



The difficulty of identifying genes in anonymous vertebrate sequences*

Jean-Michel Claverie,† Olivier Poirot and Fabrice Lopez

Structural and Genetic Information Laboratory, C.N.R.S.-E.P. 91, Institute of Structural Biology and Microbiology, 31 Chemin Joseph Aiguier, Marseille 13402, France

(Received 25 July 1996; Accepted 2 December 1996)

Abstract—The identification of genes in newly determined vertebrate genomic sequences can range from a trivial to an impossible task. In a statistical preamble, we show how “insignificant” are the individual features on which gene identification can be rigorously based: promoter signals, splice sites, open reading frames, etc. The practical identification of genes is thus ultimately a tributary of their resemblance to those already present in sequence databases, or incorporated into training sets. The inherent conservatism of the currently popular methods (database similarity search, GRAIL) will greatly limit our capacity for making unexpected biological discoveries from increasingly abundant genomic data. Beyond a very limited subset of trivial cases, the automated interpretation (i.e. without experimental validation) of genomic data, is still a myth. On the other hand, characterizing the 60 000 to 100 000 genes thought to be hidden in the human genome by the mean of individual experiments is not feasible. Thus, it appears that our only hope of turning genome data into genome information must rely on drastic progresses in the way we identify and analyse genes *in silico*. © 1997 Elsevier Science Ltd

1. INTRODUCTION

A dense genetic map has now been completed for human (Dib *et al.*, 1996) and mouse (Dietrich *et al.*, 1996) genomes, and YAC-based physical maps are already available for the full human genome (Chumakov *et al.*, 1995) as well as a number of human chromosomes (Gemmill *et al.*, 1995; Krauter *et al.*, 1995; Doggett *et al.*, 1995; Collins *et al.*, 1995). To meet the last challenge of the genome project, a number of groups is now scaling-up to start sequencing the complete human genome. If everything proceeds according to plan, 3000 megabases of (mostly) anonymous genomic sequence may be at hand by the year 2005 (Jordan and Collins, 1996). Until now, most large-scale genomic sequencing efforts have been targeted around the loci of disease genes, as the final steps in positional cloning strategies. In most of the successful cases, a large experimental effort was also devoted to the gene hunt:

1. validation of the computer predictions by PCR on cDNA libraries, Northern blot, zoo blots;

2. isolation of full length cDNA; and
3. usage of concurrent experimental methods such as exon trapping or cDNA selection.

In addition, mutation analysis (using families of patients) is often used as the final test to select the right gene among putative candidates, as is phenotypic information. While computer analysis methods have been central to the identification of many vertebrate genes, a true assessment of their sensitivity, specificity, and accuracy has never been done. It is, however, already clear that none of the most popular methods is 100% sensitive, specific or accurate (see reviews by Konopka, 1994; Fickett, 1996; Claverie, 1996a; Lopez *et al.*, 1994; Brunak *et al.*, 1991; Buset and Guigo, 1996). This means:

1. that some real genes will only be partially detected;
2. that some false identifications will occur; and
3. that gene parts (e.g. promoter, transcription start site, exons, 5' end) will not always be perfectly delineated.

With a team of highly motivated molecular biologists behind them, the best computer methods currently available may appear good enough to allow the detection of most vertebrate genes. In this paper, we present a few selected examples to illustrate how lacking these methods would be when analysing a large anonymous vertebrate sequence, without the help of detailed experimental validation.

* From a lecture presented to the International Symposium on Theoretical and Computational Genome Research, 24–27 March 1996, Heidelberg, Germany.

† Author to whom correspondence should be addressed. E-mail: jmc@igs.cnrs-mrs.fr; Fax: +(33)491164549.

1.1. Methods for Locating Exons: GRAIL Analysis and Database Similarity

Coding regions in the genome exhibits certain statistical biases that have been progressively identified. The nature of these biases and the methods designed to take advantage of them have been reviewed (Konopka, 1994; Fickett and Tung, 1992). According to Fickett and Tung (1992), the most efficient methods involve the statistics of nucleotide hexamers (Claverie and Bougueleret, 1986; Borodovsky *et al.*, 1986; Brendel *et al.*, 1986; Claverie *et al.*, 1990). A popular and successful implementation of these ideas is GRAIL (Uberbacher and Mural, 1991; Xu *et al.*, 1994), a neural network combining the most discriminate coding measures and trained to recognize vertebrate genes. An increasingly efficient way of locating exons within anonymous sequences simply uses exhaustive similarity searches against public databases. Either the protein sequence database (Bairoch and Boeckmann, 1994) or the Expressed Sequence Tag (EST) (Adams *et al.*, 1991) database (Boguski *et al.*, 1993) can be used. Because the alignments between exons and database sequences are expected to be short and local, programs of the BLAST (Altschul *et al.*, 1990; Gish and States, 1993) suite are well adapted to this task (Claverie, 1992). However, the huge contribution of ubiquitous repeats (e.g. Alu) and the pathological effect of low-entropy sequences require the systematic use of masking procedures (Claverie and States, 1993; Claverie, 1994a, 1994, 1996b) (e.g. the XNU and XBLAST programs).

1.2. Methods for Detecting Vertebrate Transcription Start Sites

Full-length cDNA clones, or full-length cDNA sequences are still difficult to obtain experimentally. Computer methods capable of helping in the identification of promoter regions and transcription start sites are thus highly valuable. PROMOTER-SCAN (Prestridge, 1995) is the best known method currently available. This program uses a mixture of profile, consensus or pairwise sequence matches to RNA polymerase II promoter elements previously recognized at the 5' end of primate genes. The final diagnostic is based on the accumulation of several transcription-factor binding-site-like sequences within a sliding window of 250 nucleotides. Results similar to PROMOTERSCAN can be obtained using a succession of profile searches with a variety of position weight matrices (PWM) (Bucher, 1990) defining the relevant regulatory elements (TATA, CCAAT and GC boxes, CAP site) or many others as collected in the TRANSFAC database (Kneuppel *et al.*, 1994). Since explicit probabilities can be associated with any PWM scores (Claverie, 1994c; Claverie and Audic, 1996), they can be combined to compute the overall statistical significance of the final promoter predictions (Audic and Claverie, 1997a, in preparation).

2. THE STATISTICAL INSIGNIFICANCE OF INDIVIDUAL GENE FEATURES

Owing to the lack of rigorous statistical foundation, most methods used to identify sequence

patterns or similarities rely on empirical score thresholds to discriminate meaningful (i.e. enriched in biological significance) signals from "noise". Establishing the expected random distribution of these signals (i.e. their statistical significance) is a good way to a priori assess the quality of detection. Using a few simple examples, we will show here that the individual features upon which the recognition of eukaryote genes can be based are remarkably (statistically) non-significant.

2.1. A Simple Exercise: the Distribution of ORF in Random DNA

We define an open reading frame (ORF) as any sequence interval starting after any stop codon and ending with one. The statistical distribution of ORF lengths in random (with an equal density of A, C, T and G) DNA is simply established as follows. In a given sequence interval, let us denote $N(c)$ the number of ORFs of at least length c (in codons). The average number of ORFs of at least length $c + 1$ is given by:

$$N(c + 1) = \frac{61}{64} N(c)$$

as there are only 3 "stop" codons in the standard genetic code. The finite difference expression:

$$N(c + 1) - N(c) = -\frac{3}{64} N(c) \equiv \frac{\Delta N}{\Delta c}$$

can be approximately integrated as

$$N(c) = N(0)e^{-\frac{3}{64}c}.$$

In a random sequence of L_c codons, we expect $3L_c/64$ stop codons, each of them defining the origin of $3L_c/64$ ORFs of at least length 0, hence the value of $N(0)$. For each reading frame, the ORF length distribution is then:

$$N(c) = \frac{3L_c}{64} e^{-\frac{3}{64}c}.$$

Finally, If we neglect the interference of overlapping codons, the distribution of ORF for a given strand (e.g. 3 overlapping reading frames) simply becomes:

$$N(l) = \frac{3L}{64} e^{-\frac{l}{64}}$$

where the lengths are now expressed in nucleotides ($L = 3L_c$, $l = 3c$).

Given our crude assumptions, it is remarkable how well this formula fits the distribution of ORF as computed on actual vertebrate genomic sequences (Fig. 1). This result predicts that ORF length, in the absence of supplementary information (such as ORF similarities with already known protein sequences) will not be helpful when detecting protein coding regions in anonymous sequences (see also Fickett, 1994).

In contrast, the genes of microorganisms (bacteria, yeast) are never or rarely interrupted by introns, and

coding regions are a single span, most often beginning with the initiation codon ATG. Including this additional constraint leads to a marked reduction in the number of "candidate ORFs". Our computations on the randomized 315 341 nucleotide sequence of yeast chromosome III (Oliver *et al.*, 1992) fit the empirical formula:

$$N_{\text{atg}}(l) = \frac{L}{90} e^{-\frac{l}{50}}.$$

Thus, ORF_{ATG} (beginning with ATG) and at least 300 nucleotides long is expected to randomly occur every 36 kb on a single strand of DNA. Three hundred nucleotides (i.e. a hundred-amino-acid protein) is a threshold usually adopted for the analysis of anonymous microorganism sequences (Oliver *et al.*, 1992; Dujon *et al.*, 1994). Out of both strands of

the 10 megabase yeast genome and a total of 6000 predicted genes this might correspond to more than 500 random "hypothetical proteins". However, the distribution of ORF_{ATG} size computed on the actual yeast sequence (data not shown) is characteristic of a densely packed genome in which ORF_{ATG} at least 300 nucleotide long are 10 times more frequent than expected by chance.

2.2. The Case of Vertebrate Coding Exons

Predicting coding regions is expected to be especially difficult in vertebrate genomic sequences, where internal coding exons have no longer the constraint of beginning by ATG and are only 150 nucleotides long, on average (Hawkins, 1988; Snyder and Stormo, 1995). Bona fide candidates must now consist of an ORF flanked by reasonable splice

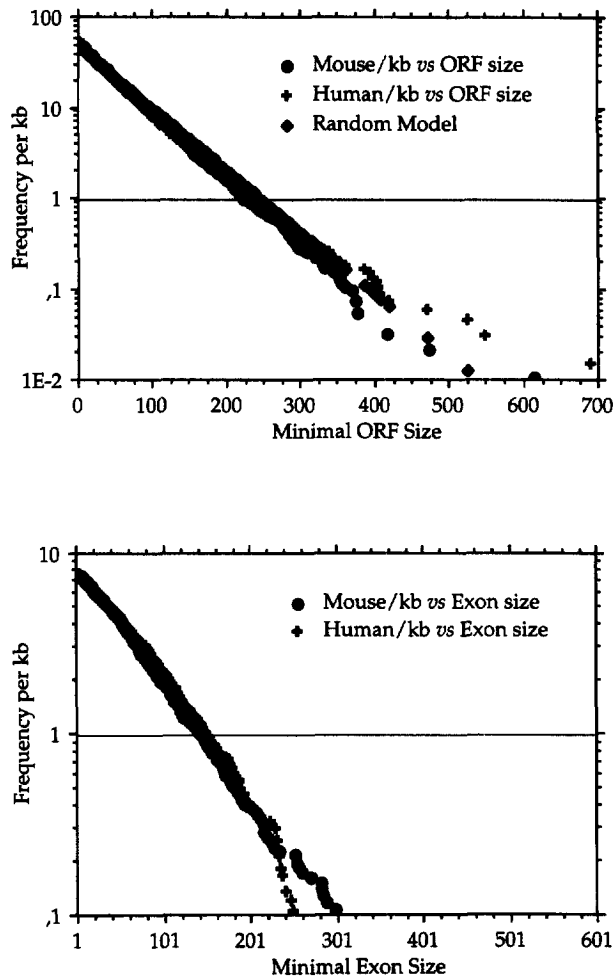


Fig. 1. Open reading frames (ORF) and candidate exons in actual sequences. Top: open reading frame (ORF) size distribution per kilobase (kb) in two unrelated long vertebrate genomic sequences. The mouse sequence is from a 94 kb contig from the X chromosome. The human sequence is from a 67 kb contig from the Xp22.3 region. Nucleotide compositions are {A: 30%, T: 27.5%, C: 21%, G:21.5%} for the murine sequence and {A: 30%, T: 31%, C: 19.5%, G:19.5%} for the human sequence. The two distributions are very close, and fit a simple theoretical model (see text) for length up to 350 nt. The random occurrence of at least one ORF of size ≥ 250 nt is expected for each kb. Actual protein coding ORFs are thus indistinguishable from random noise. Bottom: candidate exon size distribution per kilobase in the same sequences. Reasonable splice acceptor and donor consensus (Fig. 2) are now required to flank the 5' and 3' extremities of the ORF. The murine and human sequences exhibit very similar length distributions. An average of one candidate exon of size ≥ 150 nt is found for each kilobase.

acceptor and donor sites, the consensus (Senapathy *et al.*, 1990) of which can be represented by a PWM, as in Fig. 2. All candidate internal coding exons in two large mouse and human genomic sequences were identified allowing up to 4 mismatches in the pyrimidine stretch of the splice acceptor (5') site consensus (YYYYYYYY-YAG) and 2 mismatches in the positions flanking GT in the splice donor (3') site (GGTRRG) (Fig. 2). Their length distributions per kilobase are plotted in Fig. 1. For lengths below 250 nucleotides, the two distributions are nearly identical and fit the empirical formula:

$$N_{\text{exon}}(l) = 9 \times 10^{-\frac{l}{153}}$$

Thus we expect one random candidate exon with size ≥ 150 nucleotides (i.e. the average size of internal exons) in every kilobase of genomic sequence. According to reasonable estimates on gene density (one gene per 30 kb, 6 exons per gene), this would correspond to a ratio of 5 candidate exons for one real. Detecting coding exons on the sole basis of their canonical "signal" (i.e. 5' splice site-ORF-3' splice site) properties is thus impossible. Some sort of *similarity measure* (database matches, codon bias, hexamer frequencies) must be added for this purpose with, as a consequence, the loss of the universality of the method and the danger of an unknown fraction of "atypical" genes remaining undetected.

Splice Signal Position Weight Matrices

	Splice acceptor signal											
	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	10	0
C	1	1	1	1	1	1	1	1	0	2	0	0
G	0	0	0	0	0	0	0	0	0	0	0	10
T	1	1	1	1	1	1	1	1	0	1	0	0

Minimal score: 26

	Splice donor signal					
	Position					
	1	2	3	4	5	6
A	0	0	0	1	1	0
C	0	0	0	0	0	0
G	1	10	0	1	1	1
T	0	0	10	0	0	0

Minimal score: 22

Fig. 2. Position weight matrices (PWM) of vertebrate splice sites. Top: splice acceptor (5') site (YYYYYYYY-YAG). A minimal score threshold of 26 was required for the candidate exons analysed in Fig. 1. The relationships between score value, specificity and sensitivity, on one hand, and statistical significance, on the other hand, is shown in Fig. 3. Bottom: splice donor (3') site (GGTRRG). A minimal score threshold of 22 was used in Fig. 1.

2.3. The Distribution of Splice Sites in Random DNA

One might hope to enrich the predictions in favour of actual exons by increasing the required stringency on the splice signals. We have quantitatively analysed this question for the 5' acceptor site, the consensus of which (Fig. 2) contains the most information. From a standard data set of human genes prepared by Kulp and Reese (1996), we have isolated 1486 windows of 250 nucleotides each, centred on 1486 proven 5' splice sites. Using the PWM scanning program DBSITE (Claverie, 1994c) and the matrix shown in Fig. 2, we have recorded all matches with score larger or equal to a given threshold, and computed the sensitivity (proportion of actual splice sites correctly located) and the specificity (one minus the proportion of false identification) of detection. The results are plotted in Fig. 3. As previously noted (Brunak *et al.*, 1991) the threshold needed to achieve 100% sensitivity (i.e. locating all actual sites) corresponding to a specificity of 10%! (That is, 9 out of 10 identified sites are not real.) Even the requirement of the highest stringency (maximal score = 30) achieves only about 70% specificity, but with the price of a low 20% sensitivity (i.e. only 1 out of 5 sites are detected). The best compromise is found for scores equal to or larger than 29 for which approximately 50% of real sites are detected, and 1 out of 2 matched sites is not real. Clearly, locating coding exons is not an easy task.

Such a depressing result is expected from a purely theoretical standpoint. Given any PWM, we have recently shown (Claverie, 1994c) how to compute the probability of random occurrences of scores equal to or larger than a given threshold. Figure 3 (bottom) shows that 50% of any random 250-nucleotide windows are expected to match the 5' splice site PWM with a score equal to or greater than 28, and that the random probability of observing the maximal score is 1.5% (i.e. even a "perfect" match is not statistically significant at the 1% level). Thus, there is a very good agreement between the statistical significance that we can compute a priori, and the results of analysing real data. Both indicate that the window of 250 nucleotides centred on an actual splice sites does not exhibit a reliable signature. Being able to assess the a priori statistical significance of matching any signal consensus at any stringency is especially useful when experimental validations (such as for the 5' splice sites) is not practical. This is the case when only a limited set of proven examples is available, and/or when the discrimination between "false positives" (i.e. this site is never active), "false negatives" (i.e. the biological context in which this signal is active has not yet been tested) is difficult.

2.4. The Distribution of Transcription Elements in Random DNA

Experimentally identifying the 5' end of transcription units is a notoriously difficult task, and this is often a limiting step at the end of positional cloning strategies. Libraries rarely contain full-length cDNA, and the isolation, sequencing, and assembly of multiple partial clones is often necessary in the reconstruction of a transcript. One could hope that the knowledge of the genomic sequence would help

locating the transcription start sites (TSS) and even assist the detection of genes by the identification of the promoter region. Unfortunately, the situation here is as difficult as it is for identifying actual splice signals. Two of the most common elements found in vertebrate promoter regions of genes transcribed by RNA polymerase II are the TATA and CCAAT boxes. The PWMs for these two elements have been optimized by Bucher (Bucher, 1990) and are widely used. In Fig. 4, we have computed the probability distribution of scores for Bucher's TATA box matrix scanned against a random 250-nucleotide window. The theoretical probabilities have been computed for high, medium and low G + C content. The optimization process followed by Bucher (Bucher, 1990), allowed him to propose an optimal score

threshold of -8.16 , for the detection of TATA boxes. We can directly read from this figure that such a threshold is bound to detect many false positives. In fact, 70% of every 250-nucleotide random sequence window will exhibit a match at least as good as Bucher's PWM in sequences of even A:C:G:T content. This means that if a gene (with a TATA box containing promoter) is contained within a 30 kb genomic sequence, and if we have no other source of information, only 1 out of 84 TATA box signals will correspond to the actual one. In a high G + C content region, the odds will be much better, at 1 in 24. Conversely, the odds will be 1 in 120 for a low G + C genomic contig. The score threshold needed to achieve a near 100% specificity ranges from -4.7 (high G + C) to -2.2 (low G + C). However, using

**Detection of Splice Acceptor Sites:
Sensitivity, Specificity, Significance**

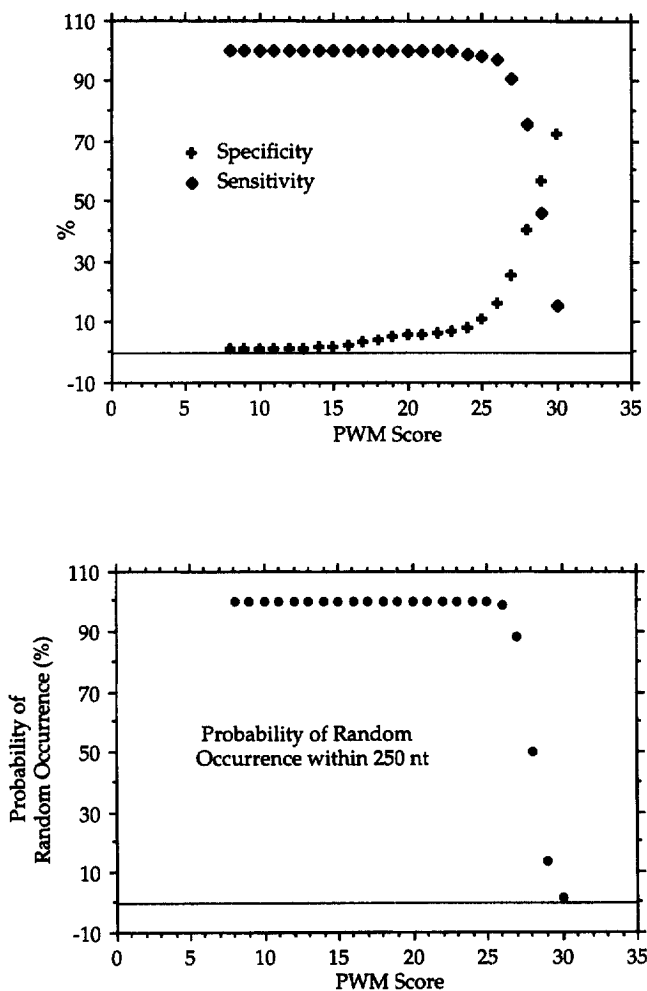


Fig. 3. Sensitivity, specificity, and statistical significance of the detection of splice acceptor sites. Top: specificity (100 - % of false detections) and sensitivity (% of actual sites detected) as a function of the score threshold. The matrix used is shown in Fig. 2. The statistics were computed using a collection of splice sites extracted from a standard set of 286 human gene sequences (Kulp and Reese, 1996). Bottom: probability of random occurrence (e.g. level of statistical significance) (PWM) scores are greater than or equal to a given value, within a range of 250 nt. The scores are derived from the PWM shown in Fig. 2.

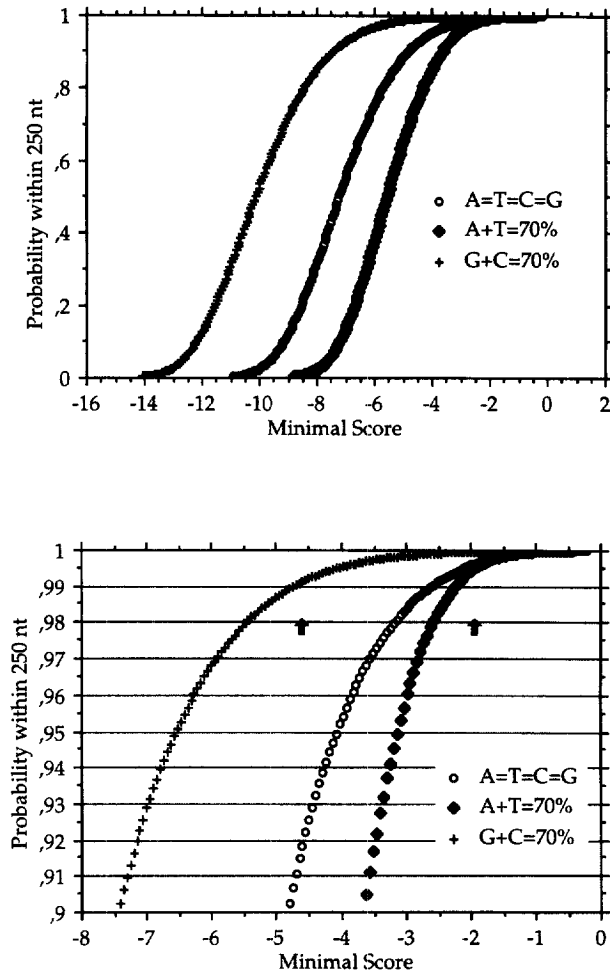


Fig. 4. Statistical significance of the TATA box transcription signal. The TATA box matrix of Bucher (1990) was used. Top: 1-probability of occurrence of scores greater than or equal to a given value, within a range of 250 nt. The influence of high, medium, and low (G + C) content on the distribution is shown. The threshold score recommended by Bucher (1990) is -8.16 , to ensure the detection of 79% of actual TATA boxes. Bottom: significant (1%) score thresholds (indicated by 2 arrows) range from -4.6 (high G + C) to -2.2 (low G + C). For high, medium and low (G + C) contents, a threshold score of -8.16 is predicted to be associated with "false positive" rates of 15%, 65% and 95%, respectively.

such thresholds will significantly decrease the sensitivity of the detection.

Figure 5 shows similar computations performed using the optimized CCAAT box PWM. In the same conditions as discussed for the TATA box, the threshold of -4.5 proposed by Bucher (Bucher, 1990) would correspond to a ratio of 1:20 between an actual CCAAT signal and a false positive. However, combining a threshold of -8.16 for the TATA box, and -4.5 for the CCAAT box will now predict (for a medium G + C content) a ratio of 1:12 between actual and false positive promoters both containing a signal.

Combining the search for many transcription signals is the principle behind current promoter detection programs such as PROMOTERSCAN. The detection of promoters combining many different transcription-factor binding sites can thus be achieved with a false positive rate of 1 for 5.6 kb (Prestridge, 1995). However, the sensitivity of these

methods is bound to decrease as their specificity increases. Furthermore, a number of "too simple" promoters (i.e. involving too few transcription elements) will forever remain undetected and undetectable by such methods. When we used PROMOTERSCAN to locate the promoter of a newly found gene (for which mouse and rat cDNA have been isolated) contained entirely in a 94 kb mouse contig (see below), three predictions were made, but none of them were located in the 5' region of the only gene detected so far in this genomic sequence.

3. GENE IDENTIFICATION USING GRAIL

3.1. A Case Study: Finding a New Gene in the Mouse *Xic* Region

One of the two X chromosomes in each somatic cell of mammalian females becomes transcriptionally

silent in early development. This mechanism compensates for the dosage difference between males (X,Y) and females (2X). A precise genetic locus, *Xic*, has been implicated in the spreading of inactivation of the X chromosome (Brockdorff *et al.*, 1992; Ballabio and Willard, 1992; Kay *et al.*, 1993). This region contains the *Xist* gene (encoding the X-inactivated specific transcript), a 15 kb "pseudo-mRNA" (i.e. not coding for a protein). As part of a systematic study of the X inactivation centre in the mouse, a 94 kb genomic sequence telomeric to *Xist* was determined and analysed in searching for new transcription units (Simmler *et al.*, 1996). This approach resulted in the identification of a new gene, and the isolation of the corresponding cDNA in both mouse and rat. On this anonymous sequence, GRAIL (Uberbacher and Mural, 1991; Xu *et al.*, 1994) was helpful in predicting some of the exons, as shown in Fig. 6. However, despite the overall success of the project, it is clear that the exon prediction was far from perfect. The many false predictions, several missed exons, as well as two errors in strand assignments, could not have been resolved without the help of many experiments.

3.2. GRAIL Performance in Detecting *Tsx* Exons

According to the GRAIL manual, the various versions (1, 1a, and 2) of the program recognize about 90% of all coding exons 100 nucleotide long or greater, when used separately. Our previous analysis of the consistency of the predictions already suggested that the overall performance of GRAIL was rather of the order of 50% (Claverie, 1996a). Figure 6(B) illustrates the performance of GRAIL during the identification of the new gene *Tsx*. Including the predictions obtained with all GRAIL versions, the success rate on the proven exons of the *Tsx* gene is 14%, or 43% if predictions overlapping two exons, but pointing to the *wrong strand*, are considered. Such a departure from the published performance is not an isolated case, and is in fact typical of the analysis of large contigs and/or high A + T content sequences (Claverie, 1996a; Lopez *et al.*, 1994; Burset and Guigo, 1996). With GRAIL as for many other programs "educated" on specific training sets, we have found with others that, for reasons that are not altogether clear, accuracy may be considerably lower than originally thought,

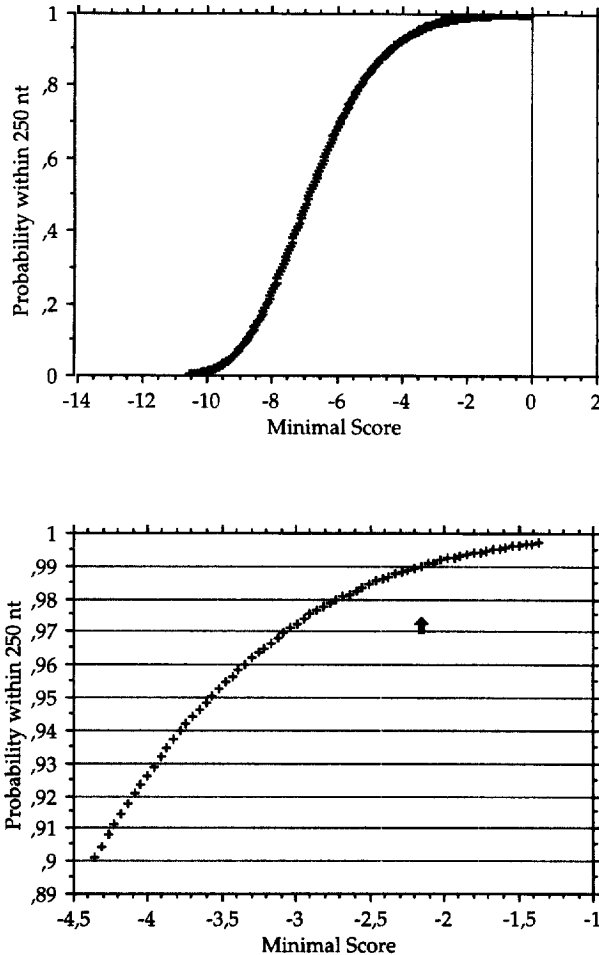


Fig. 5. Statistical significance of the CCAAT box transcription signal. The CCAAT box matrix of Bucher (1990) was used. The abundance of each nucleotide is 25%. Top: 1-probability of occurrence of scores greater than or equal to a given value, within a range of 250 nt. The threshold score recommended by Bucher (1990) is -4.54 , to ensure the detection of 87.2% of actual CCAAT boxes. Bottom: zoom on the region of higher scores. The significant (1%) score threshold is -2.2 (indicated by an arrow).

particularly on genes recently discovered. In the case of GRAIL, the use of a training set constituted of exon-dense, short (15 kb or less) genomic sequences is probably part of the problem. Typical human genes span an average of 30 kb in gene-dense regions, and may often extend over several hundred kilobases. Also, some genes with fast evolution rates may elude detection because they lack the statistical hexamer biases on which the detection is based (Claverie, 1996a). Training-dependent methods are essentially conservative, and will not be able to discover atypical genes. Without the experimental serendipity responsible for the discovery of the *Xist* pseudo-RNA,

the two *Xist* exons included in the 94 kb genomic sequence (Fig. 6(B)) would have remained forever undetected.

4. GENE IDENTIFICATION USING SIMILARITY SEARCHES

4.1. A Case Study

As the final step of a positional cloning approach to the gene responsible for Kallmann's syndrome (an X-linked defect in Gn-Rh neurone migration and olfactory neurone fasciculation in the developing

"Successful" Exon Prediction for a 94 kb genomic contig in the mouse *Xist* region

[A] Grail output

Location	Strand	Quality	Assayed	Reality overlap
5731 - 5731	f	marginal	-	
87391 - 87471	r	marginal	-	
16601 - 16621	f	marginal	-	
18221 - 18231	f	marginal	-	
21061 - 21071	f	marginal	-	
25701 - 25731	f	marginal	-	
26529 - 26561	f	excellent	+	
63941 - 63971	r	marginal	-	
61601 - 61641	r	marginal	-	
33251 - 33281	f	marginal	-	
58871 - 58901	r	good	+	-
42201 - 42251	f	good	+	-
43561 - 43601	f	marginal	-	
50681 - 50721	r	good	+	+
48545 - 48611	r	excellent	+	-
46051 - 46091	f	good	+	+(r?)
47451 - 47551	f	excellent	+	+(r?)
44721 - 44761	r	good	+	+
35581 - 35711	r	excellent	+	-
11701 - 11721	r	marginal	-	
10971 - 10971	r	marginal	-	

[B] Experimentally proven exons

Location	Strand	Grail overlap	known exons
457 - 2233	f	-	<i>Xist</i> (RNA)
5366 - 5122	f	-	<i>Xist</i> 3' exon
44625 - 44790	r	-	5' <i>Tsx</i> exon
46113 - 46213	r	-	<i>Tsx</i> , coding
46894 - 46931	r	+(r?)	<i>Tsx</i> , coding
48350 - 48435	r	+(r?)	<i>Tsx</i> , coding
50661 - 50762	r	+	<i>Tsx</i> , coding
52138 - 52205	r	-	<i>Tsx</i> , coding
54418 - 54740	r	-	3' <i>Tsx</i> exon

Fig. 6. Example of a "successful" exon identification using the GRAIL program. A 94 kb murine anonymous genomic sequence was submitted to the GRAIL E-mail server (grail@ornl.gov) using GRAIL 1, 1a and 2. A cDNA (*Tsx* transcript) was subsequently identified spanning the 44625-54740 region (Simmler *et al.*, 1996). (A) GRAIL prediction, tested candidates, and results. Three of the seven exons of *Tsx* overlapped a prediction, although two of them (indicated by "r?") were predicted on the wrong strand. (B) All identified exons (from the *Xist* and *Tsx* genes) within the contig and associated GRAIL predictions. The *Xist* exons do not encode a protein. Three of the seven protein encoding exons of the newly discovered gene are detected by GRAIL analysis (using all GRAIL versions). The protein encoded by *Tsx* has no homologue in the public databases.

brain), a 67 kb genomic sequence in the Xp22.3 region was determined (Legouis *et al.*, 1991). Starting from the identification of the last two coding exons (totalling less than 450 nucleotides, less than 0.7% of the total genomic contig!), the sequence of a full cDNA was finally assembled (Legouis *et al.*, 1991). One coding exon was identified on the basis of its similarity to a fibronectin type III domain of N-CAM L1 (see Fig. 7(A)), the other by using the hexamer coding measures (Claverie *et al.*, 1990). The two coding exons were located within 2 kb of one extremity of the contig. Once assembled, the cDNA revealed an unusually large 3' terminal untranslated exon of 3 kb, also contained in the genomic contig, separated from the last coding exon by a 5 kb intron. In 1990–1991, when this gene was first identified, no computer method could have taken advantage of the huge 3' untranslated region to locate the gene. Today, a simple BLASTN search against the EST database (Boguski *et al.*, 1993) clearly identifies this terminal exon, as shown in Fig. 7(B). Although the transcript of the Kallmann's syndrome gene is not very abundant, it is now represented by nine different ESTs.

4.2. How Good, Really, are Similarity Searches?

Database similarity searches offer the most direct way of identifying exons. However, they do suffer from their own drawbacks. The most serious problems arise from the pollution of protein and EST databases by Alu-derived sequences. These problems have been discussed in detail elsewhere (Claverie, 1992, 1994a Claverie, 1994b, 1996a, 1996b; Claverie and States, 1993; Claverie and Makalowski, 1994). Another problem is the pathological behaviour of low-entropy (protein) or "simple" (DNA) sequence segments with a local alignment program. *Ad hoc* masking of the sequences prior to a similarity search offers a satisfactory solution to these problems. However, some sensitivity is lost in the process. Some real and important proteins, such as transcription factors (Brendel and Karlin, 1989), do contain low-entropy segments (Wootton, 1994). These segments cannot serve as reliable targets for exon detection. ESTs are also contaminated by unspliced messenger RNA, or even plain genomic sequences. Since most database ESTs are generated from poly dT-primed cDNA libraries, the 3' untranslated region of the gene are becoming the most likely target for exon identification (Fig. 7(B)). However, nothing but an experiment can distinguish these matches from those caused by genomic contamination. The most confident identifications are certainly provided by the match of a translated putative exon with a protein (or the translation of an ORF) of an evolutionary distant organism (Claverie, 1996c) most likely yeast, *C. elegans*, or a bacterium. Well-conserved amino-acid sequences, encoded by highly divergent human and yeast DNA, present very strong evidence for an exon. However, too well-conserved DNA sequences should be a warning for an ever-possible contamination! What fraction of newly determined human genes do have a match in the database? In a systematic study of the ancestral gene families predating the metazoan radiation (600 million years ago), we found that most of them were already

represented in protein databases (Green *et al.*, 1993). However, those ancestral proteins might only represent about 50% of all human proteins (Claverie, 1993). While the Kallmann's syndrome gene clearly encodes several fibronectin type III domains, the new gene we identified within the 94 kb mouse contig has no significant similarity in the database, and does not correspond to an EST. At the moment, the fraction of new genes of any organism [e.g. yeast (Dujon *et al.*, 1994), or *H. influenza* (Fleischmann *et al.*, 1995)] exhibiting a significant similarity in the databases is about 50%. A similar value has been reported for the proportion of new vertebrate genes finding a match in the EST database (Boguski *et al.*, 1993, 1994).

5. DISCUSSION

In this paper, we have rapidly reviewed the main computer methods used to identify vertebrate (mostly human and mouse) genes in anonymous genomic sequences. Looking for candidate exons on the basis of open reading frames and splice sites, using GRAIL, database similarity searches or the identification of regulatory nucleotide motifs, constitute a standard set of tools. Backed with the proper experimental support, one or several of these methods are often successful in locating genes. However, we are far from the point where a completely automated computer system could take a human anonymous genomic sequence and turn it into a correctly annotated database entry. Performances are still very bad for locating promoter regions, small coding exons, coding exons of any size in A + T-rich isochores, or transcription termination signals. More importantly, no program can systematically detect transcribed regions *not coding for a protein*. This includes 5'- and 3'-untranslated regions of regular protein coding genes, as well as functional RNA such as the product of *Xist* (Brockdorff *et al.*, 1992; Kay *et al.*, 1993). As a practical consequence, determining the 5'-end of transcripts is now the limiting step in the characterization of new genes. Another remaining challenge is the prediction of alternative forms of a transcript (start, splice and end variants).

Discussions and commentaries on the success rate (and incremental progresses) of leading methods for gene detection tend to hide their fundamental limitation: to be based on the simplistic assumption that the kind of genes they are looking for are in a way similar to what has been encountered before and stored in databases. Of course, this is explicit for bona fide "similarity search" methods nowadays accounting for 30–50% of gene "discoveries" (Green *et al.*, 1993; Claverie, 1993, 1996c; Boguski *et al.*, 1994). In this context, it is worth recalling the well-documented (Claverie and States, 1993; Claverie, 1996a, 1996b; Altschul *et al.*, 1994) pathological behaviour of BLAST when dealing with sequences of "unusual" statistical properties.

"Looking for more of the same" is also the principle behind all programs educated on "training sets". This includes all leading neural-network and Markov chain-based programs. Audic and Claverie (1997b) have recently applied a simple Markov chain formalism to the detection of vertebrate RNA polymerase II promoters. Although no explicit

knowledge about promoter structure was incorporated into the program, a recognition of up to 93% of the training set could be achieved. However, the program still exhibited the deceptive characteristics of all training set-based methods: the recognition rate fell to 48% when tested on previously untested promoter regions. These numbers are comparable to what the program observed for exon detection with GRAIL (Fig. 6; Claverie, 1996a).

Prior to knowing the extent of natural variability of a given type of sequence-encoded biological signal, training set-based methods are at a great risk of confining us to a narrow and conservative view—mostly a linear combination of the already known cases. A high detection specificity can be achieved (i.e. low rate of false alarms), but the (usually unknown) sensitivity (fraction of functional signals actually detected) can be low. An alternative approach is to build detection programs from the combination of well-identified, proven features, in a rigorous probabilistic framework. We have seen here (Figs 1–5) that this approach suffers from the surprisingly low significance of these individual signals. How then does the cell manage to express correctly its genome? There are two answers to this question—one pessimistic and the other optimistic. The pessimistic view will state that only a small fraction of the genome is accessible to the transcription machinery, owing to epigenetic mechanisms (imprinting, methylation) as well as its chromatin structure (for which there is already good evidence, see Bickmore and Oghene, 1996). Sequence-based methods would thus be condemned to some degree of inaccuracy. The optimistic view is that many more sequence-encoded signals are still yet to be found, the combination of which will eventually lead to near-flawless delineation of the genomic functional domains by future programs. In any case, the spectacular successes that bioinformatics has enjoyed so far, and the praises we occasionally receive from the biologists with whom we collaborate, should not hide the fact that huge progresses have to be made to meet the challenge of interpreting the 3000 million nucleotides of the human genome without the experimental validations of each individual gene prediction.

Acknowledgements—We thank Drs Daniel Gautheret and Stéphane Audic for stimulating discussions and for reading the manuscript.

REFERENCES

- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. and Moreno, R. F. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651.
- Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nature Genetics* **6**, 119.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403.
- Audic, S. and Claverie, J.-M. (1997a) Paper in preparation.
- Audic, S. and Claverie, J.-M. (1997b) Detection of eukaryotic promoters using Markov transition matrices. *Computers and Chemistry* **21**, 223–227.
- Bairoch, A. and Boeckmann, B. (1994) The SWISS-PROT protein sequence database: current status. *Nucleic Acids Research* **22**, 3578.
- Ballabio, A. and Willard, H. F. (1992) Mammalian X-chromosome inactivation and the XIST gene. *Current Opinions in Genetic Development* **2**, 439.
- Bickmore, W. A. and Oghene, K. (1996) Visualizing the spatial relationships between defined DNA sequences and the axial region of extracted metaphase chromosomes. *Cell* **84**, 95.
- Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993) dbEST, database for “expressed sequence tags”. *Nature Genetics* **4**, 332.
- Boguski, M. S., Tolstoshev, C. M. and Bassett, D. (1994) Gene discovery in dbEST. *Science* **265**, 1993.
- Borodovsky, M. Y., Sprizhitsky, Y. A., Golavanov, E. I. and Alexandrov, A. A. (1986) Statistical patterns in primary structures of the functional regions of the genome in *Escherichia coli*. III. Computer recognition of coding regions. *Molecular Biology (Moscow)* **20**, 1390.
- Brendel, V., Beckmann, J. S. and Trifonov, E. N. (1986) Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics* **4**, 11.
- Brendel, V. and Karlin, S. (1989) Association of charge clusters with functional domains of cellular transcription factors. *Proceedings of the National Academy of Sciences of the U.S.A.* **86**, 5698.
- Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. and Rastan, S. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515.
- Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequences. *Journal of Molecular Biology* **220**, 49.
- Bucher, P. (1990) Weight matrix description of four eukaryotic RNA Polymerase II promoter elements derived from 5023 unrelated promoter sequences. *Journal of Molecular Biology* **212**, 563.
- Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics* **34**, 353.
- Chumakov, I. M., Rigault, P., Le Gall, I., Bellané-Chantelot, C., Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E. and Gros, I. *et al.* (1995) A YAC contig map of the human genome. *Nature* **377**, 175.
- Claverie, J.-M. (1992) Identifying coding exons by similarity search: Alu-derived and other potentially misleading protein sequences. *Genomics* **12**, 838.
- Claverie, J.-M. (1993) Database of ancient sequences. *Nature* **364**, 19.
- Claverie, J.-M. (1994a) Large scale sequence analysis. In *Automated DNA Sequencing and Analysis Techniques*, eds M. D. Adams, C. Fields and J. C. Venter, Chap. 36, pp. 267–279. Academic Press, New York.
- Claverie, J.-M. (1994b) A streamlined random sequencing strategy for finding coding exons. *Genomics* **23**, 575.
- Claverie, J.-M. (1994c) Some statistical properties of position/weight matrix scoring systems. *Computers and Chemistry* **18**, 287.
- Claverie, J.-M. (1996a) Progress in large scale sequence analysis. In *Advances in Computational Biology*, ed. H. Villar, Vol. 2, pp. 161–208. JAI Press, London.
- Claverie, J.-M. (1996b) Effective large scale sequence similarity searches. *Methods in Enzymology* **266**, 212.
- Claverie, J.-M. (1996c) Exploring the vast territory of uncharted ESTs. In *Genomes, Molecular Biology and Drug Discovery*, pp. 55–71. Academic Press, London.

- Claverie, J.-M. and Audic, S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Computers and Applications in Biosciences* **12**, 431.
- Claverie, J.-M. and Bougueleret, L. (1986) Heuristic informational analysis of sequences. *Nucleic Acids Research* **14**, 179.
- Claverie, J.-M. and Makalowski, W. (1994) Alu alert. *Nature* **371**, 752.
- Claverie, J.-M., Sauvaget, I. and Bougueleret, L. (1990) k-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods in Enzymology* **183**, 237.
- Claverie, J.-M. and States, D. J. (1993) Information enhancement methods for large scale sequence analysis. *Computers and Chemistry* **17**, 191.
- Collins, J. E., Cole, C. G., Smink, L. J., Garret, C. L., Leversha, M. A., Soderlund, C. A., Maslen, G. L., Everett, L. A., Rice, K. M. and Coffey, A. J. *et al.* (1995) A high-density YAC contig map of human chromosome 22. *Nature* **377**, 367.
- Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J. and Seboun, E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152.
- Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M. J., Ingalls, K. A. and O'Connors, T. J. *et al.* (1996) A comprehensive genetic map of the mouse genome. *Nature* **380**, 149.
- Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H.-C. and Marrone, B. L. *et al.* (1995) An integrated physical map of human chromosome 16. *Nature* **377**, 335.
- Dujon, B., Alexandraki, D., Andre, B., Ansoerge, W., Baladron, V., Ballesta, J. P., Banrevi, A., Bolle, P. A., Bolotin-Fukuhara, M. and Bossier, P. *et al.* (1994) Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371.
- Fickett, J. W. (1994) Inferring genes from open reading frames. *Computers and Chemistry* **18**, 203.
- Fickett, J. W. (1996) The gene identification problem: an overview for developers. *Computers and Chemistry* **20**, 103.
- Fickett, J. W. and Tung, C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Research* **20**, 6441.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A. and Merrick, J. M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* **269**, 496.
- Gemmil, R. M., Chumakov, I., Scott, P., Waggoner, B., Rigault, P., Cypser, J., Chen, Q., Weissenbach, J., Gardiner, K. and Wang, H. *et al.* (1995) A second-generation YAC contig map of human chromosome 3. *Nature* **377**, 299.
- Gish, W. and States, D. J. (1993) Identification of protein coding regions by database similarity search. *Nature Genetics* **3**, 266.
- Green, P., Lipman, D. J., Hillier, L., Waterston, R., States, D. and Claverie, J.-M. (1993) Ancient conserved regions in new gene sequences and the protein databases. *Science* **259**, 1711.
- Hawkins, J. D. (1988) A survey of intron and exon lengths. *Nucleic Acids Research* **21**, 9893.
- Jordan, E. and Collins, F. S. (1996) A march of genetic maps. *Nature* **380**, 111.
- Kay, G. F., Penny, G. D., Patel, D., Ashworth, A., Brockdorff, N. and Rastan, S. (1993) Expression of Xist during mouse development suggests a role in the initiation of X chromosome inactivation. *Cell* **72**, 171.
- Knueppel, R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994) The TRANSFAC database. *Journal of Computational Biology* **1**, 191.
- Konopka, A. K. (1994) Sequences and codes: fundamentals of biomolecular cryptology in biocomputing. In *Informat-ics and Genome Projects*, ed. D. W. Smith, pp. 119-174. Academic Press, New York.
- Krauter, K., Montgomery, K., Yoon, S.-J., LeBlanc-Straceski, J., Renault, B., Marondel, I., Herdman, V., Cupelli, L., Banks, A. and Lieman, J. *et al.* (1995) A second-generation YAC contig map of human chromosome 12. *Nature* **377**, 321.
- Kulp, D. and Reese, M. (1996) Standard data set for the prediction of genes from human DNA sequences. Available by anonymous ftp from: www-hgc.lbl.gov/in/pub/genesets.
- Legouis, R., Hardelin, J.-P., Levilliers, J., Claverie, J.-M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D. and Caterina, D. *et al.* (1991) The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. *Cell* **67**, 423.
- Lopez, R., Larsen, F. and Prydz, H. (1994) Evaluation of the exon prediction of the GRAIL software. *Genomics* **24**, 133.
- Oliver, S. G., van der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. P. and Benit, P. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38.
- Prestridge, D. S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **249**, 923.
- Senapathy, P., Shapiro, M. B. and Harris, N. L. (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods in Enzymology* **183**, 252.
- Simmler, M.-C., Cunningham, D. B., Clerc, P., Vermet, T., Caudron, B., Cruaud, C., Pawlak, A., Szpirer, C., Weissenbach, J., Claverie, J.-M. and Avner, P. (1996) A 94 kb genomic sequence 3' to the murine *Xist* gene reveals an AT rich region containing a new testis specific gene *Tsx*. *Human Molecular Genetics* **5**, 1713.
- Snyder, E. E. and Stormo, G. D. (1995) Identification of protein coding regions in Genomic DNA. *Journal of Molecular Biology* **248**, 1.
- Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in DNA sequences by a multiple sensor-neural approach. *Proceedings of the National Academy of Sciences of the U.S.A.* **88**, 11261.
- Wootton, J. C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers and Chemistry* **18**, 269.
- Xu, Y., Einstein, J. R., Mural, R. J., Shah, M. B. and Uberbacher, E. C. (1994) Recognizing exons in genomic sequence using grail II. In *Genetic Engineering: Principles and Methods*, ed. J. Setlow. Plenum Press, New York.