BIOINFORMATICS

Vol. 14 no. 1 1998
Pages 14–19

Algorithms and software for support of gene identification experiments

Sing-Hoi Sze[1,7], Michael A. Roytberg[3], Mikhail S. Gelfand[4], Andrey A. Mironov[5,6], Tatiana V. Astakhova[3] and Pavel A. Pevzner[1,2]

[1]Departments of Computer Science and [2]Mathematics, University of Southern California, Los Angeles, CA 90089-1113, USA, [3]Institute of Mathematical Problems of Biology and [4]Institute of Protein Research, Russian Academy of Sciences, Puschino, Moscow, Russia, [5]Laboratory of Mathematical Methods, National Center for Biotechnology NIIGENETIKA, Moscow, Russia and [6]ArkaGene, Inc., San Mateo, CA 94403, USA

Received on July 18, 1997; revised on September 8, 1997; accepted on September 12, 1997

Abstract

Motivation: *Gene annotation is the final goal of gene prediction algorithms. However, these algorithms frequently make mistakes and therefore the use of gene predictions for sequence annotation is hardly possible. As a result, biologists are forced to conduct time-consuming gene identification experiments by designing appropriate PCR primers to test cDNA libraries or applying RT-PCR, exon trapping/amplification, or other techniques. This process frequently amounts to 'guessing' PCR primers on top of unreliable gene predictions and frequently leads to wasting of experimental efforts.*

Results: *The present paper proposes a simple and reliable algorithm for experimental gene identification which bypasses the unreliable gene prediction step. Studies of the performance of the algorithm on a sample of human genes indicate that an experimental protocol based on the algorithm's predictions achieves an accurate gene identification with relatively few PCR primers. Predictions of PCR primers may be used for exon amplification in preliminary mutation analysis during an attempt to identify a gene responsible for a disease. We propose a simple approach to find a short region from a genomic sequence that with high probability overlaps with some exon of the gene. The algorithm is enhanced to find one or more segments that are probably contained in the translated region of the gene and can be used as PCR primers to select appropriate clones in cDNA libraries by selective amplification. The algorithm is further extended to locate a set of PCR primers that uniformly cover all translated regions and can be used for RT-PCR and further sequencing of (unknown) mRNA.*

Availability: *The programs are implemented as Web servers (GenePrimer and CASSANDRA) and can be reached at http://www-hto.usc.edu/software/procrustes/*

Contact: *ssze@hto.usc.edu*

[7]*To whom correspondence should be addressed*

Introduction

In the absence of accurate gene prediction programs, gene identification and exon annotation in genomic DNA usually amount to sequencing the corresponding mRNA. This mRNA can be found by direct screening of cDNA libraries, Northern blot analysis or hybrid selection of cDNA. The limiting factor in these techniques is non-specific hybridization (Hattier *et al.*, 1995; Timmermans *et al.*, 1996). The problem of non-specific hybridization is usually addressed by pre-hybridization to repeats and restricting the analysis to genomic DNA fragments already suspected to contain potential exons found by zoo blotting, CpG island selection, exon trapping, exon amplification, etc. (for a review, see Parrish and Nelson, 1993; Parimoo *et al.*, 1995). Comparisons of different techniques for the identification of transcribed sequences in the particular case of chromosome 17 *BRCA1* region containing more than 26 genes were carried out in Hattier *et al.* (1995) and Brody *et al.* (1995).

A serious limitation of these techniques is the low signal-to-noise ratio in hybridization experiments and high false-positive rate of splicing-based exon amplification methods (Church *et al.*, 1993; North *et al.*, 1993). Furthermore, some of the techniques cannot be applied to intronless or single-intron genes (Parimoo *et al.*, 1995). For example, these techniques failed during analysis of the *FMR2* gene in the FRAXE fragile site on the X chromosome associated with mild mental retardation (Gu *et al.*, 1996) and the *RPGR* gene mutated in X-linked retinitis pigmentosa (Meindl *et al.*,

1996). In addition, many of these methods are very labor intensive (Parrish and Nelson, 1993; Selleri *et al.*, 1995).

An alternative strategy is *in silico* gene prediction. However, no existing gene recognition algorithm provides accuracy sufficient for gene identification and exon annotation. The accuracy of gene predictions drastically depends on the availability of a related protein. If a related mammalian protein of an analyzed human gene is known, the accuracy of gene predictions in this fragment is as high as 97–99%, and it is 95, 93 and 91% for related plant, fungal and prokaryotic proteins, respectively (Gelfand *et al.*, 1996a; Mironov *et al.*, 1998; Sze and Pevzner, 1997). On the contrary, recognition of genes having no relatives in sequence databases is difficult, and the accuracy falls significantly for genes with many exons or with unusual codon usage (Burset and Guigó, 1996).

Although insufficient for final annotation, such predictions are used in practice to decrease the noise and to limit the experimental analysis to promising regions. In addition to the above examples, this approach was used, in particular, to select cDNAs of the gene for X-linked Kallmann syndrome (Legouis *et al.*, 1991), the gene for DiGeorge syndrome (Budarf *et al.*, 1995; Goldmuntz *et al.*, 1996), *Caenorhabditis elegans* muscle-specific gene *unc-89* (Benian *et al.*, 1996), as well as analysis of alternative splicing in the *Drosophila* gene *zipper* encoding non-muscle myosin II heavy chain (Mansfield *et al.*, 1996). Computer predictions were also used to perform single-strand conformation polymorphism mutation analysis in X-linked myotubular myopathy (Laporte *et al.*, 1996). Thus, the use of computer predictions in experimental practice is limited to the construction of oligonucleotide probes for Southern and Northern analyses, and the construction of PCR primers for clone detection in cDNA libraries or RT-PCR. Since the reliability of a predicted gene is not known, PCR primers for experimental gene identification are either selected at random, or guessed on top of unreliable exon predictions. This procedure often leads to wasting of experimental efforts.

This paper describes a different approach to this problem. Instead of trying to develop a universal gene prediction procedure, we use simple combinatorial techniques to make predictions needed in particular experimental schemes. We analyze open reading frames to find regions including long potential exons, thus reducing the noise level in hybridization experiments (zoo blotting, Southern and Northern analyses). The results of this step can be used for hybridization-based analyses and preliminary mutation analysis if a disease gene is studied. They also serve as the base for further processing.

If the cDNA libraries contain a clone corresponding to the analyzed gene, it can be identified more effectively not by hybridization, but by PCR selection. To do that, a biologist needs a set of candidate PCR primers guaranteed to contain a given number of primers to translated regions. Such a set is found by analyzing the coding potential of open reading frames or chains of candidate exons.

The most complicated experiments are necessary to sequence tissue-specific or low-copy mRNAs not represented in cDNA libraries. Such mRNAs are identified and amplified simultaneously by RT-PCR. We introduce the primer cover problem which models the primer selection in this case and attempts to find a small set of primers uniformly covering the (unknown) mRNA corresponding to the (known) genomic sequence. In contrast to conventional gene recognition algorithms, the primer cover algorithm helps a biologist to identify a gene experimentally without attempting to predict all exons explicitly.

Data

The algorithms were tested on a set of 257 human genomic fragments containing non-homologous complete genes (Mironov *et al.*, 1998), and on a set of 133 single-gene *Arabidopsis thaliana* DNA fragments from Korning *et al.* (1996).

Maximal open reading frames

A genomic sequence can be read in three frames (in each direction). An open reading frame is a region of a genomic sequence with no stop codon in frame. For each frame, stop codons partition a genomic sequence into non-overlapping regions called maximal open reading frames (MORFs). Each translated exon in a genomic sequence is contained inside a MORF. Although not every MORF contains an exon, the longest MORF contains an exon in 72% of cases in our human sample.

Since a random sequence in a given frame contains a stop codon approximately every 60 nucleotides, 'random' MORFs are relatively short. A typical genomic sequence contains a large number of short MORFs which are unlikely to contain exons. Our algorithm discards short MORFs (<150 nucleotides) and tries to find reliable PCR primers inside long MORFs. Although some short exons are lost after this procedure, it does not create serious problems since such exons can be recovered by PCR experiments with primers from long MORFs.

Coding windows

Given a long MORF, we would like to find a short region within this MORF that is likely to overlap with an exon. This region may be used later for PCR primer selection. A number of measures (coding potentials; reviewed in Fickett and Tung, 1992; Gelfand, 1995) correlate with the likelihood of the region being coding. We use the simplest definition of coding potential requiring only information about codon usage. Let $f(abc)$ be the frequency of the codon *abc* in the learning sample. The coding potential of a fragment consist-

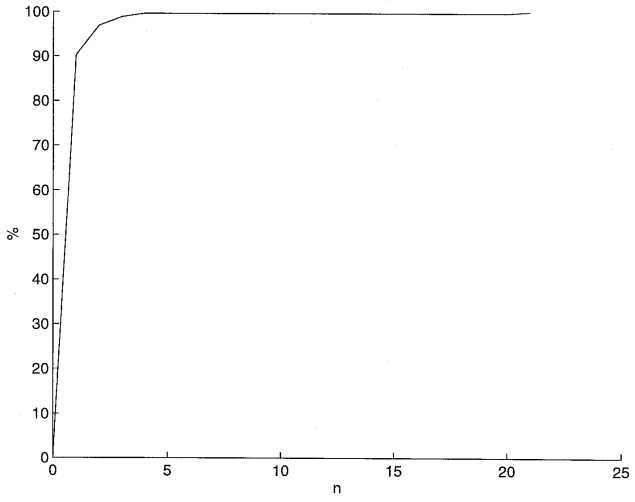

Fig. 1. Percentage of human genes when the top n windows (150 nucleotides) are sufficient to get a window that overlaps an exon.

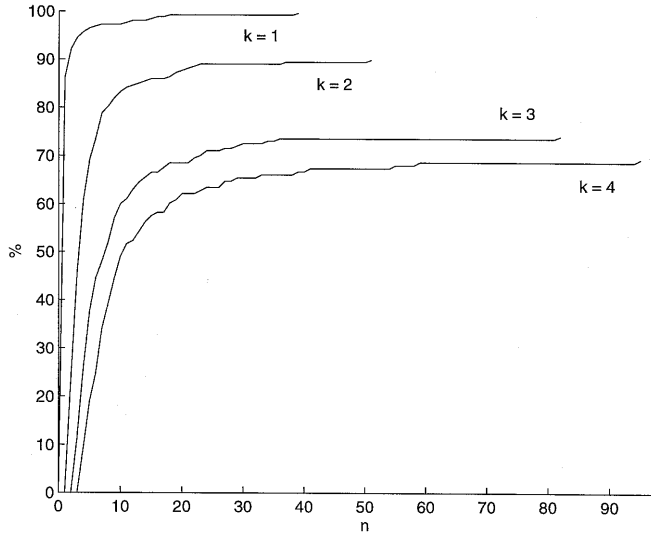

Fig. 2. Percentages of human genes (with at least k exons) when the top n primers (20 nucleotides) are sufficient to have primers inside k exons.

ing of n codons is $S(a_1b_1c_1 \ldots a_nb_nc_n) \sum_{i=1}^{n} \log f(abc)$. We select a window with the highest coding potential within a MORF and assume that this window overlaps an exon. Alternatively, a window with the largest difference between the coding potential in one frame and the maximum coding potential in the other two frames (difference criteria) can be selected. Computational experiments indicated that the difference criteria give slightly better results.

An algorithm for finding windows that are likely to overlap an exon is as follows. Given a genomic sequence, generate all long MORFs (150 nucleotides). For each such MORF, find the best window (of length 150 nucleotides) with respect to the difference criteria. Return the windows in decreasing order of their difference values. Figure 1 shows the percentage of human DNA fragments when the top n windows are sufficient to get a window that overlaps an exon. One window was sufficient in 90% of cases, whereas four windows were sufficient in all cases with one exception. The exception is the retinoblastoma gene (of 180 kb) which required 21 windows.

PCR primers

Since coding windows are likely to overlap with exons, we select the central 20-nucleotide region of a coding window as candidate PCR primers likely to be inside an exon. Figure 2 shows the percentages of human genes when the top n primers are sufficient to have primers inside k exons. There was only one case when all exons were missed by this procedure. In this case, all exons are shorter than 100 nucleotides.

In particular, for $k = 1$, the first primer was contained in an exon in 86% of cases and one of the top five primers was contained in an exon in 96% of cases. With k increasing, many more primers are needed to detect more exons and the above approach does not always find primers inside all exons, leading to exon misses. No exons were missed in 37% of cases, one exon was missed in 28% of cases, while two exons were missed in 16% of cases. In 91% of cases, no more than three exons were missed. In 99% of cases, less than eight exons were missed. There are two exceptional cases with 14 and 38 exons missing (collagen and retinoblastoma, respectively). A large number of missing exons is an indication of the complexity of the gene recognition problem (gene recognition programs frequently miss exons). This problem is overcome by designing a primer cover of genomic sequences which can be used for experimental gene identification.

Primer cover

Let P be a set of primers in a genomic sequence. Some of these primers may be contained in the corresponding mRNA (valid primers), while others may not. For a valid primer p, define *left*(p) as the valid primer preceding p in the cDNA or (if p is the leftmost valid primer) as a fictitious primer corresponding to the beginning of the cDNA. Similarly, define *right*(p) as the valid primer following p in the cDNA or (if p is the rightmost valid primer) as a fictitious primer corresponding to the end of the cDNA. Given a threshold r indicating the maximum length of potential PCR products, a set of primers P is a cover of a genomic sequence if for every valid primer from P, the distances from *left*(p) to p and from p to *right*(p) are less than r. Intuitively, in this formulation,

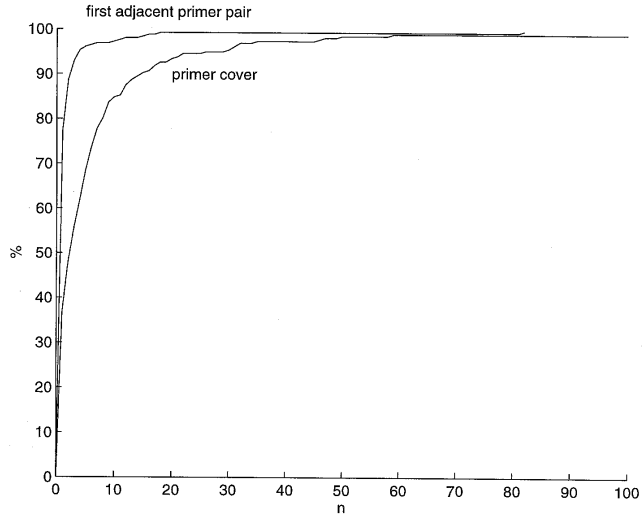

Fig. 3. Percentages of human genes when n primers (20 nucleotides) are sufficient to have a primer cover or a primer pair separated by at most $r = 500$ nucleotides (in addition to the two fictitious primers at the start and at the end of the cDNA).

primers are undirected fragments that can be used to construct a set of primers for PCR amplification, each undirected fragment corresponding to two PCR primers, one in each direction. Given a genomic sequence, the goal is to find a primer cover of minimal size. In reality, some adjustments in the positions of primers are necessary to avoid PCR artifacts.

We implemented a simple algorithm to find a primer cover. For each MORF, a set of primers is constructed as follows. Find a primer p in the middle of the best window as before. To the left of p and to the right of p, add primers every $r/2$ nucleotides as long as the primers are inside the corresponding MORF and there are no primers already put within r nucleotides. We position primers every $r/2$ nucleotides to ensure that when primers are found inside all exons, the resulting set of primers is guaranteed to form a primer cover. Consider MORFs in sorted order as before, return primers in a MORF in increasing order of the distance from p.

Figure 3 shows the percentages of human genes when n primers are sufficient to have a primer cover for $r = 500$ (in addition to the two fictitious primers at the start and at the end of the cDNA). There are only two exceptional cases when the algorithm fails to construct a primer cover. When the algorithm constructs a primer cover successfully, only one primer was needed in 37% of cases. At most eight primers were needed to cover 80% of cases, while at most 14 primers were needed to cover 90% of cases.

To find a clone corresponding to the analyzed gene in a cDNA library, biologists often use experimental protocols based on PCR amplification. In this situation, they need a relatively small set of PCR primers containing at least two primers to the cDNA in question. In other words, we are interested in the number of primers needed to get the first adjacent pair of primers separated by at most r nucleotides (primer pair problem). If a set of primers contains such a pair, then the PCR product corresponding to the primers from the pair leads to the identification of the corresponding part of the gene. There is only one case when the approach fails to find such an adjacent pair (Figure 3). Otherwise, only one primer (in addition to the two fictitious primers at the start and at the end of the cDNA) was needed to have a primer pair in 75% of cases, at most three primers were needed in 90% of cases and at most 16 primers were needed in 99% of cases.

Alternative approach for finding single probes and primer pairs

Below we describe a different approach specifically for finding single probes and primer pairs based on exon chains. Let S be a set of suboptimal exons or exon chains, and let each chain $p \in S$ be ascribed a statistical weight $R(p)$ [here we used 3-exon chains weighed by the function from Gelfand *et al.* (1996b) measuring coding potential and splice site strengths]. For each position b, let $S(b)$ be the subset of chains coming through b. The score of a candidate primer corresponding to positions $b_1 \ldots b_k$ is defined as:

$$W(b_1 \ldots b_k) = \frac{1}{k} \sum_{i=1}^{k} \sum_{p \in S(b_i)} C^{R(p)} \qquad (1)$$

where C is a constant. The candidate primers are sorted by decreasing order of their scores and a fixed number of highest scoring primers are retained with the additional requirement that the distance between the primers exceeds some given threshold.

The algorithm was tested on samples of human and *Arabidopsis* genes (Tables 1 and 2, respectively). We predicted single probes of length 150 nucleotides (a probe was accepted if at least 100 nucleotides could hybridize with the coding region of cDNA), and pairs of primers of length 30 nucleotides.

On the first sample, the highest scoring candidate probe was coding in 92% of cases and a set of five candidate probes almost always contained a coding one. It was possible to construct a primer pair in all but two cases (99%). The two exceptions are genes with one or two short exons. Two candidates were sufficient in 81% of cases and three candidates were sufficient in 89% of cases.

On the second sample, the highest scoring candidate probe was coding in 97% of cases, and three probes were sufficient in all cases. A primer pair was constructed in all but one case and it always contained no more than four candidates (two candidates in 93% of cases, three candidates in 98% of cases).

Table 1. Number of genes in each category from predictions of 257 human genes

Type of prediction	Candidates needed							
	1	2	3	4	5	6–10	>10	No prediction
Single probe	237	11	4	1	1	3	0	0
Primer pair	–	209	19	9	2	12	4	2

Table 2. Number of genes in each category from predictions of 133 *Arabidopsis* genes

Type of prediction	Candidates needed				
	1	2	3	4	No prediction
Single probe	129	3	1	0	0
Primer pair	–	124	6	2	1

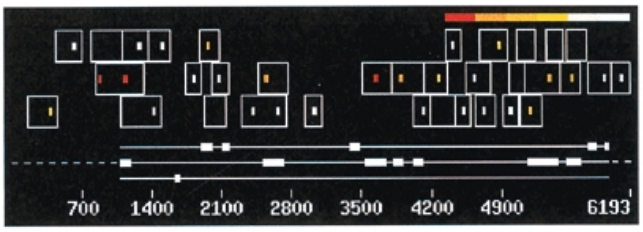

Fig. 4. Output of GenePrimer software for human gene l33842. Rectangles denote MORFs in different frames. Primers within MORFs are shown in 'successive' colors in the order of their priorities with each color showing three primers (the first color is red). The real gene structure is also shown with each exon in its respective frame.

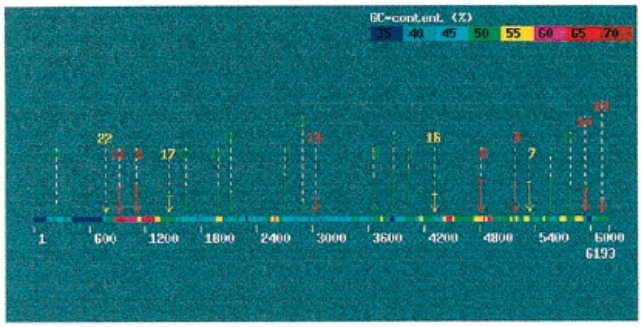

Fig. 5. Output of CASSANDRA software for human gene l33842. The predicted segments are shown as arrows pointing to their positions on the sequence line. The number above each arrow is the exon position in the candidate list. The height of an arrow's solid part is proportional to the candidate score, so that the arrow for the most probable segment is the longest one. The color of an arrow indicates self-hybridization (red) or cross-hybridization to some other segment (yellow). Green lines correspond to segments for which the program does not expect any hybridization artifacts.

Discussion

The above algorithms provide computational support for gene identification experiments. We have tested the algorithms on two samples of human and plant genes, and demonstrated that the reliability of predictions is extremely high. The programs are implemented as Web servers (GenePrimer for the primer cover problem; CASSANDRA for the alternative approach for finding single probes and primer pairs) and can be reached at http://www-hto.usc.edu/software/procrustes/. Figures 4 and 5 show sample graphical outputs of the two programs.

Of course, biological experiments are rather diverse, and the proposed algorithms do not cover all experimental approaches to gene identification. An appealing feature of the approach is its simplicity and combinatorial flexibility, making it easy to modify or optimize the program for different experimental schemes.

The algorithms described are based on two types of analyses: (i) information about long open reading frames and (ii) information about the intersection of predicted exons. Further developments can be based on merging these data in the following way. First, anchor primers are generated either at the middle of the window with the highest coding potential, or as the points where most highest scoring exons (or exon chains) intersect. Then the set of anchor primers is modified to generate a primer cover (by taking all primers within MORFs conforming to PCR requirements), a primer pair (by considering only highest scoring anchor primers), or whatever set of primers is needed.

Acknowledgements

We are grateful to Paul Hardy for many helpful comments. This work is supported by Department of Energy grant DE-FG02-94ER61919. The work of M.S.G. is also partially supported by Russian Fund of Basic Research grant

94-04-12330 and grant MTW300 from ISF and the Russian Government. M.S.G. and A.A.M. are partially supported by the Russian State Program 'Human Genome'.

References

Benian,G.M., Tinley,T.L., Tang,X. and Borodovsky,M. (1996) The *Caenorhabditis elegans* gene *unc-89*, required for muscle M-line assembly, encodes a giant modular protein composed of Ig and signal transduction domains. *J. Cell Biol.*, **132**, 835–848.

Brody,L.C. *et al*. (1995) Construction of a transcription map surrounding the *BRCA1* locus of human chromosome 17. *Genomics*, **25**, 238–247.

Budarf,M.L. *et al*. (1995) Cloning a balanced translocation associated with DiGeorge syndrome and identification of a disrupted candidate gene. *Nature Genet.*, **10**, 269–278.

Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction algorithms. *Genomics*, **34**, 353–375.

Church,D.M., Banks,L.T., Rogers,A.C., Graw,S.L., Housman,D.E., Gusella,J.F. and Buckler,A.J. (1993) Identification of human chromosome 9 specific genes using exon amplification. *Hum. Mol. Genet.*, **2**, 1915–1920.

Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **21**, 2837–2844.

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

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

Gelfand,M.S., Astakhova,T.V. and Roytberg,M.A. (1996b) An algorithm for highly specific recognition of protein-coding regions. In Akutsu,T., Asai,K., Hagiya,M., Kuhara,S., Miyano,S. and Nakai,K. (eds), *Genome Informatics 1996* (Proceedings of the 7th Workshop on Genome Informatics, December 1996, Tokyo, Japan). Universal Academy Press, Tokyo, pp. 82–87.

Goldmuntz,E., Wang,Z., Roe,B.A. and Budarf,M.L. (1996) Cloning, genomic organization, and chromosomal localization of human citrate transport protein to the DiGeorge/velocardiofacial syndrome minimal critical region. *Genomics*, **33**, 271–276.

Gu,Y., Shen,Y., Gibbs,R.A. and Nelson,D.L. (1996) Identification of *FMR2*, a novel gene associated with the *FRAXE* CCG repeat and CpG island. *Nature Genet.*, **13**, 109–113.

Hattier,T., Bell,R., Shaffer,D., Stone,S., Phelps,R.S., Tavtigian,S.V., Skolnik,M.H., Shattuck-Eidens,D. and Kamb,A. (1995) Monitoring the efficacy of hybrid selection during positional cloning. *Mamm. Genome*, **6**, 873–879.

Korning,P.G., Hebsgaard,S.M., Rouze,P. and Brunak,S. (1996) Cleaning the GenBank *Arabidopsis thaliana* data set. *Nucleic Acids Res.*, **24**, 316–320.

Laporte,J., Hu,L.J., Kretz,C., Mandel,J.-L., Kioschis,P., Coy,J.F., Klauck,S.M., Poustka,A. and Dahl,N. (1996) A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast. *Nature Genet.*, **13**, 175–182.

Legouis,R. *et al*. (1991) The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. *Cell*, **67**, 423–435.

Mansfield,S.G., Al-Shirawi,D.Y., Ketchum,A.S., Newbern,E.C. and Kiehart,D.P. (1996) Molecular organization and alternative splicing in *zipper*, the gene that encodes the *Drosophila* non-muscle myosin II heavy chain. *J. Mol. Biol.*, **255**, 98–109.

Meindl,A. *et al*. (1996) A gene (*RPGR*) with homology to the *RCC1* guanine nucleotide exchange factor is mutated in X-linked retinitis pigmentosa (RP3). *Nature Genet.*, **13**, 35–42.

Mironov,A.A., Roytberg,M.A., Pevzner,P.A. and Gelfand,M.S. (1998) Performance guarantee gene predictions via spliced alignment. *Genomics*, in press.

North,M.A., Sanseau,P., Buckler,A.J., Church,D., Jackson,A., Patel,K., Trowsdale,J. and Lehrach,H. (1993) Efficiency and specificity of gene isolation by exon amplification. *Mamm. Genome*, **4**, 466–474.

Parimoo,S., Patanjali,S.R., Kolluri,R., Xu,H., Wei,H. and Weissman,S.M. (1995) cDNA selection and other approaches in positional cloning. *Anal. Biochem.*, **228**, 1–17.

Parrish,J.E. and Nelson,D.L. (1993) Methods for finding genes: A major rate-limiting step in positional cloning. *Gene Anal. Tech. Appl.*, **10**, 29–41.

Selleri,L., Smith,M.W., Holmsen,A.L., Romo,A.J., Thomas,S.D., Paternotte,C., Romberg,L.C.R., Wei,Y.H. and Evans,G.A. (1995) High-resolution physical mapping of a 250-kb region of human chromosome 11q24 by genomic sequence sampling (GSS). *Genomics*, **26**, 489–501.

Sze,S.-H. and Pevzner,P.A. (1997) Las Vegas algorithms for gene recognition: suboptimal and error-tolerant spliced alignment. *J. Comp. Biol.*, **4**, 297–309.

Timmermans,M.C.P., Das,O.P. and Messing,J. (1996) Characterization of a meiotic crossover in maize identified by a restriction fragment length polymorphism-based method. *Genetics*, **143**, 1771–1783.