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

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 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

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

A *Comparison of GENSCAN with other gene prediction programs*

Program	Sequences	Accuracy per nucleotide				Accuracy per exon				
		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*

Subset	Sequences	Accuracy per nucleotide				Accuracy per exon				
		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,

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

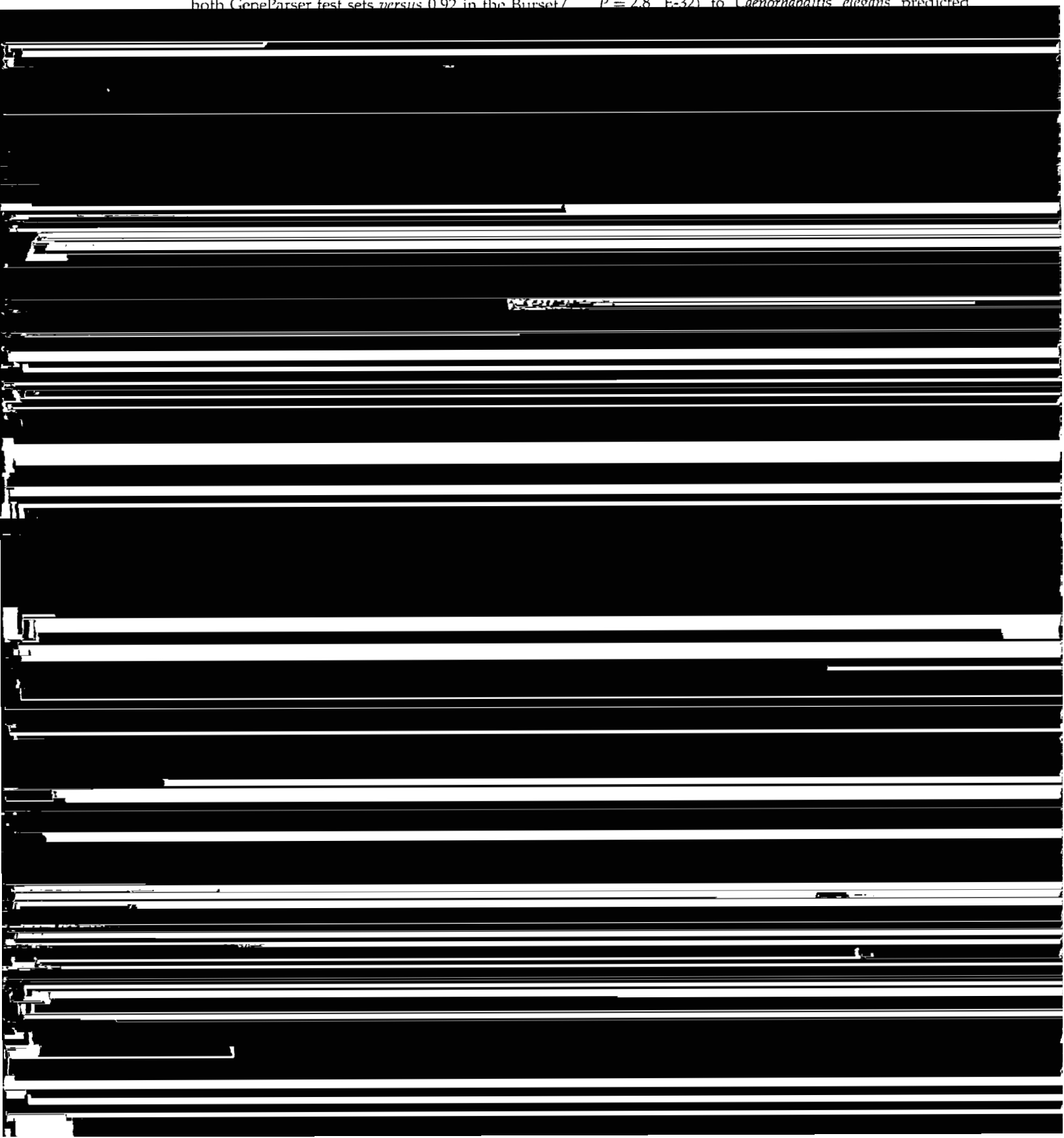
The rationale for each of these definitions is discussed by Burset & Guigó (1996). At the exon level, predicted exons (*PP*) are compared to the actual exons (*AP*) from the annotation; true positives (*TP*) is the number of predicted exons which exactly match an actual exon (i.e. both endpoints exactly 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

curacy 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

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

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/

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



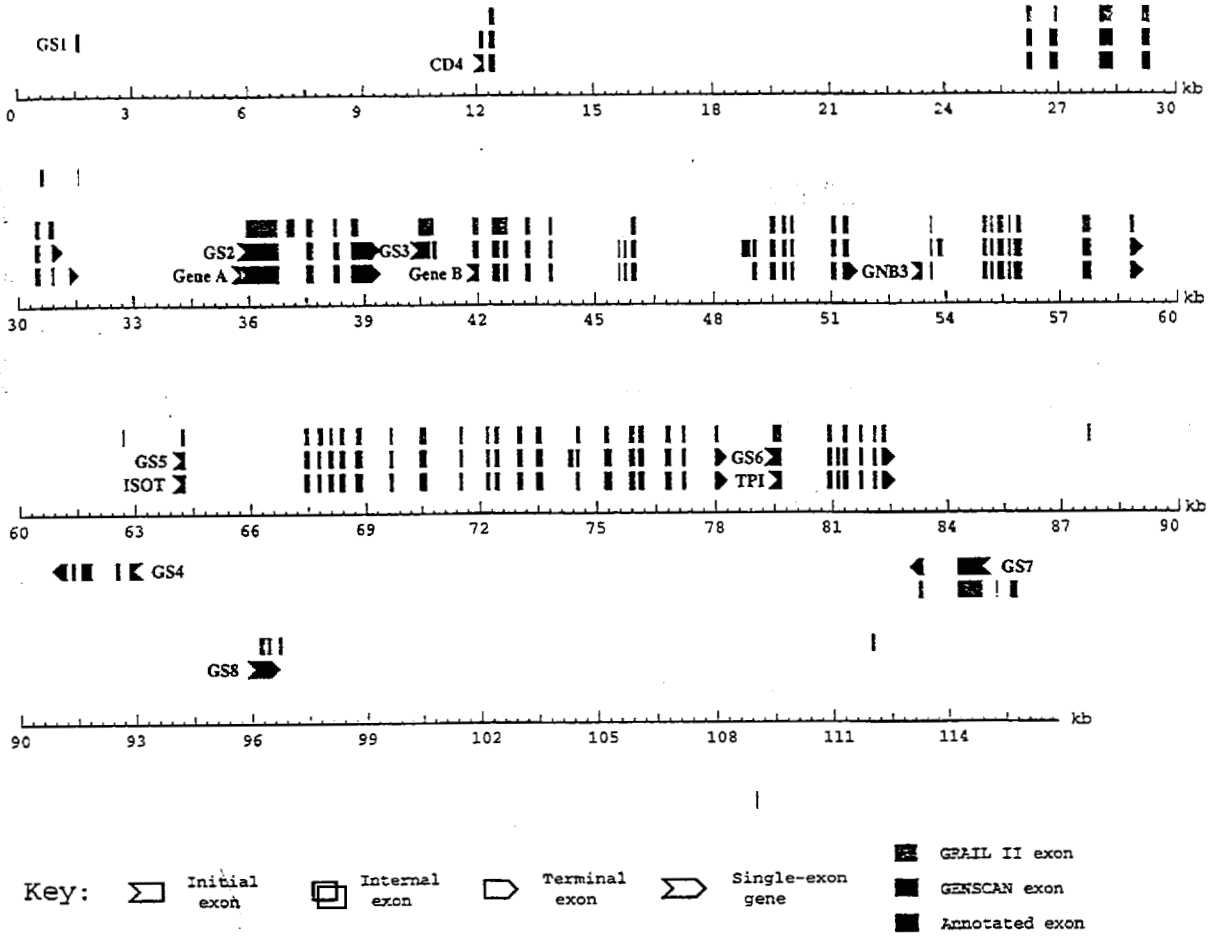


Figure 1. A diagram of GenBank sequence HSU47924 (accession no U47924, length 116,879 bp) is shown with annotated coding exons (from the GenBank CDS features) in black, GENSCAN predicted exons in dark gray, and GRAIL predicted exons in light gray. Exons on the forward strand are shown above the sequence 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

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

ferred over A_3 at some step in splicing subsequent to donor site selection.

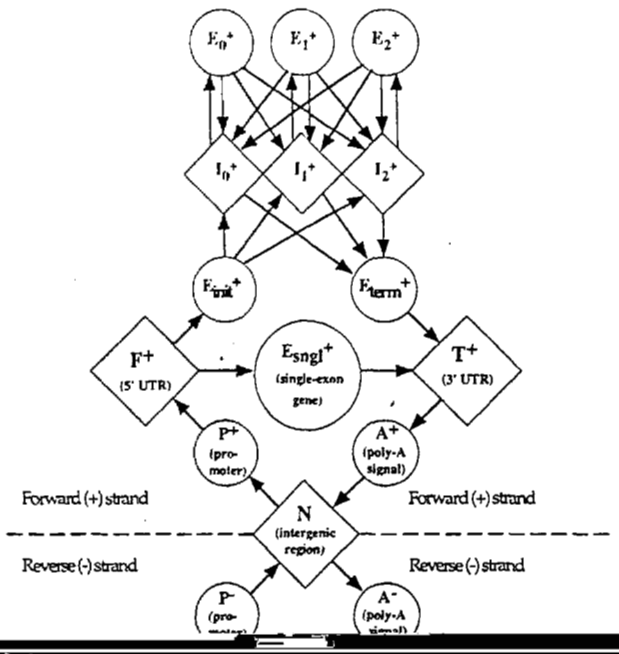
Methods

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 $\mathcal{L}$ (gene list available upon request). The total size of the set is 2,580,965 bp: the multi-exon genes in $\mathcal{L}$ 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 $\mathcal{L}$ to form a set of 1999 complete coding sequences totaling



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 in-

(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 [illegible]. Note that the

ated 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 algor-

[illegible]

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

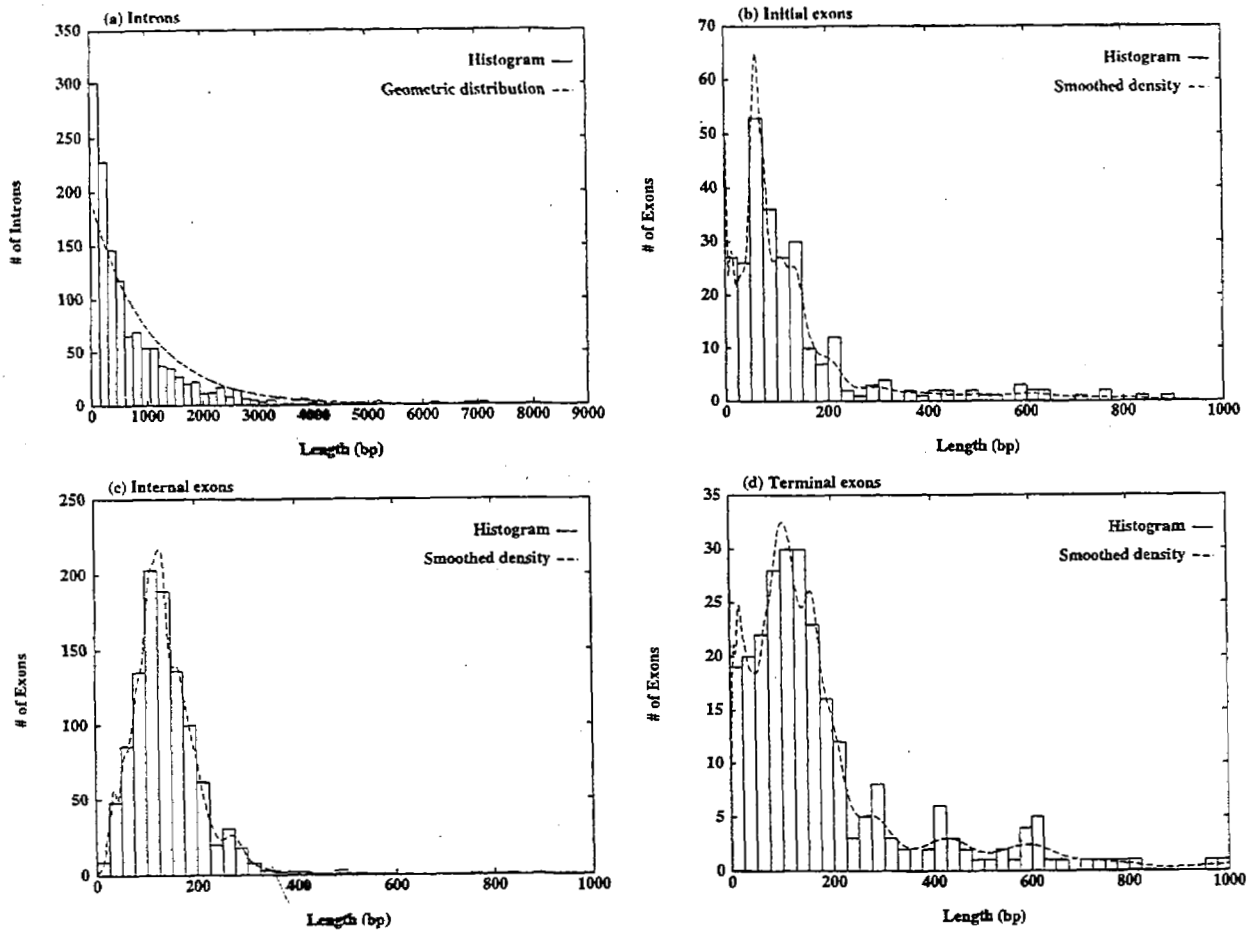


Figure 4. Length distributions are shown for (a) 1254 introns; (b) 238 initial exons; (c) 1151 internal exons; and (d) 238 terminal exons from the 238 multi-exon genes of the learning set $\mathcal{L}$. Histograms (continuous lines) were derived with a bin size of 300 bp in (a), and 25 bp in (b), (c), (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

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

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 $P\{X\} = \Pi_{i=1}^{n} p_{x_i}^{(i)}$ 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 (corre-

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

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%

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.

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.

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

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

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)