G$, then X_{-1} is generated from the (conditional) WMM model for the subset G_5. (3a) If ($X_5 = G$ and) $X_{-1} \neq G$, then the WMM model for subset G_5H_{-1} is used. (3b) If ($X_5 = G$ and) $X_{-1} = G$, X_{-2} is generated from the model for G_5G_{-1}; and so on, until the entire 9 bp sequence has been generated. Biological factors related to the MDD model are addressed in the Discussion.

Acceptor splice site model

The first step in the MDD procedure was also applied to the 1254 acceptor sites from the multi-exon genes of $\mathscr{L}$, but dependencies between positions were found to be much weaker than for donor sites and those that existed were mostly between adjacent positions (data not shown). Therefore, we apply a modified WAM method to model this signal. Specifically, bases −20 to +3 relative to the intron/exon junction, encompassing the pyrimidine-rich region and the acceptor splice site itself, are modeled by a first-order WAM model as by Zhang & Marr (1993). The branch point region is notoriously difficult to model, since even the most degenerate branch point consensus is present in only a fraction of acceptor sequences. For example, YYRAY was present in the appropriate region [−40, −21] in only 30% of acceptor sequences in our data set; similarly low frequencies of branch point consensus sequences have been observed previously, e.g. Harris & Senapathy (1990). To model this region, we introduce a "windowed second-order WAM model". (WWAM), in which nucleotides are generated conditional on the nucleotides at the previous two positions. In order to have sufficient data to estimate these conditional probabilities reliably, we averaged the conditional frequencies over a span of five positions, i.e. the WAM entries for position i are formed by averaging the appropriate conditional frequencies at positions $i-2$, $i-1$, i, $i+1$ and $i+2$. This model captures the weak but detectable tendency toward YYY triplets as well as certain branch point-related triplets such as TGA, TAA, GAC, and AAC in this region, without requiring the occurrence of any specific branch point consensus sequence.

Exon models, non-coding state models

Coding portions of exons are modeled using an inhomogeneous 3-periodic fifth-order Markov model as by Borodovsky & McIninch (1993); see also Gelfand (1995). In this approach, separate fifth-order Markov transition matrices are determined for hexamers ending at each of the three codon positions, denoted c_1, c_2, c_3, respectively; exons are modeled using the matrices c_1, c_2, c_3 in succession to generate each codon. These transition probabilities were derived from the set $\mathscr{C}$ of complete coding sequences described previously. In regard to this choice of coding sequence model, we note that Fickett & Tung (1992) have shown that frame-specific hexamer measures are generally the most accurate compositional discriminator of coding *versus* noncoding regions. We found, as have others, that A + T-rich genes are often not well predicted using such bulk hexamer-derived parameters. Accordingly, a separate set of fifth-order Markov transition matrices was derived for C + G composition group I regions (<43% C + G). Specifically, the coding sequences of all group I genes from $\mathscr{L}$ were combined with all cDNAs of <48% C + G from $\mathscr{C}$ (observing that cDNAs are on average about 5% richer in C + G than the genomic region from which they derive): this subset comprised 638 sequences totaling approximately 1.139 Mb.

In our model, the disruption of coding regions by introns in multi-exon genes is dealt with by keeping track of intron/exon phase, ensuring that a consistent reading frame is maintained throughout a gene. Specifically, initial exons begin with codon position c_1 and end with codon position c_j such that $j = i$ mod 3† is the phase of the subsequent intron state; terminal exons will end with codon position c_3 and begin with codon position c_{i+1}, where i is the phase of the previous intron; and internal exons E_i begin with codon position c_{i+1} and end with codon position c_j, where $k \equiv j$ mod 3 is the phase of the subsequent intron. This treatment of the coding portions of multi-exon genes is essentially equivalent to the "in-frame scoring". plus "in-frame assembly" approach described by Wu (1996), which he has shown gives somewhat better accuracy than alternative methods of gene scoring/assembly, e.g. those used by GeneParser (Snyder & Stormo, 1995) and by GRAIL II (Xu *et al.*, 1994). The non-coding states F, T, N and I_k are modeled using a homogeneous fifth-order Markov model, with transition probabilities derived from the non-coding portions of the genes in $\mathscr{L}$. As for coding regions, a separate fifth-order Markov matrix was derived from the genes of group I for use in sequences of <43% C + G.

Reverse-strand states

Sequence generating models for the reverse strand states are derived from the corresponding forward strand models by the simple operation of

† i mod 3 indicates the remainder when i is divided by 3.

inverse complementation. For example, if the forward strand termination signal model generates the triplets *TAG*, *TAA* and *TGA* with probabilities p_1, p_2 and p_3, respectively, then the reverse strand termination model will generate the triplets *CTA* (inverted complement of *TAG*), *TTA* and *TCA*, with probabilities p_1, p_2 and p_3. Equivalently, the forward-strand model is used to generate a stretch of sequence, and then the inverse complement of the sequence is taken.

Acknowledgements

We gratefully acknowledge Drs D. Vollrath, V. Brendel, J. Kleffe, L. Brocchieri and J. Mrazek for helpful comments on the manuscript and M. G. Reese for discussions related to the datasets used. S.K. and C.B. are supported in part by NIH grants 5R01HG00335-09 and 2R01GM10452-32 and NSF grant DMS-9403553-002.

References

Altschul, S. F., Gish, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Ansari-Lari, M. A., Muzny, D. M., Lu, J., Lu, F., Lilley, C. E., Spanos, S., Malley, T. & Gibbs, R. A. (1996). A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human. chromosome 12p13. *Genome Res.* 6, 314–326.

Bernardi, G. (1989). The isochore organization of the human genome. *Annu. Rev. Genet.* 23, 637–661.

Bernard, G (1993). The vertebrate genome: isochores and evolution. *J. Mol. Evol.* 10, 186–204.

Black, D. L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes Dev.* 5, 389–402.

Boguski, M. S. (1995). The turning point in genome research. *Trends Biochem. Sci.* 20, 295–296.

Borodovsky, M. & McIninch, J. (1993). GENMARK: parallel gene recognition for both DNA strands. *Comp. Chem.* 17, 123–133.

Brendel, V. (1992). PROSET-a fast procedure to create non-redundant sets of protein sequences. *Math. Comp. Modeling*, 16, 37–43.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563–578.

Burge, C. (1997). Identification of complete gene structures in human genomic DNA. PhD thesis. Stanford University, Stanford, CA.

Burset, M. & Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, 34, 353–367.

Chen, I. T. & Chasin, L. A. (1994). Large exon size does not limit splicing in vivo. *Mol. Cell. Biol.* 14, 2140–2146.

Dominski, Z. & Kole, R. (1991). Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol.* 11, 6075–6083.

Duret, L., Mouchiroud, D. & Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in CG-rich isochores. *J. Mol. Evol.* 40, 308–317.

Fickett, J. W. (1996). Finding genes by computer: the state of the art. *Trends Genet.* 12(8), 316–320.

Fickett, J. W. & Tung, C.-S. (1992). Assessment of protein coding measures. *Nucl. Acids Res.* 20, 6441–6450.

Forney, G. D. (1973). The Viterbi algorithm. *Proc. IEEE*, 61, 268–278.

Gelfand, M. S. (1995). Prediction of function in DNA sequence analysis. *J. Comp. Biol.* 2(1), 87–115.

Gelfand, M. S. & Roytberg, M. A. (1993). Prediction of the intron-exon structure by a dynamic programming approach. *BioSystems*, 30, 173–182.

Gelfand, M. S., Mironov, A. A. & Pevzner, P. (1996). Gene recognition via spliced alignment. *Proc. Natl Acad. Sci. USA*, 93, 9061–9066.

Gish, W & States, D. J. (1993). Identification of protein coding regions by data base similarity search. *Nature Genet.* 3, 266–272.

Guigó, R., Knudsen, S., Drake, N. & Smith, T. (1992). Prediction of gene structure. *J. Mol. Biol.* 226, 141–157.

Harris, N. L. & Senepathy, P. (1990). Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucl. Acids Res.* 18, 3015–3019.

Hawkins, J. D. (1988). A survey on intron and exon lengths. *Nucl. Acids Res.* 16, 9893–9908.

Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. (1996). CENSOR-a program for identification and elimination of repetitive elements from DNA sequences. *Comp. Chem.* 20(1), 119–122.

Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. (1996). A generalized Hidden Markov Model for the recognition of human genes in DNA. In *Proceedings of the Fourth International Conference on Intelligent System for Molecular Biology.* AAAI Press, Menlo Park, CA.

Lasser, A. B., Buskin, J. N., Lockshon, D., Davis, R. L., Apone, S., Hauschka, S. D. & Weintraub, H. (1989). MyoD is a sequence-specific DNA binding protein requiring a region of myc homology to bind to the muscle creatine kinase enhancer. *Cell*, 58, 823–831.

Lopez, R., Larsen, F. & Prydz, H. (1994). Evaluation of the exon predictions of the GRAIL software. *Genomics*, 24, 133–136.

McKeown, M. (1993). The role of small nuclear RNAs in RNA splicing. *Curr. Opin. Cell Biol.* 5, 448–454.

Prestridge, D. S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* 249, 923–932.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, 77(2), 257–285.

Robberson, B. L., Cote, G. J. & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* 10, 84–94.

Sankoff, D. (1992). Efficient optimal decomposition of a sequence into disjoint regions, each matched to some template in an inventory. *Math. Biosci.* 111, 279–293.

Snyder, E. E. & Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* 248, 1–18.

Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucl. Acid. Res.* 22, 5156–5163.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.* **12**, 505–519.

Sterner, D. A., Carlo, T. & Berget, S. M. (1996). Architectural limits on split genes. *Proc. Natl Acad. Sci. USA*, **93**, 15081–15085.

Stormo, G. D. & Haussler, D. (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 47–55, AAAI Press, Menlo Park, CA.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory*, **IT-13**, 260–269.

Wieringa, B., Hofer, E. & Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit B-globin intron. *Cell*, 37, 915–925.

Wu, T. (1996). A segment-based dynamic programming algorithm for predicting gene structure. *J. Comp. Biol.* **3(3)**, 375–394.

Xu, Y., Einstein, J. R., Mural, R. J., Shah, M. & Uberbacher, E. C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 376–384, AAAI Press, Menlo Park, CA.

Zhang, M. Q. & Marr, T. G. (1993). A weight array method for splicing signal analysis. *Comp. Appl. Biol. Sci.* **9(5)**, 499–509.

Edited by F. E. Cohen

(Received 7 November 1996; received in revised form 29 January 1997; accepted 30 January 1997)