Computational identification of promoters and first exons in the human genome

Ramana V. Davuluri[1,2], Ivo Grosse[1] & Michael Q. Zhang[1]

Published online: 26 November 2001, DOI: 10.1038/ng780

The identification of promoters and first exons has been one of the most difficult problems in gene-finding. We present a set of discriminant functions that can recognize structural and compositional features such as CpG islands, promoter regions and first splice-donor sites. We explain the implementation of the discriminant functions into a decision tree that constitutes a new program called FirstEF. By using different models to predict CpG-related and non-CpG-related first exons, we showed by cross-validation that the program could predict 86% of the first exons with 17% false positives. We also demonstrated the prediction accuracy of FirstEF at the genome level by applying it to the finished sequences of human chromosomes 21 and 22 as well as by comparing the predictions with the locations of the experimentally verified first exons. Finally, we present the analysis of the predicted first exons for all of the 24 chromosomes of the human genome.

Introduction

The publication and preliminary analysis of the human genome sequence[1,2] marks a significant milestone in the field of biology. One of the main goals of the Human Genome Project is to provide a complete list of annotated genes to serve as a 'periodic table' for biomedical research[3]. The National Center for Biotechnology Information (NCBI), Ensembl and Golden Path have provided the initial annotations, but the process of annotation is expected to go on for many years. Most of the current gene annotations refer to protein-coding regions and do not provide much information about the noncoding and regulatory regions of genes. Although programs for delineating the internal coding exons of a gene (including Genscan[4], FGENES[5] and MZEF[6]) have reached a high degree of sophistication and accuracy, with a sensitivity and specificity higher than 90% at the nucleotide level[7], finding first exons—particularly noncoding exons—and promoters still remains a challenge, except where the true full-length mRNA sequences are available[8,9]. Unfortunately, most of the available mRNA sequences are incomplete at their 5′ ends and do not provide information about the first exons and promoter regions.

Traditional gene-finding programs treat the translation start site as the 5′ boundary of a gene, and there are currently no computational tools to predict the noncoding first exons or noncoding portion of a first exon. Based on our current first-exon data, approximately 40% of the human genes have completely noncoding first exons. For most of these genes, the promoter region occurs well upstream of the translation start codon (ATG) because the first intron tends to be longer than average[10]. Moreover, if the start codon occurs very close to the splice-donor site, the coding portion may be too small to be identified by currently available gene-finding programs. In these cases, we have no way of knowing where the promoter is located or where the 5′ end of the gene is likely to reside without experimental data such as full-length 5′ UTR sequences. Existing computational tools that predict DNA polymerase II promoters, such as PromoterInspector[11], are far from satisfactory, typically averaging one false positive per several thousand base pairs, with a sensitivity of approximately 50%. To fill this gap, we have developed a new program, FirstEF, dedicated to the task of identifying promoter regions and first exons in the human genome, which may also be useful for the annotation of other mammalian genomes.

Results

First-exon database

An important requirement for building a classification model using statistical pattern recognition methods is a high-quality dataset for training the model. Because there were not many GenBank records with experimentally verified first-exon annotations, we had to adopt an indirect approach to build a first-exon database. We created a collection of first exons and promoters of 2,139 known genes by mapping full-length 5′ UTRs[12] to their genomic sequences. Each of the first-exon sequences is flanked by a 5′ region (proximal promoter) 500 bp in length and a 3′ region (largely introns) 500 bp in length. Of these first-exon sequences, 1,315 (61%) were partially coding, the remaining 824 (39%) being completely noncoding. The mean length of partially coding first exons is 348 bp, whereas that of completely noncoding first exons is 151 bp.

First exons and CpG islands

Stretches of DNA sequences greater than 200 bp in length with a high G+C content and a frequency of CpG dinucleotides close to the expected value based on the mononucleotide frequencies are known as CpG islands[13]. As many human promoters are located near CpG islands[14], we classified these sequences as CpG-related and non-CpG-related based on a CpG score defined as follows:

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.* [2]*Present address: Human Cancer Genetics Program, The Ohio State University, 420 W. 12th Avenue, TMRF 524, Columbus, Ohio 43210, USA. Correspondence should be addressed to M.Q.Z. (e-mail: mzhang@cshl.org).*

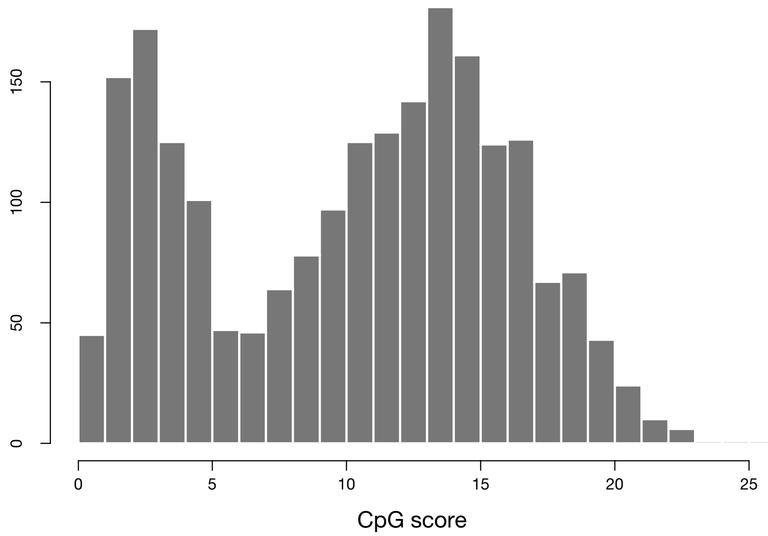

Fig. 1 Histogram of CpG scores for all of the 2,139 first exons of our first-exon database. We move a sliding window 201 bp in length along the first exon region, ranging from –500 bp upstream of the transcription start site to +500 bp downstream of the first splice-donor site. For each position of the sliding window, we compute the CpG dinucleotide percentage and define the maximum of all of the percentages as the CpG score. The CpG score has a bimodal distribution, the left mode at 3 and the right at 13 being separated by a valley at approximately 6.5. This bimodal distribution suggests that the set of first exons may stem from two distinct classes, which we denote as CpG-related and non-CpG-related. On average, CpG-related first exons have a CpG score of 13.5, whereas that of non-CpG-related first exons is 3.1. If the collection of first exons from the first-exon database were representative of the entire human genome, we could extrapolate that approximately 70% of the first exons in the human genome might be CpG-related.

we used a sliding window 201 bp in length and calculated the CpG dinucleotide percentage for each window, defining the maximum of these CpG percentages as the CpG score and the corresponding window as the CpG window.

The CpG score has a bimodal distribution (Fig. 1) that divides the set of first exons into CpG-related (with a mode of 14, a mean (μ_1) of 13.5 and a standard deviation (σ_1) of 3.4) and non-CpG-related (with a mode of 3, a mean (μ_2) of 3.1 and a standard deviation (σ_2) of 1.7). By assuming a bimodal normal distribution and an overlap of 10%, 5% from each group, we selected 6.5 as a cutoff value because it satisfied the condition $\mu_1+2\sigma_1<6.5<\mu_2-2\sigma_2$. We classified those sequences with a CpG score of 6.5 or more as CpG-related and those sequences with a CpG score of less than 6.5 as non-CpG-related; approximately 70% of the first exons in the first-exon database were therefore CpG-related.

For the CpG-related first exons, the histogram of the relative distance between the 5′ end of the CpG windows and the splice-donor site (Fig. 2) is unimodal with a mean of approximately 0.5 kb and a standard deviation of approximately 0.3 kb. The CpG window overlapped with the first exon of 76.3% of the genes in the first-exon database. If we extended the first exon by 200 bp at the 5′ end, the CpG window overlapped with the extended region ranging from –200 bp upstream of the transcription start site (TSS) to the splice-donor site of 93.8% of the genes in the first-exon database. This indicates that, in general, the CpG window overlaps with either the first exon or the proximal promoter region, even though the CpG island, typically 0.5 to 2 kb in length, may extend well upstream and/or downstream of the first exon.

First-exon finder

We developed the program FirstEF to predict the first exons and promoter regions in the human genome. FirstEF consists of different discriminant functions structured as a decision tree. The probabilistic models are designed to find potential first splice-donor sites and CpG-related and non-CpG-related promoter regions based on discriminant analysis. For every potential first splice-donor site and upstream promoter region, FirstEF decides whether the intermediate region could be a potential first exon based on a set of quadratic discriminant functions. For training and testing the different discriminant functions, we used the first exons and promoter regions from the first-exon database.

We tested the accuracy of FirstEF in two ways. First, we performed a systematic cross-validation analysis, using the data in the first-exon database; second, we ran the program on the complete sequences of human chromosomes 21 and 22 (refs. 15,16). For the cross-validation analysis, we trained the algorithm on 90% of the randomly selected data and tested it on the remaining 10%. We then estimated the sensitivity, specificity and correlation coefficient (Table 1) by repeating this process ten times. We counted predicted first exons as true positives if the predicted first splice-donor sites were identical to the real first splice-donor sites and the predicted TSSs fell within the region between –500 and +200 around the real TSS. Pseudo-exons (see Methods) predicted as first exons were counted as false

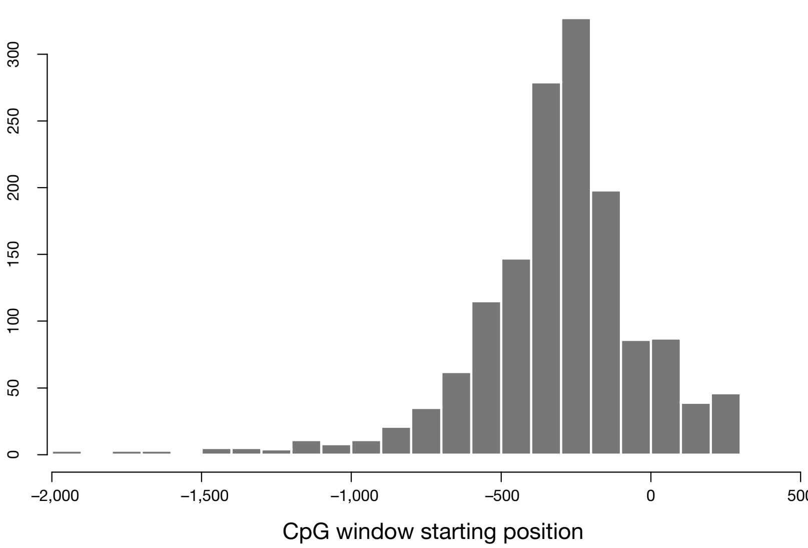

Fig. 2 Histogram of the relative distance between the 5′ end of the CpG window and the splice-donor site for CpG-related first exons. A negative value for the relative distance (plotted along the abscissa) indicates that the 5′ end of the CpG window lies upstream of the splice-donor site, whereas a positive value indicates that it lies downstream of the splice-donor site. We found that 76.3% of the first exons overlapped with the CpG window, and that for 93.8% of the first exons, the CpG window overlapped with the region ranging from –200 kb of the transcription start site to the splice-donor site.

positives, missed first exons were counted as false negatives and missed pseudo-exons were counted as true negatives. We found that FirstEF predicted 92% of the CpG-related first exons with 4% false positives and 74% of the non-CpG-related first exons with 40% false positives. Overall, FirstEF predicted 86% of all of the first exons with 17% false positives.

Table 1 • Accuracy of FirstEF based on cross-validation

Exon type	S_n[a]	S_p[b]	CC[c]
CpG-related	0.92	0.97	0.94
not CpG-related	0.74	0.60	0.65
all exons	0.86	0.83	0.83

[a] S_n (sensitivity) = TP/(TP+FN), [b] S_p (specificity) = TP/(TP+FP),

[c] CC (correlation coefficient) = $\dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

where TP, TN, FP and FN denote number of true positives, true negatives, false positives and false negatives, respectively. We find that the accuracy of FirstEF is significantly higher for CpG-related than non-CpG-related genes.

To study the performance of FirstEF on the genome scale, we ran it on human chromosomes 21 and 22 and compared the predictions with the experimentally verified first exons. We first performed extensive searches of GenBank and collected all full-length mRNAs and promoter sequences of the genes on chromosomes 21 and 22. Next, we downloaded the assembled sequences (both the original and the repeat-masked sequences from the 12 December 2000 dataset) of chromosomes 21 and 22 from the UCSC genome server (http://genome.ucsc.edu). We define 'repeat-masked sequences' as those in which known families of repeats were masked by the computer program RepeatMasker (http://ftp.genome.washington.edu). We aligned the mRNA sequences to the chromosomes by using the local alignment program BLAT (http://genome.ucsc.edu/). We identified 121 first exons and promoter regions on both chromosomes (42 on chromosome 21 and 79 on chromosome 22). FirstEF predicted 106 (88%) of these first exons (37 on chromosome 21 and 69 on chromosome 22); (Web Tables A and B). The novel feature of FirstEF is its ability to identify completely noncoding first exons, some of which occur well upstream of the annotated translation start codon. Of the 121 experimentally verified first exons, 42 were completely noncoding; FirstEF predicting 33 (79%) of these.

Because FirstEF also predicts a proximal promoter 570 bp in length (see Methods) as the 5′ boundary of the first exon, we compared the promoter-prediction accuracy of FirstEF with that of PromoterInspector, the best currently available promoter recognition program[11]. As PromoterInspector is a commercial software package and its use is restricted, we could analyze only a set of 58 randomly selected genomic sequences. Each genomic sequence consisted of an experimentally verified first exon flanked by sequences of 20 kb at both ends. Using the same criteria as Scherf *et al.*[11], we counted a predicted promoter region as true positive if the transcription start site was located within or up to 200 bp downstream of the predicted promoter region; otherwise, we considered the promoter region to be a false positive. PromoterInspector predicted the location of promoters with 48% sensitivity and 43% specificity (Table 2), which is consistent with the published results[11]. Using the same criteria, the sensitivity and specificity of FirstEF were 79% and 54%, respectively.

Annotation of first exons on human chromosomes 21 and 22

Although extensive gene annotations are available from the EBI and UCSC human genome servers, most annotations do not contain information about noncoding exons and regulatory regions. To demonstrate the efficacy of FirstEF in finding first exons and to provide the annotation of potential first exons and promoters, we used the finished sequences of chromosomes 21 and 22. We downloaded release 2.3 (6 March 2001) of the annotated gene transcripts from the Sanger Center Chromosome 22 web server (http://www.sanger.ac.uk/HGP/Chr22) and Chromosome 21 Sequencing Consortium (http://eri.uchsc.edu/chromosome21). We aligned these transcripts to their respective chromosomes using BLAT and identified the coding regions of the genes.

We scanned a 15-kb region upstream of each gene and localized putative first exons and promoter regions using FirstEF predictions. FirstEF reports all those first exons with a donor probability of 0.4 or greater, a promoter probability of 0.4 or greater and a first-exon probability of 0.5 or greater (see Methods). We post-processed the output of FirstEF by selecting the first exon that had the maximum *a posteriori* probabilities of exon, donor and promoter in that order. For chromosome 21, FirstEF predicted first exons for 141 of 218 known genes, 3 of 5 pseudogenes and 36 of 46 predicted genes (Table 3). For chromosome 22, FirstEF predicted 322 of the 341 known mRNA genes, 103 of the 152 pseudogenes, 88 out of 112 related genes, 88 out of 109 predicted genes and 18 out of 118 gene segments (Table 3). FirstEF missed more first exons on chromosome 21 than on chromosome 22, which may be because there are fewer CpG-related first exons on chromosome 21 than on chromosome 22. The relative locations of promoters from the annotated translation start codon are shown in Fig. 3.

Annotations of first exons in the human genome

We ran FirstEF on the assembled sequences (1 April 2001 GenBank freeze) of each of the 24 chromosomes that we downloaded from the UCSC Human Genome Project working draft (http://genome.ucsc.edu). FirstEF can predict alternative first exons and ranks them based on *a posteriori* probabilities; if two consecutive predictions are separated by fewer than 1,000 nucleotides, FirstEF treats them as one cluster of alternative first exons that belong to the same gene. FirstEF predicted 68,645 first-exon clusters in the human genome (32,786 on the Watson strand and 35,859 on the Crick strand), of which 39,643 were CpG-related and the remaining 29,002 non-CpG-related. The chromosomal positions of the predictions of first exons, promoters and associated CpG windows are available at http://www.cshl.org/mzhanglab.

Table 2 • Promoter prediction accuracy of PromoterInspector and FirstEF

Program	True positives	False positives	Sensitivity (%)	Specificity (%)
PromoterInspector	28	37	48.3	43.1
FirstEF	46	40	79.3	53.5

Comparison of the promoter prediction accuracy of PromoterInspector and FirstEF on 58 randomly chosen first exons with experimentally verified TSS. Approximately 70% of those 58 first exons are non-CpG-related.

Discussion

The human genome contains a vast number of *cis*-regulatory elements responsible for directing the spatial and temporal patterns of gene expression in response to metabolic requirements, developmental programs and external stimuli[17]. The localization of these regulatory regions is important for understanding large-scale gene expression data, such as those from microarray experiments. Moreover, precisely identifying the 5′ boundaries and noncoding exons of genes in higher eukaryotic genomes has been a challenge of bioinformatics for years[18]. Lack of high-quality data has delayed the development of computational tools to predict first exons, especially noncoding exons. The Eukaryotic Promoter Database[19] (http://www.epd.isb-sib.ch), which holds information on previously characterized promoter sequences, contains only 273 human promoters. The large collection of high-quality data for this study enabled us to classify the set of first exons into CpG-related and non-CpG-related first exons. We used this classification to build a combination of multiple discriminant models for FirstEF. Recent research results in statistical pattern recognition[20] demonstrate the effectiveness of combining multiple models of the same or different types for improving the accuracy of predictive modeling.

Owing to the phenomenon of CpG suppression in mammalian genomes, CpG dinucleotides account for only about 1% of the human genome. There are approximately 50,000 CpG islands in the human genome and 29,000 in the repeat-masked human genome, most of which reside near the first exons of genes. Earlier studies on CpG islands used the definition of Gardiner-Garden and Frommer[13], which required the G+C content to be greater than 50%. Motivated by the fact that GC-poor genomes maintain unmethylated CpG islands[21,22], we considered only the CpG dinucleotide percentage and did not put any restriction on the G+C content in classifying CpG-related versus non-CpG-related first exons. Notably, all of the CpG-related first exons in the first-exon database have GC-rich (G+C content greater than 50%) CpG windows, except the first exon of *LOH11CR2A* (GenBank accession number NM_014622), which is a putative tumor-suppressor gene.

We considered a first exon to be a genomic region bordered by a promoter region (5′ boundary) 570 bp in length and the first splice-donor site (3′ boundary). We have developed a set of discriminant functions that recognize CpG islands, promoter regions and first splice-donor sites, which we implemented in FirstEF. Important features of FirstEF include: (i) the capacity to predict both partially coding and completely noncoding first exons; (ii) the use of distinct models to capture differences in CpG-related and non-CpG-related first exons; and (iii) the use of a specific model for identifying first splice-donor sites that was exclusively trained on first-exon data.

The accuracy of FirstEF in predicting CpG-related first exons (70% of the first exons in the human genome) was very high, with a specificity and sensitivity greater than 90%. The CpG score and the location of the CpG window relative to the first splice-donor site contributed to the more accurate prediction of CpG-related first exons. The accuracy is lower for non-CpG-related first exons because the statistical composition of these is similar to that of

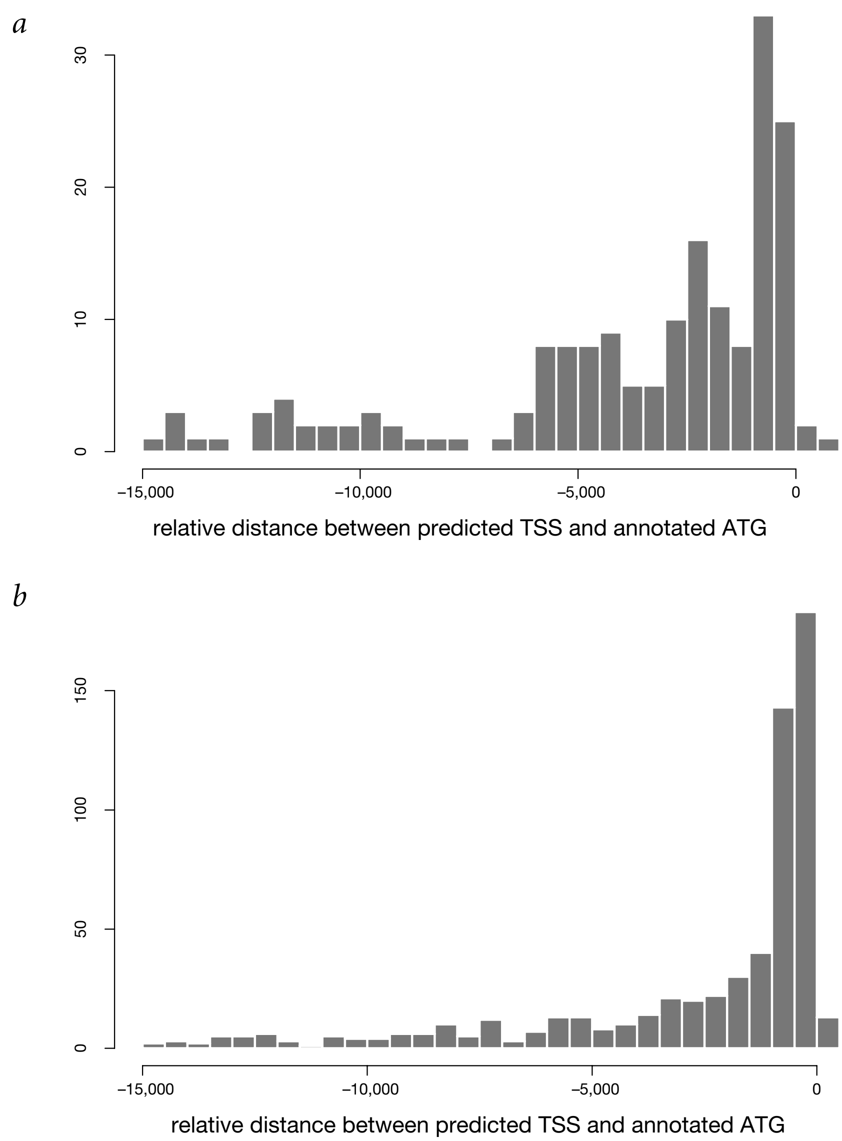

Fig. 3 Histograms of the relative distance between the 5′ end of the predicted first exon (TSS) and the annotated translation start codon (ATG) for human chromosome 21 (Fig. 3a) and human chromosome 22 (Fig. 3b). Negative values for the relative distance (plotted along the abscissa) indicate that the predicted TSS lies upstream of the annotated ATG, whereas positive values indicate that it lies downstream of the annotated ATG. Although positive values for the relative distance should in principle not occur (as the TSS should always lie upstream of the translation start codon ATG), there are three reasons why this could happen: (i) there might be alternative transcription or translation start sites; (ii) the prediction of the location of the promoter region and hence the TSS is not 100% accurate and may be wrong; and (iii) the annotation of TSSs is not 100% accurate and may be wrong. We find that nearly 60% of promoters lie within the first 1,000 bp upstream of the annotated start codon, with 5% lying more than 10 kb upstream of the annotated start codon. The two bars around 0 indicate the number of the predicted promoter regions 570 bp in length that overlap with the annotated first exon.

Table 3 • Predicted first exons for chromosomes 21 and 22

Gene type*	Number of annotated genes	Number of predicted first exons	Number of CpG-related first exons
Chromosome 21			
mRNA genes	218	141 (65%)	126 (58%)
pseudogenes	5	3 (60%)	1 (20%)
predicted genes	46	36 (78%)	32 (70%)
total genes	269	180 (67%)	159 (59%)
Chromosome 22			
mRNA genes	341	322 (94%)	299 (88%)
pseudogenes	152	103 (68%)	92 (61%)
related genes	112	88 (79%)	71 (64%)
predicted genes	109	88 (81%)	75 (69%)
gene segments	118	18 (15%)	4 (3%)
total genes	832	619 (74%)	541 (65%)

*Based on chromosome 21 and 22 sequencing consortium classifications.

introns and intergenic DNA. The accuracy of FirstEF was also demonstrated by its correct prediction of the experimentally verified first exons on chromosomes 21 and 22 (Web Tables A and B). In addition, we found that FirstEF could predict promoters with greater accuracy than PromoterInspector.

We annotated first exons and promoter regions on chromosomes 21 and 22 using FirstEF and the coding-region annotations of the public consortium. We initially ran FirstEF on repeat-masked sequences to avoid false positives, but this method missed several first exons near the repeat regions. We then ran FirstEF on the original (that is, non–repeat masked) sequences and analyzed the output of FirstEF in the 15-kb region upstream of the translation start codon. This strategy worked better, and we annotated most of the first exons and promoter regions of chromosomes 21 and 22. In practice, we recommend using a combination of gene-finding programs, such as Genscan and MZEF, on repeat-masked sequences to identify potential coding regions, and then applying FirstEF to sequences in which repeats have not been masked to annotate the first exons and proximal promoter regions. In the final step, the application of core-promoter recognition programs, such as Core-Promoter[23], may further localize potential transcription start sites.

The output of FirstEF (available at http://www.cshl.org/mzhanglab) on the working draft of the entire human genome could possibly stimulate the experimental exploration of putative false-positive results. There are approximately 4% false positives in CpG-related and 40% false positives in non-CpG-related first exons predicted by FirstEF. We would therefore expect approximately 1,586 CpG-related and approximately 11,601 non-CpG-related false predictions out of the total of 68,645 predictions in the human genome. These values might be small enough for experimentalists to test all of the false positives for expression.

Methods

First-exon database. We previously extracted, classified and characterized human full-length 5′ UTR sequences[12]. Here, we aligned mRNAs with full-length 5′ UTRs to the genomic sequences with the local alignment program BLAST. We retrieved the first exons and their flanking regions 500 bp in length for each gene from their respective genomic sequences. We eliminated redundant and ambiguous sequences, thereby obtaining a set of 2,139 first exons flanked by upstream regions 500 bp in length containing all or part of the proximal promoters and downstream regions 500 bp in length containing all or part of the first introns.

FirstEF algorithm. Two major obstacles to detecting the first exons and promoter regions are the low signal-to-noise ratio and the heterogeneous nature of the data. Hence, not all first exons can be considered as a single class, and a general model to differentiate real first exons from pseudo first exons will not work. We have thus built different classification models for different classes of first exons and incorporated these models into a decision tree. The major steps involved in the algorithm are as follows:

FirstEF scans the input sequence for potential first splice-donor sites (GT). During this step, the program computes, for every GT, the *a posteriori* probability of the splice-donor site given GT, $P(\text{donor site}|\text{GT})$, by a quadratic discriminant function (donor QDF), which was trained on the splice-donor sites of the first-exon database. If $P(\text{donor}|\text{GT}) \geq 0.4$, FirstEF considers this to be a candidate splice-donor site.

For every candidate splice-donor site, FirstEF scans a region 2,000 bp in length (1,500 bp upstream and 500 bp downstream of GT) for the existence of a CpG window with a CpG score of 6.5 or greater. Depending on the presence or absence of a CpG window, FirstEF decides during this step whether the first exon is CpG-related or non-CpG-related.

FirstEF uses a sliding window 570 bp in length (considering the first 500 bp (positions –500 to –1) to be proximal promoter region upstream of TSS and the following 70 bp (positions +1 to +70) to be downstream of TSS) within the region 1,500 bp upstream of the candidate splice-donor site. FirstEF decides whether the sliding window might or might not be a promoter based on the *a posteriori* probability of being a promoter given the window, $P(\text{promoter}|\text{window})$. This was evaluated using two different quadratic discriminant functions (promoter QDF), one for CpG-related and the other for non-CpG-related first exons.

If $P(\text{promoter}|\text{window}) > 0.4$, FirstEF matches the promoter region with the corresponding splice-donor site and evaluates the *a posteriori* probability of being an exon given the promoter and splice-donor site, $P(\text{exon}|\text{all})$, by using four different quadratic discriminant functions (first-exon QDFs). FirstEF reports all those exons with $P(\text{exon}|\text{all}) > 0.5$, along with the promoter region and, if it exists, the CpG window.

For a binary decision problem, it is a normal convention to select 0.5 as the *a posteriori* probability cutoff value, but we selected 0.4 as the cutoff probability for the donor in step 1 and for the promoter in step 4 in order not to miss marginal cases. The main purpose of these two steps is to filter the candidate donors and promoters for the first-exon QDF in order to speed up the computer program. Because the first-exon QDF considers all the feature variables of the donor, promoter and exon, the arbitrary selection of the donor and promoter *a posteriori* probabilities does not significantly affect the output of FirstEF.

Quadratic discriminant analysis. We performed the characterization of splice-donor sites, promoter regions and first exons by quadratic discriminant analysis[6,24]. FirstEF uses three different QDFs (donor QDF, promoter QDF and first-exon QDF) with different sets of variables. The QDF variables were obtained by experimenting with many scoring measures based on hexamer, pentamer and trimer frequencies, the CpG and G+C percentage and the presence of palindrome structures.

Discriminant functions for splice-donor site recognition. We tried many different scoring functions that incorporated various characteristics of first splice-donor sites and used the best scoring functions that discriminated first-exon splice-donor sites from pseudo-sites for building the QDFs. We assumed the position of GT in a sequence to be +1, using two different QDFs depending on whether the sequence window (1–200) was GC-rich (G+C 52% or greater) or GC-poor (G+C less than 52%). The donor QDFs use the following four variables: (D1) splice-donor site conditional weight matrix score; (D2) hexamer score in the window

(1–200); (D3) hexamer score in the window (–200 to –1) and (D4) trimer score in the window (1–64). All the *n*-mer scores are weighted averages of the *n*-mer frequencies, the weights being the log-likelihood ratios of the *n*-mer frequencies of the first exons divided by the *n*-mer frequencies of the pseudo-exons. Conditional weight matrices of splice-donor sites were calculated based on a subclassification of splice-donor sites similar to the maximal dependence method explained in Burge and Karlin[4], using a larger window (–5 to +8) than the usual one (–3 to +6); +1 indicated the position of G in the splice-donor site GT and –1 indicated the position of the nucleotide just before GT.

Discriminant functions for promoter recognition. We used two different QDFs to discriminate a promoter region 570 bp in length from other genomic regions, one for CpG-related and the other for non-CpG-related first exons. The promoter QDFs use six variables: (P1) hexamer score in window (1–250); (P2) hexamer score in window (200–450); (P3) hexamer score in window (1–450); (P4) pentamer score in window (420–500); (P5) pentamer score in window (490–570); (P6a) G+C percentage in window (1–570) and (P6b) CpG percentage in window (1–570). Variables P1–P5 are common to both of the QDFs, whereas variable P6b is used in the CpG-related promoter QDF and P6a in the non-CpG-related promoter QDF.

Discriminant functions for first-exon recognition. Four different QDFs are used to discriminate first exons from other genomic regions depending on whether or not the first intron region is GC rich and on whether or not the first exon region is CpG related. The CpG-related first exon QDF uses 12 variables: D1–D4, P1–P6, exon length and the relative distance of the CpG window starting position from the splice-donor site. The non-CpG-related first exon QDF uses 11 variables: D1–D4, P1–P6 and exon length. We used the first exons and promoters of the first-exon database for training and testing the QDFs. The training set consists of 1,949 first exons and 31,485 pseudo-exons. A pseudo-exon is defined as any region flanked by a GT dinucleotide at the 3′ end. All of the pseudo-exons were collected from inside the gene regions (the genomic regions from ATG to the stop codon) of those genes that were annotated from experimental evidence. First exons and pseudo-exons were preclassified as (i) CpG-related and GC-rich (1,178 real and 4,945 pseudo-exons), (ii) CpG-related and GC-poor (165 real and 1,402 pseudo-exons), (iii) non-CpG-related and GC-rich (219 real and 16,294 pseudo-exons) and (iv) non-CpG-related and GC-poor (387 real and 18,844 pseudo-exons). The different QDFs were trained using the classified data.

Availability of FirstEF. FirstEF can be obtained at http://www.cshl.org/mzhanglab/.

Acknowledgments

This work was supported by grants to M.Q.Z. from the National Institutes of Health, and I.G. is also supported by a CSHL Association fellowship. We thank G. Chen for setting up the web interface to FirstEF, as well as N. Banerjee, K. Hermann, H. Herzel, M. Hoffman, D. Holste, W. Li, F. Lillo, M. Ronemus, R. Sachidanandam, K. Rateitschak, A. Schmitt and Z. Xuan for valuable discussions and comments on the manuscript.

Received 3 July; accepted 19 October 2001.

1. Lander, E.S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J.C. *et al*. The sequence of the human genome. *Science* **291**,1304–1351 (2001).
3. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
4. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
5. Solovyev, V.V., Salamov, A.A. & Lawrence, C.B. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22**, 5156–5163 (1994).
6. Zhang, M.Q. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94**, 565–568 (1997).
7. Cleverie, J.M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
8. Galas, D.J. Sequence interpretation: making sense of sequence. *Science* **291,** 1257–1260 (2001).
9. Stormo, G.D. Gene-finding approaches for eukaryotes. *Genome Res.* **10**, 394–397 (2000).
10. Maroni, G. The organization of eukaryotic genes. *Evol. Biol.* **29**, 1–19 (1996).
11. Scherf, M., Klingenhoff, A. & Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**, 599–606 (2000).
12. Davuluri, R.V., Suzuki, Y., Sugano, S. & Zhang, M.Q. CART classification of human 5′ UTR sequences. *Genome Res.* **10**, 1807–1816 (2000).
13. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
14. Ioshikhes, I.P. & Zhang, M.Q. Large-scale human promoter mapping using CpG islands. *Nature Genet.* **26**, 61–63 (2000).
15. Hattori, M. *et al*. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
16. Dunham, I. *et al*. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
17. Lemon, B. & Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14**, 2551–2569 (2000).
18. Claverie, J.M. From bioinformatics to computational biology. *Genome Res.* **10**, 1277–1279 (2000).
19. Perier, R.C., Praz, V., Junier, T., Bonnard, C. & Bucher P. The eukaryotic promoter database (EPD). *Nucleic Acids Res.* **28,** 302–303 (2000).
20. Hong, S.J. & Weiss, S.M. Advances in predictive models for data mining. *Pattern Recognition Let.* **22**, 55–61 (2001).
21. Cross, S.H. & Bird, A.P. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5**, 309–314 (1995).
22. Cross, S., Kovarik, P., Schmidtke, J. & Bird, A. Non-methylated islands in fish genomes are GC-poor. *Nucleic Acids Res***. 19,** 1469–1474 (1991).
23. Zhang, M.Q. Identification of human gene core promoters in silico. *Genome Res.* **8**, 319–326 (1998).
24. Venables, W.N. & Ripley, B.D. *Modern Applied Statistics with S-Plus* (Springer, New York, 1994).