Finding the genes in genomic DNA

Christopher B Burge* and Samuel Karlin†

Genome sequencing efforts will soon generate hundreds of millions of bases of human genomic DNA containing thousands of novel genes. In the past year, the accuracy of computational gene-finding methods has improved significantly, to the point where a reasonable approximation of the gene structures within an extended genomic region can often be predicted in advance of more detailed experimental studies.

Addresses
*Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, 40 Ames Street, E17-526 Cambridge, MA 02139, USA; e-mail: cburge@mit.edu
†Department of Mathematics, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA; e-mail: sam@galois.stanford.edu
Correspondence: Samuel Karlin

Current Opinion in Structural Biology 1998, **8**:346–354

http://biomednet.com/elecref/0959440X00800346

Abbreviations

bp	base pair
EST	expressed sequence tag
HMM	hidden Markov model
IMM	interpolated Markov model
LINE	long interspersed nuclear element
MDD	maximal dependence decomposition
ORF	open reading frame
SINE	short interspersed nuclear element
TSS	transcription start site
WWAM	windowed weight array model

Introduction

Identification of all of the genes in the human genome (and in the genomes of various model organisms) is a major objective of the human genome project. Recently, as the genome project has entered the phase of large-scale sequencing, computational approaches to gene finding have begun to draw significant attention from the molecular biology and genomics community. In addition, significant advances in gene-finding methodology have taken place in the past two years, and the current methods are significantly more accurate, reliable and useful than those available in the past. Among recent reviews related to gene finding, we specifically mention a stimulating overview of the subject by Fickett [1], an excellent summary of available programs by Claverie [2•], the landmark comparative study by Burset and Guigo [3••], as well as the more technical article by Gelfand [4]. Additional information can be found in the extensive bibliographies maintained by Li (http://linkage.rockefeller.edu/wli/gene) and Gelfand [5]. Here, we summarize the recent developments in gene-finding algorithms, highlight open problems in this area, and discuss future research directions.

Finding genes in prokaryotic genomes

Gene discovery in prokaryotic genomes is a quite different problem from that encountered in eukaryotic sequences, owing to the higher gene density typical of prokaryotes and the absence of introns in their protein coding genes. These properties generally imply that most open reading frames (ORFs) encountered in a prokaryotic sequence that are longer than some reasonable threshold, such as 300 or 500 base pairs (bp) will likely correspond to genes. The primary difficulties arising from this simple approach are that very small genes will be missed and that the occurrence of overlapping long ORFs on opposite DNA strands (genes and 'shadow genes') often leads to ambiguities. To resolve these problems, several methods have been devised that use different types of Markov models (see below) in order to capture the compositional differences among coding regions, 'shadow' coding regions (coding on the opposite DNA strand), and noncoding DNA. Such methods, including ECOPARSE [6], the widely used GENMARK algorithm [7], and the recently introduced Glimmer program [8], appear to be able to identify most protein coding genes with good specificity, but still have difficulties in predicting the precise position of the start of translation.

Some degree of caution must be exercised in using such statistically-based methods in view of the relatively high frequency of genetic transformation, the occurrence of lateral gene transfer in many bacteria [9], and other factors that lead to heterogeneity in gene composition. Thus, using a principal components-type method to cluster genes on the basis of codon usage, Medigue *et al.* [10] partitioned *Escherichia coli* genes into three groups: 'typical genes' (about 70%), 'highly expressed genes' (about 15%, including ribosomal protein genes, elongation factor genes, and other genes involved in transcription and translation) and a third group (about 15%) consisting of mostly laterally transferred genes. (Similar subdivisions may apply to other bacterial genomes.) The drastic differences in codon usage observed between these three groups indicates that distinct Markov models could be used to find and classify genes of different types. Recently a system called GENMARK Genesis has been developed that automatically clusters ORFs from an uncharacterized bacterial genome and derives separate Markov models for each cluster obtained [11]. The possibility that eukaryotic genomes may also contain distinct clusters of genes is an interesting open problem [12].

Finding genes in eukaryotic sequences

The remainder of this review is devoted to the more complex problem of finding genes in eukaryotic sequences. At this point, it is convenient to define a few essential terms that have specialized meanings in the gene-finding literature. First, the accuracy of prediction of particular gene features, such as exons, coding nucleotides, and splice sites, by a gene-finding method is typically measured in

terms of the 'sensitivity', defined as the proportion of 'true' sites (e.g. exons or donor splice sites) that are correctly predicted, and the 'specificity', defined as the proportion of 'predicted' sites that are correct. It should be clear that an accurate and thorough prediction has been achieved by a method only when the sensitivity and specificity are simultaneously high. Since most methods seek primarily to predict coding exons (as opposed to exons corresponding to 5′ or 3′ untranslated regions), four types of exons are typically distinguished: 'initial exons' (initiation codon to first 5′ splice junction); 'internal exons' (3′ splice site to 5′ splice site); 'terminal exons' (3′ splice site to stop codon); and 'single-exon' (intronless) genes (initiation codon to stop codon). As will be seen below, these four types of exons present different challenges for gene-finding methods, and the methods differ significantly in their ability to predict the four exon types.

Transcriptional signals

The most natural way to find genes computationally would be to mimic as closely as possible the processes of transcription and RNA processing (e.g. splicing and polyadenylation) that define genes biologically. Although this direct approach to gene finding is not yet feasible, a number of important signals related to transcription, translation and splicing are now sufficiently well characterized as to be useful in computer predictions of the location and exon–intron organization of genes. The transcriptional signals most often used in gene finding are the initiator or cap signal, located at the transcription start site (TSS), and the A+T-rich TATA-box signal, typically located about 30 bp upstream of the TSS [13]. These core promoter elements are, however, present in only about 70% of human promoters and, even when present, are not sufficiently precise to allow reliable prediction of promoter locations (reviewed in [14]). Even when the full spectrum of characterized transcription factor-binding sites is used in a promoter recognition algorithm [15], there does not appear to be a significant improvement in the prediction of precise promoter locations when tested on novel promoter sequences [14]. This somewhat disappointing result is probably related to the variability in the location of transcription factor-binding sites relative to the location of the TSS and to the difficulty of accounting for their combinatorial activity. Other features known to play a role in promoter function, such as transcriptional enhancers and silencers, CpG methylation, chromatin structure and DNA curvature, could prove useful in prediction when they are better understood.

The polyadenylation signal appears to have a much simpler structure, comprising a consensus AATAAA hexamer sequence followed by a more complex signal (not yet completely characterized), located some 20 to 30 bp downstream [16]. Even this signal is not trivial to predict, however, and recent studies of public expressed sequence tag (EST) databases have shown that the consensus AATAAA hexamer is absent from more than half of all 3′ untranslated regions [2•]. Thus, development of improved methods for identifying the polyadenylation signal and, in particular, promoter regions, is an important challenge for the next several years.

Translational signals

The principal translational signals that have been used in gene finding are the 'Kozak signal', located immediately upstream of the initial ATG [16], and the termination codon, useful primarily for its absence (in frame) in coding exons. Since these signals contain far too little information to allow discrimination in bulk genomic DNA, reliable prediction of translation start and stop sites may not be possible until more progress has been made towards predicting the sites of transcription initiation and termination (see above), which would dramatically reduce the amount of sequence that needs to be searched. Using simple weight matrix descriptions of the Kozak and translation termination signals in the context of the integrated gene-finding program GENSCAN [17••], about two thirds (66%) of translation initiation sites and about three quarters (78%) of termination codons have been correctly predicted (Table 1a), with specificities of 84% and 91%, respectively. Although these levels of accuracy are high enough to be useful, they are significantly lower than those achieved for splicing signals (see below), and lead to poorer prediction of initial and terminal exons than has been achieved for internal exons (Table 1b).

Splicing signals

Even if one could reliably predict promoter and polyadenylation signals, and translation start and stop sites in genomic sequences, this knowledge would generally help only in predicting the location of the first and last exons of a gene. Since most vertebrate, invertebrate and plant genes contain several exons, accurate prediction of gene structure in these organisms is much more dependent upon the ability of predictions to pinpoint splice signals. Nuclear pre-mRNA introns are excised from the primary transcript by a large ribonucleoprotein complex known as the spliceosome (reviewed in [18]), which recognizes sites at the 5′ and 3′ ends of the intron (the donor and acceptor splice sites, respectively), as well as an internal site known as the branch point. With a few interesting exceptions (see [19•]), virtually all spliceosomal introns begin with GT and end with AG, and this nearly invariant rule is used by the majority of gene-finding programs to narrow the search space of possible exon and intron boundaries.

Many early gene-finding methods used simple weight matrix (independence) models of the position-specific compositional biases present in 5′ and 3′ splice sites and of the bias towards pyrimidine nucleotides upstream of 3′ splice sites. More recently, several authors have observed statistically significant dependencies between positions within both the donor and acceptor splice sites [17••,20–22]. Certain observed dependencies between donor splice site positions can be interpreted in terms of the thermodynamics of RNA duplex formation between

Table 1

Accuracy of GENSCAN for different signal and exon types.

(a) Prediction of individual splice sites and translational signals.

Type of signal	Type of exon	Annotated exons		Predicted exons	
		Number	% Correctly predicted	Number	% Correctly predicted
Initiation	Initial only	570	66	450	84
Termination	Terminal only	570	78	487	91
5′ splice site	Initial only	570	88	450	89
5′ splice site	Internal only	1510	93	1682	89
5′ splice site	Initial and internal	2080	91	2132	89
3′ splice site	Terminal only	570	81	487	92
3′ splice site	Internal only	1510	92	1682	83
3′ splice site	Internal and terminal	2080	89	2169	85

(b) Accuracy for initial, internal and terminal exons.

Exon type	Annotated exons				Predicted exons			
	Number	% Exactly	% Partially	% Missed	Number	% Exactly	% Partially	% Wrong
Initial	570	65	25	9	457	81	9	10
Internal	1510	90	5	4	1707	80	11	8
Terminal	570	76	8	15	509	84	6	8
All types	2650	81	10	8	2678	81	10	9

Accuracy statistics are shown for forward-strand exons predicted by the GENSCAN program [17••], as tested on the Burset and Guigo dataset of 570 vertebrate gene sequences [3••]. **(a)** Accuracy is shown for four types of signals: initiation codons, termination codons and 5′ and 3′ splice sites. For each signal type, the number of (true) sites according to the GenBank CDS (coding sequence) annotation and the percentage of sites predicted correctly by GENSCAN are shown in columns 3 and 4, respectively. Columns 5 and 6 show the number of sites predicted by GENSCAN and the percentage of predicted sites that were correct, respectively. For 5′ and 3′ splice sites, accuracy data are also shown separately for initial versus internal exons, and internal versus terminal exons, respectively. **(b)** Accuracy data at the exon level. The percentages of annotated exons that were predicted exactly (both endpoints correct), predicted partially (one endpoint correct) or missed (not overlapped by a predicted exon) are listed in columns 3, 4 and 5, respectively. Columns 7, 8 and 9 show the percentages of predicted exons that were exactly correct (both endpoints correct), partially correct (one endpoint correct), or wrong (not overlapping an annotated exon), respectively. In addition, five single-exon genes were predicted by GENSCAN in this set (not given a separate row, but included in the totals).

U1 small nuclear RNA (snRNA) and the 5′ splice site region of the pre-mRNA [17••]. Of the dependencies observed for human acceptor splice sites, some appear to result simply from the compositional heterogeneity of the human genome, whereas others probably relate to the specificity of pyrimidine tract-binding proteins [23]. The development of more complex splice signal models that are capable of capturing such dependencies has been a significant recent trend in the gene-finding literature: examples include the 'maximal dependence decomposition' (MDD) and 'windowed weight array' (WWAM) models [17••], hidden Markov models [20], decision tree methods [21] and multilayer neural networks [22]. These more complex models typically yield significant, but not dramatic, improvements in splice site discrimination over the simpler models which assume independence between positions. The final level of accuracy achieved depends critically on whether prediction is measured 'in isolation' or in the context of an integrated gene-finding method (see below).

This important distinction between 'isolated' and 'integrated' splice-site prediction can be illustrated by comparing the performance of the MDD 5′ and WWAM 3′ splice site models in isolation with that achieved when these models are integrated into GENSCAN [17••]. By themselves, the specificity of these models is 34% for 3′ splice sites and 36% for 5′ splice sites at a 50% sensitivity threshold (i.e. at a threshold that identifies half of the true sites) in bulk genomic sequences [23], strongly suggesting that splice-site selection is dependent on factors other than the strength of the splice signals. This conclusion is consistent with a large amount of experimental data showing that splice-site usage is often influenced by specific exonic and intronic enhancer (and repressor) signals located some distance from the splice junctions [24]. On the other hand, when additional types of information, such as the compositional properties of exons and introns and the reading frame compatibility of adjacent exons, are incorporated into the integrated GENSCAN model, the accuracy of prediction improves dramatically. In particular, sensitivity increases to 91% of 5′ splice sites and 89% of 3′ splice sites from the Burset and Guigo data set [3••] at specificity levels of 89% and 85%, respectively (Table 1a), and fully 90% of internal exons were predicted exactly with 80% specificity (Table 1b). These latter numbers may be compared with accuracy data reported using the MZEF computer program (78% sensitivity, 86% specificity), a recently introduced method designed specifically to predict internal coding exons [25•]. Although these numbers are quite promising, it should be emphasized that the accuracy for larger genomic contigs containing multiple genes and

significant amounts of intergenic DNA is likely to be somewhat lower for both programs.

Markov sequence models

Compositional differences between coding and noncoding DNA were recognized very early on and a number of methods were developed in the mid to late eighties that used such differences in order to identify putative coding regions in genomic DNA (reviewed in [2•]). The proliferation of properties reported to be useful in identifying genes prompted Fickett and Tung [26] to undertake a systematic comparison of more than twenty different compositional properties in terms of their ability to distinguish coding from noncoding DNA. The conclusions of this study were that measures based on reading-frame-specific hexamer composition gave the best discrimination, and most subsequent gene-finding methods have used hexamer composition in one form or another.

Several recent gene-finding methods use Markov models of coding and noncoding regions in order to classify sequence segments as either exons or introns. Simply stated, a Markov model of order k captures local dependencies in sequence at the level of k+1-mers, for example, a fifth-order model reflects dependencies in hexamers. In a homogeneous model, all positions are treated the same, whereas in an inhomogeneous model, different transition probabilites are used at different positions. GENSCAN [17••] uses a homogeneous fifth order Markov model of noncoding regions, and both GENSCAN and GENMARK [7] use a three periodic (inhomogeneous) fifth order Markov model of coding regions (illustrated in Figure 1a). In the former model (illustrated in Figure 1b), the conditional probability of the identity of the next nucleotide depends on the identities of the previous five bases, thus incorporating biases in the hexamer composition. The latter model (Figure 1a) differs from this in that separate conditional probabilities are used for nucleotides that occur in the three distinct codon positions, resulting in a model that can account for the differential usage of hexamers in the two out of frame positions. This relatively complicated model thus incorporates a combination of biases related to amino acid usage, codon usage, di-amino acid and dicodon usage, as well as other factors (see also [12]).

The parameters for such models are typically estimated using the maximum likelihood method, that is, by using the observed conditional frequencies of an appropriate training set of known genes to estimate the corresponding conditional probabilities. An alternative approach, proposed recently by Salzberg *et al.* [8], is to use 'interpolated Markov model' (IMM) estimation, in which the conditional probabilities of a high order Markov chain model are estimated from an average of lower order conditional frequencies. Some authors have also used separate coding and/or noncoding region models for sequences of high and low C+G composition to account for the well known 'isochore' organization of the human genome [27,28].

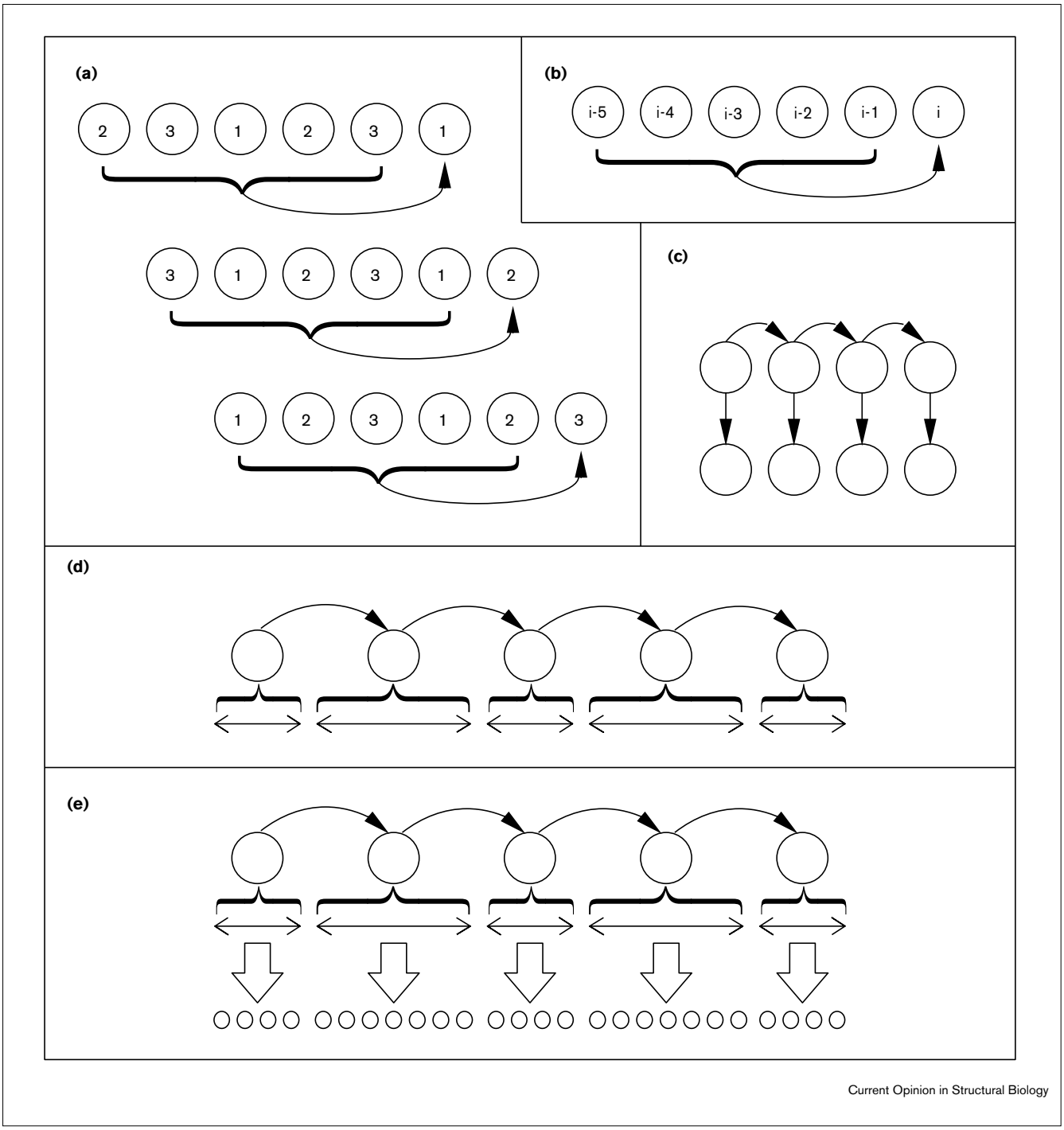
Beyond simply identifying gene components such as splice signals or coding exons in isolation, several recent methods use more elaborate models of gene architecture that require specific subcomponents of a gene to occur in the appropriate order, and allow exon and intron length distributions to be accounted for. One way to combine several different types of sequence generating models into a unified scheme is to use a 'hidden Markov model' (HMM) framework (Figure 1c). In this approach, first applied to gene finding by Haussler and colleagues [6] and also used by Henderson *et al.* [20], transitions between submodels corresponding to particular gene components are modeled as unobserved ('hidden') Markov processes (the upper circles in Figure 1c), which determine the probability of generating particular (observable) nucleotides (the lower circles in Figure 1c). The HMM architecture is in fact quite general and has been applied successfully to many problems in computational biology and in other fields [29].

All HMMs have a limitation in that blocks of the same hidden-state type (e.g. an exon considered as a block of 'coding' states) can only be modeled as geometric (exponential) random variables. Since exon and intron lengths appear to be constrained by factors related to pre-mRNA splicing (e.g. [30]), and do not exhibit geometric distributions, more general models are required to accurately account for the lengths of exons and introns in real genes. A model in which subsequent states are generated according to a Markov chain but have arbitrary (instead of fixed unit) length distributions is called a 'semi-Markov' model (e.g. [31], illustrated in Figure 1d). In the general model architecture used by GENSCAN [17••] and Genie [32•] (Figure 1e), DNA nucleotides (represented by small circles) are generated according to the probabilistic rules derived from an underlying (hidden) semi-Markov process (represented by the large circles above). Such a model structure has been described variously as an 'explicit state duration HMM' [29], a 'generalized HMM' [32•], or a 'hidden semi-Markov model' [33]. A key practical advantage of (hidden) Markov-type models is that efficient (linear time) recursions can be devised in order to determine, for example, the most likely gene structure corresponding to a given sequence [29,33,34]. Another important distinction between model architectures used in gene finding is that some programs (e.g. GENSCAN and GENMARK) use explicitly double-stranded models that allow for the occurrence of multiple genes on either or both DNA strands, whereas most others (e.g. FGENEH [35], Genie and VEIL) analyze only one strand at a time and assume that the input sequence contains a single complete gene.

Sequence similarity

Sequence similarity is a very powerful, but not infallible, type of evidence used to assign function to novel

Figure 1



The five different types of Markov models discussed in the text are illustrated. Throughout, circles represent DNA nucleotides or more general 'states' and arrows indicate dependencies. **(a)** Three periodic fifth order Markov models. Circles represent consecutive DNA bases, numbers indicate the codon position, and the arrows indicate that the next base is generated conditionally on the previous five bases and on the codon position. **(b)** Homogeneous fifth order Markov model. Circles represent consecutive DNA bases, and the arrow indicates that the next base (i) is generated conditionally on the previous five (i–5,..., i–1) in the sequence. **(c)** Hidden Markov model. The upper circles represent unobserved or 'hidden' states, perhaps corresponding to whether the position is coding or noncoding; arrows between upper circles indicate that the states are generated according to a (first order) Markov chain. The lower circles correspond to (observable) DNA bases; downward arrows indicate that each base is generated conditionally on the identity of the corresponding hidden state. **(d)** Semi-Markov model. The circles represent hidden states; single-headed arrows indicate the Markov dependence of the hidden states; double-headed arrows represent the (variable) lengths of the hidden states. **(e)** Hidden semi-Markov model. The large circles and associated arrows are as in (d). The large downward arrows indicate that the nucleotides (small circles) are generated conditionally on the identity and length of the corresponding hidden state.

sequences. In gene finding, sequence similarity can be used in at least six different ways, outlined below. First, a direct comparison of a genomic sequence with databases of expressed sequence tags (ESTs), using programs such as BLASTN 2.0 [36] and AAT [37], can identify regions of a contig that correspond to processed mRNA. Second, comparison of a genomic sequence that is translated in all six reading frames with protein sequence databases, using a program such as BLASTX 2.0 [36], can identify probable coding regions. Third, 'spliced alignment' of a genomic sequence containing a single complete gene with a homologous protein sequence, using PROCRUSTES [38], may enable reconstruction of the exon and intron organization of the gene. Fourth, comparison of predicted peptides, derived from programs such as GENSCAN or FGENEH, with protein sequence databases can be used to confirm predictions and/or to assign putative function to predicted proteins. Fifth, a comparison of a translated genomic sequence with a translated genomic or cDNA sequence using TBLASTX 2.0 [36] can identify similarities among coding regions. Finally, comparison of genomic sequences with homologous genomic sequences from closely related organisms (e.g. human versus mouse or chicken), using BLAST 2.0 [36] and pairwise alignment programs such as CLUSTAL W [39], can be used to identify conserved regions, which often correspond to coding exons or important transcriptional or splicing signals.

Each of these methods can provide useful information about gene locations, as well as clues to gene function, although similarity-based methods are currently able to identify only about half of all human genes, and this proportion is increasing rather slowly. It should also be kept in mind that similarity-based methods are only as reliable as the databases that are searched, and apparent homology can be misleading at times. For example, cDNA clones may occasionally correspond to incompletely processed messages containing one or more introns, which could lead to misclassification of a genomic segment as an exon rather than an intron. Automatic similarity-based methods may also be led astray by the inclusion of (potentially incorrect) predicted proteins in sequence databases; this problem could be easily avoided if the sequence databases would require more detailed computer readable annotations of the source of each annotated coding region (e.g. complete cDNA sequence, partial cDNA, GENSCAN and GRAIL [40] predictions, etc.). Even spliced alignment methods [38,41•] can be misleading if the target (database) protein used is not a true ortholog of the source (genomic) gene and only shares some domains, likely causing portions of the gene to be missed. One possible solution to this dilemma is to use a combination of composition-based and homology-based methods, as in the recently developed GSA program (X Huang, H Zhou, AR Kerlavage, MD Adams, personal communication), which combines database similarity information from the AAT program [37] with exons predicted by GENSCAN.

Besides sequence similarity to individual known proteins, another possible approach is to search translated genomic sequences for the occurrence of the short peptide motifs that are characteristic of common protein families, such as zinc finger motifs, ATP and GTP-binding motifs, and general PROSITE patterns. For example, some of the regular expressions characteristic of the serine protease family include: CGGS[ILMV][ILMV]; [FWY][IV][FLMV][ST]AAHC; and G[DE]SGGP[FILMV] (where [XZ] indicates the occurrence of either residue X or residue Z). For example, a genomic sequence were found to contain all three of these motifs on the same strand, and in the appropriate order, that might indicate the presence of an undetected serine protease gene. For this type of approach to become a useful tool for genome analysis will require studies of the frequencies of particular peptide patterns in (conceptually translated) noncoding genomic DNA, in order to establish appropriate benchmarks for statistical significance.

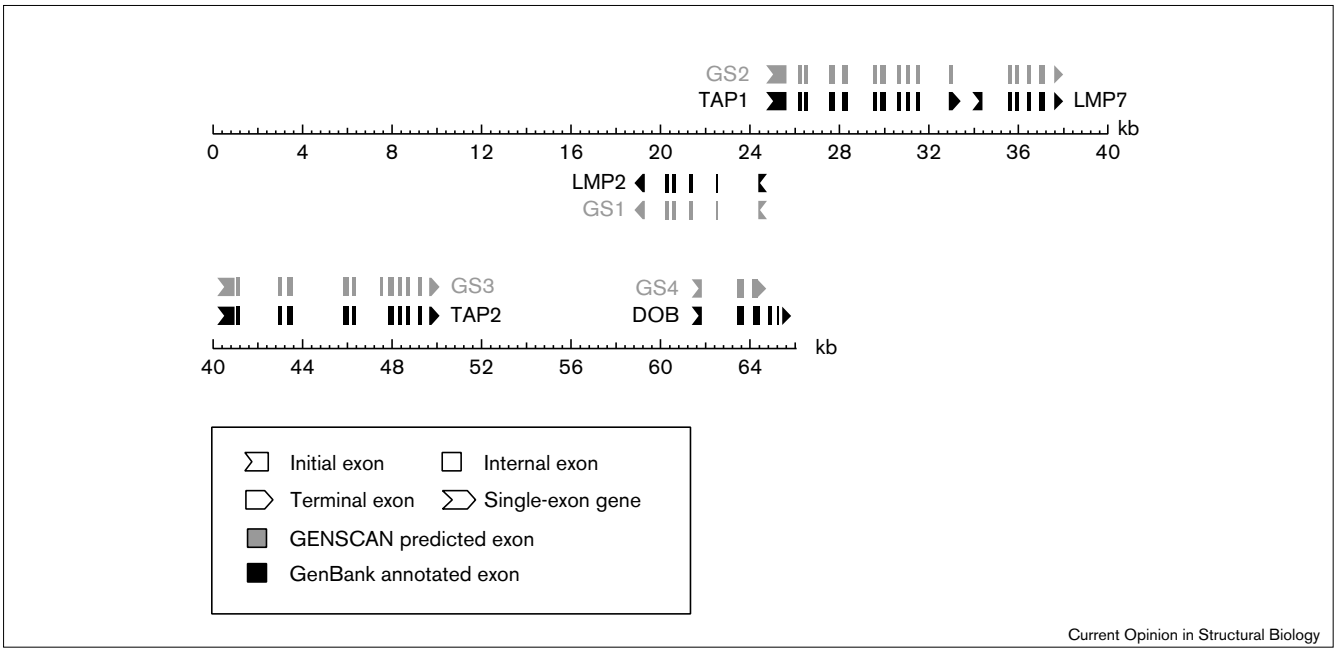
Repetitive elements

Just as sequence similarity to known proteins can help to identify probable coding regions, similarity to known classes of interspersed repeats, such as LINEs and SINEs (reviewed in [42]), can be useful in identifying probable noncoding regions. These types of elements are so abundant in the human genome that powerful repeat-finding programs such as RepeatMasker (Smit and Green, personal communication) can often classify as much as 30 to 40% of a human genomic contig as repetitive elements [42]. The presence of such repeats in the 5′ and/or 3′ untranslated regions of some cDNAs makes it absolutely essential to prescreen genomic sequences with a program such as RepeatMasker or Censor [43] before BLAST searching it against the public EST database, for example.

Open problems and future directions

Existing gene-finding programs, although signifcantly advanced over those available a few years ago, still have several important limitations. First, most programs only predict protein coding genes and not genes whose products function exclusively at the RNA level. Although specialized programs exist that identify tRNA genes [44], and rRNA genes may generally be identifiable through homology, no method has yet been developed for the identification of novel spliced or unspliced functional RNA genes such as XIST [45]. Another limitation is that no current method can deal effectively with overlapping genes in eukaryotes, and prediction of multiple genes in a single sequence is still difficult. This latter challenge is illustrated in Figure 2, which compares the GENSCAN predicted genes of a 66 kb portion of the human MHC class II region [46] with those annotated in the GenBank entry (accession number X66401). Although accuracy at the exon level is quite high (34 of 40 = 85% of annotated coding exons are predicted exactly and 34 of 37 = 92% of predicted exons are exactly correct), only one out of five genes is predicted

Figure 2



A schematic representation of predicted and annotated genes in a 66 kb portion of the human MHC II region (GenBank accession number X66401) is shown. Annotated genes are labeled according to the names given in the GenBank annotation; predicted genes are labeled GS1 through GS4 as they occur along the sequence. Genes coding on the forward and reverse DNA strands are shown above and below the sequence line, respectively. Predicted exons are shown in grey, annotated exons in black; the shape of the exon indicates its type, as shown in the key. Exon sizes and locations are drawn approximately to scale.

perfectly, and one of the predicted genes (GS2) corresponds to the fusion of exons from two annotated genes (TAP1 and LMP7 — see Figure 2).

As promoter regions often correspond to CpG islands, one approach, which might help to 'segment' a long genomic sequence into regions corresponding to single genes, would be to prescreen the sequence for CpG islands, identified using, for example, score-based statistics [47]. In this approach, scores are assigned to local sequence features (e.g. dinucleotides and trinucleotides) that are proportional to the logarithm of the ratio of their frequency in target regions (e.g. known CpG islands and coding regions) to that observed in background genomic DNA. Statistical significance thresholds are then used to define significantly high scoring regions of a sequence (e.g. the predicted CpG islands). Another approach for single-gene segmentation of a genomic region is to use the locations and polarities (DNA strands) of matches to 5′ and 3′ ESTs [48].

Finally, and perhaps most importantly, the problem of multiple protein products that correspond to a single gene through alternative splicing, alternative transcription and/or alternative translation has not yet been dealt with effectively, although some current gene-finding programs are able to predict sets of alternative exons or genes. Alternative splicing in particular is an important regulatory mechanism in higher eukaryotes, as exemplified by the elaborate regulatory cascade involved in *Drosophila* sex determination (reviewed in [18]) and by the regulation of the *fruitless* gene that is involved in courtship behavior in the male *Drosophila* [49]. Aside from a few well-studied cases, however, the rules governing alternative exon and intron choice are not well understood, presenting significant challenges to both experimental and computational biologists.

Acknowledgements

SK is supported in part by National Institutes of Health grants 5R01GM10452-33 and 5R01HG00335-10, and National Science Foundation grant DMS9403553-002.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fickett JW: **Finding genes by computer: the state of the art.** *Trends Genet* 1996, **12**:316-320.

2. • Claverie J-M: **Computational methods for the identification of genes in vertebrate genomic sequences.** *Hum Mol Genet* 1997, **6**:1735-1744.

This well-written review gives a brief history of the various methods that have been applied to computational gene identification, summarizes the methods used by current programs, and includes web addresses for most available gene finding software as well as fairly extensive references. The author also points out some of the limitations of the current methods, for example, the inability of current algorithms to handle the complexities of

overlapping genes and alternative transcription or splicing patterns, and the difficulties in predicting the beginning and end of genes.

3. •• Burset M, Guigo R: **Evaluation of gene structure prediction programs.** *Genomics* 1996, **34**:353-367.
This landmark paper provided the first large-scale, systematic, unbiased comparison of available gene-finding methods. The authors describe the construction of a large reference data set of 570 vertebrate gene sequences, critically evaluate the usefulness of a variety of predictive accuracy measures proposed previously, and introduce some new accuracy measures. They also provide the results of a systematic test of all available exon and gene prediction programs and assess the current (as of 1996) state of the gene finding problem.

4. Gelfand MS: **Prediction of function in DNA sequence analysis.** *J Comput Biol* 1995, **2**:87-115.

5. Gelfand MS: **FANS-REF: a bibliography on statistics and functional analysis of nucleotide sequences.** *Comput Appl Biosci* 1995, **11**:541.

6. Krogh A, Mian IS, Haussler D: **A hidden Markov model that finds genes in *E. coli* DNA.** *Nucleic Acids Res* 1994, **22**:4768-4778.

7. Borodovsky M, McIninch J: **GENMARK: parallel gene recognition for both DNA strands.** *Comput Chem* 1993, **17**:123-133.

8. Salzberg SL, Delcher AL, Kasif S, White O: **Microbial gene identification using interpolated Markov models.** *Nucleic Acids Res* 1998, **26**:544-548.

9. Lorenz MG, Wackernagel W: **Bacterial gene transfer by genetic transformation in the environment.** *Microbiol Rev* 1994, **58**:563-602.

10. Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A: **Evidence for horizontal gene transfer in *Escherichia coli* speciation.** *J Mol Biol* 1991, **222**:851-856.

11. Bult CJ, White O, Olsen GJ, Zhou L Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RM, Gocayne JD *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.*** *Science* 1996, **273**:1058-1073

12. Karlin S, Mrazek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262**:459-472.

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

14. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter prediction.** *Genome Res* 1997, **7**:861-878.

15. Prestridge DS: **Predicting pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **249**:923-932.

16. Kozak M: **Interpreting cDNA sequences: some insights from studies on translation.** *Mamm Genome* 1996, **7**:563-574.

17. •• Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
The authors introduce a probabilistic model for the structural and sequence compositional properties of genes in human genomic DNA and describe the application of this model to gene finding using the program GENSCAN. The model architecture employed is quite general, allowing for multiple complete or partial gene structures occurring on either or both DNA strands. The model also captures sequence properties of some of the most important *cis* elements involved in transcription, translation, and pre-mRNA splicing, as well as the length distributions of gene components such as exons and introns. The results show significant improvements in predictive accuracy over other available gene-finding methods, as measured on standard test sets of human and vertebrate genomic sequences.

18. Moore MJ, Query CC, Sharp PA: **Splicing of precursors to mRNAs by the spliceosome.** In *RNA World.* Edited by Gesteland RF, Atkins JF. Plainview, New York: Cold Spring Harbor Laboratory Press; 1993:305-358.

19. • Sharp PA, Burge CB: **Classification of introns: U2-type or U12-type.** *Cell* 1997, **91**:875-879.
The authors summarize recent research that has shown: that a very small fraction of nuclear pre-mRNA introns have AT and AC dinucleotides at their 5′ and 3′ termini, rather than the more common termini of GT and AG; that two distinct types of spliceosome, termed U2-type and U12-type, are present in both animal and plant cells; that individual introns are apparently spliced by only one type of spliceosome or the other; and that contrary to what was initially thought, the type of spliceosome used is not determined simply by the terminal dinucleotides, but instead depends on the presence or absence of specific internal consensus sequences at both the 5′ splice site and branch site of the intron. Known U12-type AT→AC introns, U2-type AT→AC introns, and U12-type GT→AG introns are tabulated and consensus patterns are described.

20. Henderson J, Salzberg S, Fasman KH: **Finding genes in DNA with a hidden Markov model.** *J Comput Biol* 1997, **4**:127-141.

21. Salzberg S, Chen X, Henderson J, Fasman K: **Finding genes in DNA using decision trees and dynamic programming.** In *Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology.* Menlo Park: AAAIPress; 1996.

22. Reese MG, Eeckman FH, Kulp D, Haussler D: **Improved splice site recognition in genie.** *J Comput Biol* 1997, **4**:311-324.

23. Burge C: **Modeling dependencies in pre-mRNA splicing signals.** In *Computational Methods in Molecular Biology*. Edited by Salzberg S, Searls DB, Kasif S. Amsterdam: Elsevier Science; 1998: 127-163.

24. Berget SM: **Exon recognition in vertebrate splicing.** *J Biol Chem* 1995, **270**:2411-2414.

25. • Zhang MQ: **Identification of protein coding regions in the human genome by quadratic discriminant analysis.** *Proc Natl Acad Sci USA* 1997, **94**:565-568.
This paper describes a program called MZEF for the prediction of internal coding exons in genomic sequences using a weighted combination of factors related to splice sites and the composition of exons and introns. The method, using quadratic discriminant analysis, is a generalization of the linear discriminant analysis approach to exon prediction used by Solovyev *et al.* [34] with the widely used HEXON/FGENEH program.

26. Fickett JW, Tung C-S: **Assessment of protein coding measures.** *Nucleic Acids Res* 1992, **20**:6441-6450.

27. Bernardi G: **The human genome: organization and evolutionary history.** *Annu Rev Genet* 1995, **29**:445-476.

28. Gardiner K: **Base composition and gene distribution: critical patterns in mammalian genome organization.** *Trends Genet* 1996, **12**:519-524.

29. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition.** *Proc IEEE* 1987, **77**:257-285.

30. Sterner DA, Carlo T, Berget SM: **Architectural limits on split genes.** *Proc Natl Acad Sci USA* 1996, **93**:15081-15085.

31. Howard RA: *Dynamic Probabilistic Systems, Volume II: Semi-Markov and Decision Processes.* New York: John Wiley & Sons; 1971.

32. • Kulp D, Haussler D, Reese MG, Eeckman FH: **A generalized hidden Markov model for the recognition of human genes in DNA.** In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology.* Menlo Park: AAAI Press; 1996.
This paper pioneered the use of 'generalized' HMMs, in which gene structure is modeled by an underlying Markov process of generalized hidden states, each of which can emit one or more nucleotides, possibly according to probabilities derived from an internal model structure such as a HMM or neural network.

33. Burge C: **Identification of genes in human genomic DNA [PhD thesis]**. Stanford: Stanford University; 1997.

34. Wu T: **A segment-based dynamic programming algorithm for predicting gene structure.** *J Comput Biol* 1996, **3**:375-394.

35. Solovyev VV, Salamov AA, Lawrence CB: **Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames.** *Nucleic Acids Res* 1994, **22**:5156-5163.

36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

37. Huang X, Adams MD, Zhou H, Kerlavage AR: **A tool for analyzing and annotating genomic sequences.** *Genomics* 1997, **46**:37-45.

38. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci USA* 1996, **93**:9061-9066.

39. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.

40. Xu Y, Mural RJ, Uberbacher EC: **Constructing gene models from accurately predicted exons: an application of dynamic programming.** *Comput Appl Biosci* 1994, **10**:613-623.

41. Sze S-H, Pevzner PA: **Las Vegas algorithms for gene recognition:**
• **suboptimal and error-tolerant spliced alignment.** *J Comput Biol* 1997, **4**:297-309.

The authors describe some variations and extensions to the 'spliced alignment' algorithm introduced by Gelfand *et al.* [38] and implemented in the PROCRUSTES program. The basic idea of PROCRUSTES is to identify the exon and intron structure of a gene in a genomic sequence by searching for the set of genomic segments (predicted exons) that maximize a global similarity measure to a pre-specified homologous protein. These homology-based methods may be extremely accurate when a sufficiently similar protein is available (e.g. human genomic DNA versus orthologous mouse protein).

42. Smit AFA: **The origin of interspersed repeats in the human genome.** *Curr Opin Genet Dev* 1996, **6**:743-748.

43. Jurka J, Klonowski P, Dagman V, Pelton P: **CENSOR - a program for identification and elimination of repetitive elements from DNA sequences.** *Comput Chem* 1996, **20**:119-122.

44. Lowe TM, Eddy SR: **tRNAcan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964.

45. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF: **The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus.** *Cell* 1992, **71**:527-542.

46. Beck S, Kelly A, Radley E, Khurshid F, Alderton RP, Trowsdale J: **DNA sequence analysis of 66 kb of human MHC class II region encoding a cluster of genes for antigen processing.** *J Mol Biol* 1992, **228**:433-441.

47. Karlin S: **Statistical studies of biomolecular sequences: score-based methods.** *Phil Trans R Soc Lond Biol* 1994, **344**:391-402.

48. Xu Y, Uberbacher EC: **Automated gene identification in large-scale genomic sequences.** *J Comput Biol* 1997, **4**:325-338.

49. Heinrichs V, Ryner LC, Baker BS: **Regulation of sex-specific selection of fruitless 5′ splice sites by transformer and transformer-2.** *Mol Cell Biol* 1998, **18**:450-458.