EGPred: Prediction of Eukaryotic Genes Using Ab Initio Methods After Combining With Sequence Similarity Approaches

Biju Issac and Gajendra Pal Singh Raghava

Genome Res. 2004 14: 1756-1766
Access the most recent version at doi:10.1101/gr.2524704

EGPred: Prediction of Eukaryotic Genes Using Ab Initio Methods After Combining With Sequence Similarity Approaches

Biju Issac and Gajendra Pal Singh Raghava[1]

Institute of Microbial Technology, Sector 39A, Chandigarh-160036. India

EGPred is a Web-based server that combines ab initio methods and similarity searches to predict genes, particularly exon regions, with high accuracy. The EGPred program proceeds in the following steps: (1) an initial BLASTX search of genomic sequence against the RefSeq database is used to identify protein hits with an E-value <1; (2) a second BLASTX search of genomic sequence against the hits from the previous run with relaxed parameters (E-values <10) helps to retrieve all probable coding exon regions; (3) a BLASTN search of genomic sequence against the intron database is then used to detect probable intron regions; (4) the probable intron and exon regions are compared to filter/remove wrong exons; (5) the NNSPLICE program is then used to reassign splicing signal site positions in the remaining probable coding exons; and (6) finally ab initio predictions are combined with exons derived from the fifth step based on the relative strength of start/stop and splice signal sites as obtained from ab initio and similarity search. The combination method increases the exon level performance of five different ab initio programs by 4%–10% when evaluated on the HMR195 data set. Similar improvement is observed when ab initio programs are evaluated on the Burset/Guigo data set. Finally, EGPred is demonstrated on an ~95-Mbp fragment of human chromosome 13. The list of predicted genes from this analysis are available in the supplementary material. The EGPred program is computationally intensive due to multiple BLAST runs during each analysis. The EGPred server is available at http://www.imtech.res.in/raghava/egpred/.

[Supplemental material is available online at www.genome.org and http://www.imtech.res.in/raghava/egpred/supl/.]

Gene identification programs can be classified into two categories: ab initio methods and similarity-based methods (Mathè et al. 2002). Ab initio methods derive multiple and dissimilar gene structural information based on compositional properties of exons, introns, and other gene features to predict gene locations (Fickett 1996). Similarity-based methods use localized alignment of query sequences to known genes, proteins, complementary DNA (cDNA), or expressed sequence tags (ESTs; Gish and States 1993).

Numerous ab initio gene prediction methods have been developed (Zhang 2002). Seven ab initio methods were evaluated on a nonhomologous mammalian data set by Rogic et al. (2001). They reported that among the evaluated programs, Genscan (http://genes.mit.edu/GENSCAN.html; Burge and Karlin 1997) and HMMgene (http://www.cbs.dtu.dk/services/HMMgene/; Krogh 1997) were able to predict the precise locations of 70%–80% of the coding exons with low false positives (Rogic et al. 2001). The success of these programs is believed to be dependent on the probability that the underlying gene models are correct and the training samples are not biased (Zhang 2002). The inability of these programs to predict complete gene structure for each gene sequence has motivated researchers to investigate the benefits of combining predictions from two or more programs (Murakami and Takagi 1998; Rogic et al. 2002). Four different programs were used to develop five different combination methods for combining output from two or more programs in the GeneScope client server (Murakami and Takagi 1998). However, even the best of the five methods, the OR method does not achieve the sensitivity achieved by Genscan (Murakami and Takagi 1998). Pavlovic et al. (2002) developed a Bayesian network framework that learns the dependencies between predictions obtained from several programs or experts. The framework then uses a *Combination of experts* method for predicting the genes. The combination method uses hidden Markov models that define the captured correlation to effectively combine predictions from different programs. The method has been applied for analysis of the *Adh* region of *Drosophila*. The approach is benchmarked against standard combination methods (based on the OR & AND approach). However, the approach does not achieve a uniform increase in all of the components of accuracy, particularly at the exon level (Pavlovic et al. 2002). A combination of the predictions from Genscan and HMMgene based on their probability score was used to develop three combination methods: (1) the exon union-intersection (EUI) method; (2) the exon union-intersection with reading frame consistency (EUI-Frame) method; and (3) the gene intersection (GI) method. These methods have been implemented as a Web server, GeneComber, for prediction of genes using Genscan and HMMgene output (http://bioinformatics.ubc.ca/genecomber/). These methods increase the accuracy of exon prediction by 5%–10% (Rogic et al. 2002). However, more improvement is needed for these programs to be reliable in each single instance.

The similarity search programs, such as BLASTX and Sim4, are very effective in improving the accuracy of gene prediction (Gish and States 1993; Florea et al. 1998). Similarities with three different types of sequences—proteins, cDNA/ESTs transcripts, and genomic DNA—can provide information about exon/intron locations (Mathé et al. 2002). Similarity searches using programs such as BLASTX detect similarities between genomic DNA and

[1]Corresponding author.
E-MAIL raghava@imtech.res.in; FAX +91-172-269-0632 or +91-172-269-0585.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2524704.

14:1756–1766 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

database sequences. These programs indicate only the approximate locations of many coding exons. Moreover, they do not identify every exon and do not accurately delineate exon boundaries (Gish and States 1993). Similarity can often be misleading, as sometimes conservation in regions may be in only some part of coding exons, or the similarity may extend to the introns and/or the UTR regions of genes. However, a previous study suggests that more than 50% of newly sequenced vertebrate genes have similar sequences in the sequence databases (Claverie 1997). Most genes in the newly sequenced genomes are annotated based on similarity to known sequences (Lander et al. 2001; Waterston et al. 2002).

The recent trend in computational gene prediction is to combine similarity information with ab initio methods (Mathè et al. 2002; Ashurst and Collins 2003). Usually, ab initio gene prediction and similarity searches are run independently with the output from these two approaches being manually integrated for gene annotation. Many attempts to automate the integration have been made. Procrustes improves detection of exon boundaries based on close protein homolog for that gene using a 'spliced alignment' program (Gelfand et al. 1996). Similarly, GeneWise combines an HMM model for gene prediction with protein-profile HMM (Birney and Durbin 2000). GeneWise requires close homologues to identify complete genes (Guigo et al. 2000). A new version of the program, called GenomeWise, uses cDNA and ESTs to define spliced gene structure. GenomeWise and GeneWise are used extensively in the Ensembl genome annotation pipeline (Birney et al. 2004). Like most alignment methods, these programs are computationally intensive and require a preliminary scan with BLASTX or other search programs to identify the candidate regions. The TwinScan program finds genes in high-throughput genomic (HTG) sequences containing an unknown number of genes by exploiting the similarity between genomes of closely related organisms (Korf et al. 2001). A highly integrative approach is used in the EuGèneHom program that combines predictions from the gene predictor NetGene2 and the splice sites program SplicePredictor (Foissac et al. 2003). It integrates output of TBLASTX analysis on multiple homologous sequences from closely related organisms, start codon and splice site prediction, and a coding/noncoding probabilistic model to improve gene prediction quality. However, the program is currently tuned to only plant sequences (specifically angiosperms). Another method, the Combiner program, describes three different scoring strategies for combining evidence available from typical annotation pipelines including ab initio gene finders, protein sequence alignments, EST and cDNA alignments, and splice site predictions (Allen et al. 2004). The advantage of this program lies in its ability to integrate evidence from more than two gene finders and other extraneous sources.

The program GenomeScan (http://genes.mit.edu/genomescan.html) was developed for predicting genes; it is an extension of Genscan and incorporates similarity with a protein detected by BLASTX (Yeh et al. 2001). The method first derives information from BLASTX into a set of probabilistic statements which are then used to increase the likelihood of parses from Genscan that are consistent with similarity information, and reduces the likelihood of those parses that have no similarity. GenomeScan is able to predict coding regions missed by both Genscan and BLASTX when used alone (Mathè et al. 2002). The integration of similarity information within the GenomeScan model significantly improves the accuracy of gene prediction (increase in exon sensitivity by 10%; see Results) over Genscan, from which it is derived. Initial comparisons by the authors of GenomeScan against GeneWise and Procrustes favor GenomeScan, mainly because the latter methods truncate the predicted exons close to the end of the aligned regions, irrespective of location of splice sites or initiation/termination signals (Yeh et al. 2001). In the present study we made an attempt to further improve the accuracy of gene prediction, using a new similarity search strategy. The quality of BLASTX predictions was improved in two steps. We performed a BLASTN search of genomic DNA against an intron database in order to filter spurious exons obtained from a BLASTX search against the RefSeq database. The NNSPLICE server was used to scan genomic DNA at terminal regions of probable coding exons to assign the correct splice sites. Exons predicted using an ab initio and similarity-based approach were then combined based on the relative score (see Methods) of gene structural signals. We computed the performance of our approach and of the existing gene prediction methods on two independent data sets. A Web-based server, EGPred, was developed that implements the approach described above. The results of the evaluation are available as Supplemental material.

METHODS

Data Sets Used for Evaluation

Two different gene sequence data sets were used in this study for evaluating different programs, including our EGPred. The first data set is the HMR195 sequence data set developed by Rogic et al. (2001), which contains a total of 195 human (103), mouse (82), and rat (10) sequences. The HMR195 data set contains 948 exons altogether in 43 single-exon genes and 152 multi-exon genes. The second data set is the Burset/Guigo, which contains 570 multi-exon vertebrate genes containing a total of 2649 coding exons, developed by Burset and Guigo (1996). Both of the data sets contain sequences that have one-gene-per-sequence.

Databases for Similarity Searches

In this study we used the RefSeq (Pruitt and Maglott 2001) protein databases for the BLASTX searches. Initially the SWISS-PROT database (Boeckmann et al. 2003) was also considered but was not taken for extensive analysis, because a truly representative protein database was required for this study (see Supplemental material). The RefSeq database provides a comprehensive, integrated, nonredundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. The RefSeq protein database includes only representative sequences from the following organisms: *Drosophila melanogaster*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Sachharomyces cerevisiae*. All of the RefSeq proteins were obtained from ftp://ftp.ncbi.nih.gov/refseq/. To remove any unfair advantage to similarity-based methods over ab initio methods, all of the RefSeq protein sequences that are coded by HMR195 data set sequences were removed from the two databases (see Supplemental material).

The intron database (Sakharkar et al. 2002) was obtained for detecting potential intron regions that are then used for removing wrongly predicted exons. The intron database contains intron sequences derived from experimentally validated eukaryotic protein coding genes recovered from GenBank eukaryotic entries. Intron sequences that were included in the HMR195 data set genes were removed from the intron database to prevent unfair advantage to the similarity methods over ab initio methods during the comparison.

First BLASTX

Query genomic DNA sequence is searched against the RefSeq sequence database using the BLASTX program (Altschul et al. 1997). Initially, the expectation value (E-value) is kept at 1 to derive all probable strong and weak hits. The word length 2 is

chosen to make an extensive search, and the '−I' option is used to derive GI numbers of probable BLASTX hits from the database. Predictions are obtained only for the forward strand (+). Default parameters are selected for other options during the BLASTX search (e.g., BLOSUM62). GI numbers extracted from the BLASTX report are saved in a file for use in the second BLASTX analysis.

Second BLASTX

All proteins identified in the first BLASTX are extracted. A second BLASTX search of the genomic sequence against these limited protein sequences with relaxed parameters (E-value <10) is used to retrieve all probable coding exon regions without masking low-complexity regions. The '−l' option is used to provide the BLASTX program the list file containing the GI numbers of all hits from the initial search. Similar to the first run for BLASTX, predictions are obtained for the forward (+) strand only.

Detection of Exons From BLASTX Output

To derive probable coding exons from the BLASTX report, we used the following steps: (1) all high-scoring segment pairs (HSPs) without any inframe stop codons or four consecutive gaps are considered exons; (2) HSPs that have an inframe stop codon are split at the stop codon, and the longest segment is considered the exon; (3) HSPs with four consecutive gaps are split at the gap, and the longest segment is considered the exon; (4) multiple overlapping BLASTX hits are pruned in a preprocessing step to keep only the strongest hit (lowest E-value).

Significance of Exons

To compute the significance of probable exons assigned above, we derive four different exon parameters—the length of exon, sequence identity/similarity, score and E-value of the HSP. The sequence identity/similarity are recalculated if any exon is a subregion derived from an HSP.

BLASTN Search Against Introns

A BLASTN search of genomic sequences against the intron database is run to identify potential intron regions in the query sequence. A low E-value, $<10^{-50}$, and a very stringent word length, 15, are used to obtain more reliable intron hits to query sequence. Regions of low complexity are masked during BLASTN searches. After the search, only introns above 60 bp in length (average minimum intron length) are considered for our method.

Filtering of Exons

All of the probable exons and introns obtained from similarity searches were compared in order to identify overlapping exon and intron predictions. Different thresholds were studied for filtering exons using introns. Based on the data, all exons having 40% or more overlapping length with introns were removed from the list of probable exons (data not shown). Boundaries of exons were reassigned if (1) only their terminal ends have overlapping introns, and (2) the overlap covers only 5% to 40% of exon length. Overlapping introns of below 5% of particular exon length are not considered, because exon-intron/intron-exon boundaries are correctly defined by similarity methods.

Reassigning the Splice Site and Start/Stop Signal Positions

In this step, an attempt is made to detect the positions of start/stop codon and splice sites (e.g., intron–exon and exon–intron). We used the Web-based NNSPLICE program (Reese et al. 1997) for detecting the splice sites in a region +/− 50 bp upstream and downstream of probable exon ends (http://www.fruitfly.org/seq_tools/splice.html). NNSPLICE employs separate feedforward neural networks with one layer of hidden units to recognize acceptor and donor sites (Reese et al. 1997). Positions other than the original site and having more probability of being a splice site depending on the score from the neural network are assigned as the correct splice site for that category (i.e., acceptor/donor). In cases of no score from the network, the original positions are assumed to be the correct sites. For start and stop signals, the positions of HSPs that are near the N- or C-terminals when aligned along the length of database protein sequence are used as a template to predict the positions of these signals. After reassigning the splice sites, a filtration of exons from the similarity method is done to remove candidate exons that are less than 30% in length of the aligned region of database protein sequence.

Predictions Using Ab Initio Programs

In this study, we used the programs Genscan (Burge and Karlin 1997), HMMgene (Krogh 1997), and the EUI, EUI-Frame, and GI methods (Rogic et al. 2002). We also compared the performance of the EGPred method with GenomeScan, which also incorporates similarity search information (Yeh et al. 2001). All programs were used for prediction with the default parameters as suggested by their developers. In the case of GenomeScan, we used the Web server for prediction. Predictions using GenomeScan are obtained by providing each gene from the data set, the set of protein hits that are obtained for that gene from BLASTX search against the RefSeq database used above. This provides all methods with the same pool from which to derive information about gene structure.

Combination of Ab Initio and Similarity-Based Predictions

The exons predicted from an ab initio method and a similarity search approach are divided into five groups. The first group contains all exons that are predicted exactly the same by both approaches and are considered true exons. The second group contains exons from both methods where only the 5′ end matches. The positions from ab initio predictions are replaced with positions from similarity predictions at the 3′ end on the following conditions: (1) the ab initio method predicts the exon as a 'last exon' type, and the similarity method predicts an overlapping 'internal exon' with donor site. The identity of exon from the similarity-based method should be above 90% to database sequence, and the score of NNSPLICE prediction should be above 0.5 to allow the replacement of 3′ end; (2) the ab initio method predicts a 'single exon' type, and the similarity method predicts an 'initial exon' type with donor site. The 3′ end is replaced if the identity of the similarity-derived exon is more than 95% to database sequence and the score of donor site from NNSPLICE is above 0.9; (3) The ab initio method predicts an 'initial exon' type, and the similarity method predicts a 'single exon' type with very high significance of E-value below 10^{-200}; (4) both the ab initio and similarity-based methods predict an 'internal exon' type, and the NNSPLICE predicts a donor site with score above 0.5. The third group contains exons from both methods where only the 3′ end matches. The position of the 5′ end from ab initio predictions is replaced on the following conditions: (1) both the ab initio and similarity-based methods predict an 'initial exon' type and very high significance of E-value below 10^{-50}; (2) both the approaches predict a 'single exon' type, and the exon from the similarity method has a very high significance of E-value $<10^{-200}$; (3) both the approaches predict an 'internal exon' type, and NNSPLICE predicts an acceptor site with a score above 0.5; (4) the ab initio method predicts an 'internal exon' type, and the similarity method predicts an 'initial exon' type

with identity above 90% with database sequence; (5) the ab initio method predicts a 'single exon' type, and the similarity-based method predicts an 'internal exon' type with high significance of E-value $<10^{-100}$, and NNSPLICE predicts an acceptor site with a score above 0.5. The fourth group contains exons from both methods that overlap each other with both ends being dissimilar. The positions of 5′ ends of such exons are replaced with positions from the similarity-based method on the following conditions: (1) the ab initio method predicts an 'initial exon', and the similarity approach predicts a 'single exon' type with more than 90% identity; (2) the ab initio method predicts a 'single exon' type and the similarity approach predicts an 'internal exon' with high significance of E-value <10–50, and NNSPLICE predicts an acceptor site with a score above 0.5; (3) both approaches predict an 'internal exon' type, and NNSPLICE predicts an acceptor site with a score above 0.5. The positions at 3′ ends of overlapping exons are replaced with positions from the similarity approach on the following conditions: (1) both approaches predict an 'internal exon' type, and NNSPLICE predicts a donor site with a score above 0.5; (2) both approaches predict an 'initial exon', with NNSPLICE predicting a donor site with a score above 0.5. The fifth group contains exons predicted only by the similarity method (E-value $<10^{-50}$) or by the ab initio program (probability >0.5) alone are considered true exons. For the EUI, EUI-Frame, and GI methods, the probability score is not considered, as they are combination methods.

Rules for Combination of Exons From Both Approaches

In this study, the rules were derived based on accuracy achieved on predictions from the HMR195 data set. The boundaries of predicted exons from ab initio methods were modified based on occurrence of conflicting evidence from the similarity search against protein sequences or intron sequences and evidence of different positions for splice sites predicted by the NNSPLICE program. The modification is affected only when additional evidence from the BLAST or NNSPLICE programs are above a cutoff threshold. Cutoffs for percent identity (PID) and expectation value (E-value) in cases of BLAST-predicted exons or introns and the cutoff for scores in cases of splice sites from NNSPLICE were derived from the HMR195 data set. The cutoffs are dependent on types, exons—initial, internal, terminal, and single exons—that are predicted by both similarity-based and ab initio methods.

Accuracy Measures

The accuracy measures for evaluating the different methods used in this study were previously reported by Burset and Guigo (1996) and Rogic et al. (2001). We used the excellent script developed by Rogic et al. (2001) for computing accuracy measures that is available at their Web site. However, modifications were made in the script to perform averaging over all sequences in the data set instead of averaging only over those sequences for which predictions were obtained. This prevents overestimation of specific accuracy measures. All of the predictive accuracy measures at the nucleotide and exon levels were calculated as given below.

Nucleotide-Level Accuracy

We define TP (true positives) as the number of coding nucleotides predicted as coding; TN (true negatives) as the number of noncoding nucleotides predicted as noncoding, FP (false positives) as the number of noncoding nucleotides predicted as coding, and FN (false negatives) as the number of coding nucleotides predicted as noncoding. Accordingly,

$$Sen(sensitivity) = \frac{TP}{TP + FN} \text{ and } Spe(specificity) = \frac{TP}{TP + FP}$$

These are widely used measures of accuracy for gene prediction programs. Two measures that capture both specificity and sensitivity are the *CC* (correlation coefficient) and *AC* (approximate correlation). However, *AC* has the advantage over *CC*, because the latter is not defined in cases where the subject sequence lacks either coding regions or noncoding regions. *AC* is defined by Burset and Guigo (1996) as:

$$AC = \left(\left\{\frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right)\right\} - 0.5\right) * 2$$

Exon-Level Accuracy

Exon-level sensitivity (ESEN) is defined as the proportion of actual exons that are accurately predicted as exon. Exon-level specificity (ESPE) is defined as the proportion of predicted exons that are accurately predicted. The average of exon-level accuracy (EAVG) is defined as the average of exon-level sensitivity and exon-level specificity. Usually this is used as a reliable measure of a program's exon-level accuracy. To get a better estimate of prediction accuracy of the analyzed programs, we also computed the number of CRs (correct exons), WEs (wrong exons), MEs (missed exons), PCs (partially correct exons) and OLs (overlapping exons), where CRs represent those predicted exons that are correctly predicted, WEs are the noncoding regions predicted as exons, MEs are the exons predicted as noncoding regions, PCs are those predicted exons that have at least one end of exon correctly predicted, and OLs are predicted exons that overlap actual exons.

Application to the Human Chromosome 13

The region of human chromosome 13 sequence of ~95 Mbp in length (from 17,918,001 to 114,093,021 bp) that was sequenced and analyzed recently (Dunham et al. 2004) was obtained from http://www.sanger.ac.uk/HGP/Chr13/. The annotation available in the public domain for the segment of human chromosome 13 analyzed was obtained from http://vega.sanger.ac.uk/Homo_sapiens/exportview.

Because EGPred is critically dependent on its component program for accurate prediction, it is necessary to use programs that can predict multiple genes in long DNA sequences. Therefore, genes were predicted in chromosome 13 using EGPred methods that use the Genscan and HMMgene programs in combination with a similarity-based approach, as these ab initio programs are capable of predicting more than one-gene-per-sequence. EGPred by default predicts genes only in the forward strand. Therefore, the chromosome sequence is re-analyzed for the reverse strand by reverse-complementing the entire genomic DNA before analysis using EGPred. Because the EGPred method predicts genes only in the forward strand, the entire sequence was reverse-complemented and then used for prediction on the reverse strand. The predictions for the reverse strand are computed from the last base towards the first base, whereas predictions for the forward strand imply that positions of exons are computed from the first base towards the last base. The annotation available in the public domain was compared against the predictions from EGPred.

RESULTS

Gene Prediction Using Similarity Search

HMR195 Data Set

The performance of similarity search methods using BLASTX, second BLASTX, incorporation of intron information, and splice

Table 1. BLASTX Performance on HMR195 and Burset/Guigo Data Set on Adding Similarity Information

Program	No. genes	Nucleotide level				Exon level							
		SEN	SPE	AC	CC	CR	PC	OL	ME	WE	ESEN	ESPE	EAVG
HMR195 data set													
BLASTX (1st cycle)	1	0.88	0.64	0.69	0.66	36	277	643	98	1920	0.04	0.02	0.03
BLASTX (2nd cycle)	0	0.91	0.72	0.76	0.75	75	317	453	117	962	0.18	0.12	0.15
BLASTX+NNSPLICE	0	0.91	0.73	0.77	0.76	569	226	52	116	961	0.59	0.40	0.49
BLASTX+INTRON	0	0.90	0.84	0.84	0.84	75	316	446	124	380	0.18	0.13	0.16
BLASTX+INTRON+NNSPLICE	0	0.91	0.89	0.87	0.87	565	221	34	132	192	0.58	0.56	0.57
Burset/Guigo data set													
BLASTX (1st cycle)	0	0.92	0.61	0.67	0.66	123	746	1737	246	5697	0.04	0.02	0.03
BLASTX (2nd cycle)	0	0.90	0.75	0.77	0.76	168	1050	1070	380	1867	0.06	0.04	0.05
BLASTX+NNSPLICE	0	0.91	0.77	0.79	0.78	1677	515	106	376	1858	0.62	0.48	0.55
BLASTX+INTRON	4	0.89	0.86	0.85	0.84	168	1041	1030	421	584	0.06	0.06	0.06
BLASTX+INTRON+NNSPLICE	7	0.89	0.93	0.89	0.88	1698	433	46	472	258	0.64	0.67	0.66

Only the forward (+) strand exons from default output of programs tested were compared to GenBank annotated exons for each sequence. The standard measures of predictive accuracy were averaged over all sequences in the data set: SEN, nucleotide level sensitivity; SPE, nucleotide level specificity; AC, approximate correlation; CC, correlation coefficient; ESEN, exon level sensitivity; ESPE, exon level specificity; EAVG, (ESEN + ESPE)/2; ME, number of missed real exons; WE, number of predicted wrong exons; CR, number of correctly predicted exons that are correct at both ends; PC, number of predicted exons that are partially correct; OL, number of predicted exons overlapping actual exons; No. genes, number of genes where no predictions were made by the programs.

site information from the NNSPLICE program is shown in Table 1. The second BLASTX search of query sequence against the hits from the previous BLASTX on the HMR195 data set increases the number of correctly (CR) predicted exons from 36 to 75, and the number of partially correct (PC) exons increases from 277 to 317. This also reduces the number of wrongly predicted (WE) exons from 1920 to 962. On the other hand, the second BLASTX reduces the number of overlapping (OL) exons from 643 to 453, and increases the number of missed (ME) exons from 98 to 117. There is significant improvement in sensitivity (ESEN, from 0.04 to 0.18) and specificity (ESPE, 0.02 to 0.12) at the exon level. Incorporation of intron information reduces WE from 962 to 380, whereas ME is increased from 117 to 124. Modifications of the splice site positions using the NNSPLICE program significantly increase the exon-level sensitivity (ESEN) from 0.18 to 0.58, and exon-level specificity (ESPE) from 0.13 to 0.56. Because the derivation of rules for margining conflicting evidence from more than one source is based entirely on the HMR195 data set, there might be a probable bias in performance for the HMR195 data set. To study the effect of these rules on independent data sets, the performance of EGPred was evaluated on the Burset/Guigo data set.

Burset/Guigo Data Set

A similar trend was observed in the Burset/Guigo data set, where the results achieved validate the consistency of rules derived on the HMR195 data set to be generalized and therefore equally effectively on independent data sets. Table 3 (below) suggests an improvement of performance in the Burset/Guigo data set on adding similarity search information, where the second BLASTX increases the CRs from 123 to 168 and PCs from 746 to 1050. The number of WEs is reduced from 5697 to 1867. However, OLs are reduced from 1737 to 1070, and MEs increase from 246 to 380. The average performance (EAVG) at the exon level increases from 0.03 to 0.05. On incorporating the intron information, WEs decrease from 1867 to 574 and MEs increase from 380 to 421. The integration of splice site positions based on scores obtained from the NNSPLICE program increases the EAVG from 0.06 to 0.66 (Table 1).

Performance of Ab Initio Programs

HMR195 Data Set

The performance of the ab initio methods—Genscan, HMMgene, EUI, EUI-Frame, and GI—on the HMR195 data set is shown in Table 2. The performance of these methods after incorporating similarity search information from the protein database and after incorporating similarity information from protein and intron databases with splice site information is also shown in Table 2. The ESEN for Genscan increases from 0.70 to 0.81, the ESPE increases from 0.69 to 0.73, and the EAVG increases from 0.70 to 0.77 on incorporating similarity information against the protein database. Specificity at the exon level further increases from 0.73 to 0.74 when similarity information from intron and splice sites is incorporated. The improvement is at both the nucleotide and exon levels (Table 2). It was interesting to note that there were no missing genes when similarity information is included, whereas original Genscan missed three genes. Similarly, no genes were missed by the ab initio methods on inclusion of similarity information, unlike the original programs, where the numbers of missed genes were five, three, three, and 15 for the HMMgene, EUI, EUI-Frame, and GI methods respectively. On inclusion of similarity information from proteins, the ESEN for HMMgene increases from 0.74 to 0.80 and ESPE increases from 0.75 to 0.78. ESPE further increases to 0.78 on including similarity information from intron and splice site predictions. EAVG for HMMgene also increases from 0.75 to 0.80 on incorporating complete information from similarity searches and NNSPLICE predictions. A similar trend was observed for the EUI, EUI-Frame, and GI methods, where the performance of all methods improves significantly at the exon as well as nucleotide levels (Table 2).

The performance of GenomeScan was also evaluated on the HMR195 data set. It achieves ESEN of 0.78, ESPE of 0.71, and EAVG of 0.74. The number of genes missed by GenomeScan is one, demonstrating that the strategy adopted here is more effective than the strategy used by Yeh et al. (2001) in GenomeScan.

Burset/Guigo Data Set

The performance of the ab initio programs on the Burset/Guigo data set is shown in Table 3. The performance of Genscan on

Table 2. Performance of Ab Initio Programs on HMR195 Data Set on Adding Similarity Information

Program	No. genes	Nucleotide level				Exon level							
		SEN	SPE	AC	CC	CR	PC	OL	ME	WE	ESEN	ESPE	EAVG
GENSCAN	3	0.93	0.89	0.91	0.89	735	131	10	76	104	0.70	0.69	0.70
GENSCAN+BLASTX	0	0.98	0.89	0.92	0.92	799	110	11	36	176	0.81	0.73	0.77
GENSCAN+BLASTX+BLASTN	0	0.97	0.90	0.92	0.92	801	101	9	45	163	0.81	0.74	0.78
HMMgene	5	0.90	0.90	0.91	0.89	715	98	11	128	81	0.74	0.75	0.75
HMMgene+BLASTX	0	0.95	0.93	0.93	0.93	779	91	15	68	91	0.80	0.78	0.79
HMMgene+BLASTX+BLASTN	0	0.95	0.95	0.94	0.93	780	86	13	74	74	0.80	0.80	0.80
EUI	3	0.92	0.94	0.93	0.91	769	74	4	104	55	0.77	0.81	0.79
EUI+BLASTX	0	0.96	0.94	0.94	0.94	795	86	8	63	65	0.80	0.81	0.81
EUI+BLASTX+BLASTN	0	0.95	0.96	0.95	0.94	795	81	7	69	53	0.80	0.83	0.82
EUI − FRAME	3	0.92	0.94	0.93	0.91	762	70	5	115	46	0.77	0.82	0.79
EUI − FRAME+BLASTX	0	0.96	0.95	0.94	0.94	793	84	9	67	56	0.80	0.82	0.81
EUI − FRAME+BLASTX+BLASTN	0	0.95	0.96	0.95	0.94	793	79	8	73	48	0.80	0.83	0.82
GI	15	0.84	0.89	0.90	0.84	742	57	3	149	43	0.72	0.80	0.76
GI+BLASTX	0	0.95	0.95	0.94	0.94	796	78	7	71	56	0.80	0.83	0.82
GI+BLASTX+BLASTN	0	0.94	0.97	0.94	0.94	796	70	6	77	45	0.80	0.85	0.83
GENOMESCAN	1	0.96	0.87	0.91	0.90	781	115	16	45	213	0.78	0.71	0.74

inclusion of similarity information from proteins improved significantly, with ESEN increasing from 0.78 to 0.84, ESPE increasing from 0.79 to 0.80, and EAVG increasing from 0.79 to 0.82. The performance was further improved when information from splice site and intron search was included: ESPE increased to 0.82 and EAVG increased to 0.83. A similar trend was observed for HMMgene, where ESEN increased from 0.76 to 0.83, ESPE increased from 0.77 to 0.83, and EAVG increased from 0.77 to 0.83 when similarity information from proteins was integrated. As shown in Table 3, the performance of all methods improved significantly, particularly at the exon level, when similarity information was incorporated as described.

Our evaluation of GenomeScan on the Burset/Guigo data set showed ESEN of 0.81, ESPE of 0.78, and EAVG of 0.80. The number of genes that are completely missed by the GenomeScan program is 12. As shown in Tables 2 and 3, the performance of Genscan using similarity information as described in our study is more successful (predicting all of the genes) than that applied by GenomeScan. The performance of EGPred was higher at both the nucleotide and exon levels.

A Case Study: Gene Prediction in Human Chromosome 13

The performance of modern gene-finding algorithms must be tested on long DNA sequences in order to cope with the huge amounts of genomic DNA information coming from sequencing projects. It is important to evaluate the performance of newly developed methods in realistic situations where DNA sequence consists of multiple genes, unlike the one-gene-per-sequence model. To demonstrate the capability of EGPred, we evaluated its performance on the partial human chromosome 13 that was re-

Table 3. Performance of Ab Initio Programs on Burset/Guigo (1996) Data Set on Adding Similarity Information

Program	No. genes	Nucleotide level				Exon level							
		SEN	SPE	AC	CC	CR	PC	OL	ME	WE	ESEN	ESPE	EAVG
GENSCAN	8	0.93	0.91	0.92	0.90	2156	264	26	203	188	0.78	0.79	0.79
GENSCAN+BLASTX	0	0.98	0.91	0.94	0.93	2299	230	28	100	305	0.84	0.80	0.82
GENSCAN+BLASTX+BLASTN	0	0.98	0.93	0.94	0.94	2301	222	22	112	250	0.84	0.82	0.83
HMMgene	38	0.87	0.88	0.91	0.86	2092	239	21	308	139	0.76	0.77	0.77
HMMgene+BLASTX	2	0.95	0.95	0.94	0.94	2256	248	21	135	144	0.83	0.83	0.83
HMMgene+BLASTX+BLASTN	2	0.95	0.96	0.95	0.94	2252	234	17	154	97	0.83	0.84	0.84
EUI	20	0.90	0.92	0.93	0.90	2214	176	12	250	98	0.80	0.84	0.82
EUI+BLASTX	0	0.96	0.96	0.95	0.95	2297	213	14	129	106	0.84	0.86	0.85
EUI+BLASTX+BLASTN	0	0.96	0.97	0.95	0.95	2294	205	12	142	79	0.84	0.87	0.85
EUI − FRAME	27	0.88	0.92	0.92	0.88	2188	167	11	286	87	0.79	0.84	0.81
EUI − FRAME+BLASTX	1	0.96	0.96	0.95	0.95	2286	214	13	140	95	0.84	0.86	0.85
EUI − FRAME+BLASTX+BLASTN	1	0.95	0.97	95.00	0.90	2283	206	12	152	69	0.83	0.87	0.85
GI	43	0.84	0.90	0.91	0.85	2118	138	8	387	67	0.76	0.83	0.80
GI+BLASTX	2	0.95	0.97	0.95	0.95	2277	195	10	167	76	0.83	0.87	0.85
GI+BLASTX+BLASTN	2	0.94	0.97	0.94	0.94	2274	190	9	178	52	0.83	0.88	0.85
GENOMESCAN	12	0.95	0.89	0.92	0.90	2245	235	45	130	350	0.81	0.78	0.80

Table 4. Summary of Predictions on Human Chromosome 13 Using EGPred

	Genes				Exons						
Programs (1*)	Multi-exon (2*)	Single-exon (3*)	Partial (4*)	Total genes predicted (5*)	Total exons predicted (6*)	Exon per gene (7*)	Total exon length (8*)	Exons predicted by *ab initio* approach only (9*)	Exons predicted by similarity only (10*)	Exons predicted by both approaches (11*)	Number of matches to annotated exons (12*)
Genscan (+)	1048 (45.2)	211 (9.1)	1058 (45.6)	2317	6712	2.89	1116630 bp (1.16)	5320	0	1392	1212
HMMgene (+)	1975 (57.6)	119 (3.5)	1337 (38.9)	3431	8372	2.44	1134771 bp (1.18)	7233	0	1139	939
Genscan (−)	1077 (47.3)	195 (8.6)	1007 (44.1)	2279	6831	2.99	975084 bp (1.01)	5360	0	1471	1215
HMMgene (−)	2025 (57.9)	101 (2.9)	1368 (39.2)	3494	8206	2.35	994322 bp (1.03)	7018	0	1188	1022

*Column number.
Predictions were made using similarity-based approach against RefSeq protein database and Intron database in combination with two different ab initio predictors—Genscan and HMMgene. The figure in column headers indicates the number of that column. The figure in parentheses in columns 2–4 denotes percentage of total predicted genes, and figure in parentheses in column 8 denotes percent of total analyzed nucleotide sequence of human chromosome 13. Columns 9–11 represent the number of exons that are predicted by only the ab initio approach, the similarity-based approach, and by both the approaches, respectively. The last column shows the number of predicted exons from each category that are found to match to that provided by public domain. These matches include the exact, partial, and overlapping exon matches.

cently sequenced (Dunham et al. 2004). Human chromosome 13 is the largest acrocentric human chromosome, estimated to contain 633 genes with a total of 4266 exons excluding those from the pseudogenes (Dunham et al. 2004). An initial analysis was performed as described in Methods. A total of 96,175,021 bp was analyzed in a region from 17,918,001 to 114,093,021 bp to predict genes.

The genes were predicted using two different combinations implemented in EGPred for the Genscan and HMMgene methods. Application of these two strategies on human chromosome 13 produced four sets of putative genes (two for each strand), summarized in Table 4. A total of 2125 multi-exon genes, 406 single exon genes, and 2065 partial genes are predicted by the Genscan-based EGPred strategy. The HMMgene-based EGPred strategy produced 4000 multi-exon genes, 220 single-exon genes, and 2705 partial genes. The average numbers of exons per gene are also listed in Table 4. The total protein-coding region of the predicted exons covers slightly over 1% in each direction for both strategies out of the total chromosome segment analyzed (Table 4). Of the 10,680 exons predicted by the Genscan-based EGPred strategy, 2427 exons match those from the publicly available exon annotation data. Similarly, of the 14,251 exons predicted by the HMMgene-based EGPred strategy, only 1961 exons match the publicly available exon annotation data. Interestingly, all of the exons predicted by the similarity-based approach were also predicted by the ab initio approach (Table 4).

However, this does not suggest that only the ab initio method needs to be used for predictions. Many of the ab initio-predicted exons are modified based on evidence from a similarity-based approach. Moreover, the amount of known proteins is presently limited, and the accuracy will increase further with the increase in available data. The large number of predicted partial genes in Table 4 from both strategies reflects the fragmentation of genes in the sequenced chromosome segment. This fraction is likely to decrease with greater availability of sequence data. Some of the gene structures are likely to reflect the alternatively spliced isoforms. However, the mechanisms underlying alternative splicing are not well understood, and computational analysis of this phenomenon will require more specialized tools than those reported here.

Table 5 shows the comparison of EGPred predictions with available public domain annotation of human chromosome 13. Surprisingly, more than 70% of exons predicted by the EGPred are not reported in the available annotation. Because a considerable proportion of genes are estimated to be absent from the current databases, many of the predicted genes may be potentially novel protein coding genes. Results in Table 4 suggest that all predictions from the similarity-based approaches are also predicted by the ab initio approaches. Because EGPred uses similarity to protein sequences, a large fraction of these common predictions are likely to be protein coding genes. A direct one-to-one comparison of predicted exons from both of the EGPred-based strategies reveals that almost all exons predicted by one method are also predicted by the other method (Table 6). A large number of overlapping predictions points to the need to develop a more accurate splice site prediction program.

Table 6. Comparison of Predictions From the Two EGPred Methods

	Both ends match	One end matches	Overlap match	No match
Genscan vs. HMMgene for (+) strand	1236	742	4734	0
Genscan vs. HMMgene for (−) strand	1309	701	4821	0
HMMgene vs. Genscan for (+) strand	1236	740	6396	0
HMMgene vs. Genscan for (−) strand	1309	705	6192	0

The physical locations of all genes predicted by the EGPred method along the human chromosome 13 DNA sequence for both strands are available from http://www.imtech.res.in/raghava/egpred/supl/HChr13/. Although EGPred is demonstrated to be reliable for both short and long DNA sequences, the success of the program is critically dependent on the accuracy of underlying programs, and continued improvements in gene-prediction algorithms should improve future EGPred results.

Web Server Description

A Web server, EGPred, was developed to predict the genes in eukaryotic genomic DNA using the approach described in this study, where similarity information was integrated with different ab initio programs. The server allows users to paste or upload a nucleotide sequence in FASTA format. Although by default the server uses the parameters that show its best performance as in the present study, it also allows the users to change the various parameters. The parameters that users are allowed to change include the organism type for the Genscan program, the organism type and probability score cut-off of predicted exons for the HMMgene program, the E-value cut-off, the protein database that is searched against, and matrices and word length used during BLASTX searches. It is possible that large sequences may take substantial time for gene prediction. Therefore, we allow users to obtain results via e-mail.

EGPred presents the results in graphic format as a GIF image, where it shows the genes predicted by different strategies along the length of query sequence. Figure 1 shows graphical output from EGPred for a mouse sequence (GenBank accession no. X07625, Locus ID MMPROT1) on default parameters. Exons predicted by different methods are represented by different colors. In addition to the graphic format, the server presents results in the standard gene-finding format (GFF). The annotation shows

Table 5. Comparison of Predictions on Human Chromosome 13 Using EGPred With Available Public Domain Annotation

	Total predicted exons	Available annotated exons	Both ends match	One end matches	Overlap match	Novel exons
Genscan (+) strand	6712	2800	996	198	18	5500
Genscan (−) strand	6831	2905	1086	187	42	5516
HMMgene (+) strand	8372	2800	674	244	21	7433
HMMgene (−) strand	8206	2905	724	239	59	7184

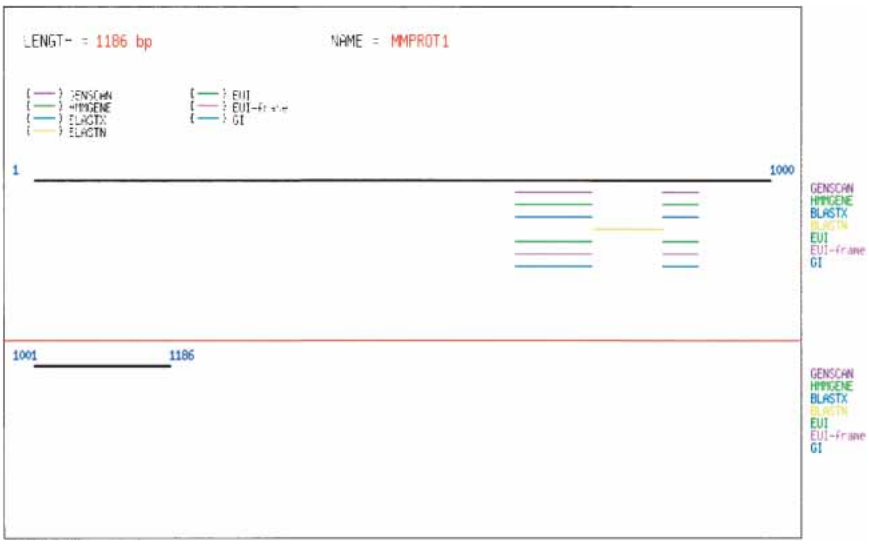

Figure 1 Graphical output from the EGPred server for a mouse gene (GenBank accession no. X07625, Locus ID MMPROT1) on default parameters. The solid black line indicates the query sequence. The predictions are represented as colored lines. The color code for each program is indicated at the *top* of the image. Longer sequences are broken into segments of 1000 bp and separated by a long red line. The length of query sequence is indicated at the *top* of the output image.

the start of exon, stop of exon, and type of exon predicted. The exon type can be any of the following: initial exon, internal exon, terminal exon, and single exon. The exons are modeled into gene structures based on the order of their physical occurrence along the sequence. The possible gene structures are single exon genes, multi-exon genes, and partial genes. Partial genes do not have complete gene structure.

DISCUSSION

One of the major challenges in the field of genome annotation is to improve the accuracy of gene prediction. Although a number of methods have been developed to predict genes, their accuracy is quite low at the exon level (Rogic et al. 2001; Zhang 2002). It has been shown that the performance of ab initio methods can be improved significantly if similarity search information is incorporated (Yeh et al. 2001). In the present study, a systematic attempt was made to improve the performance of different ab initio methods using sequence similarity information. First, we evaluated the performance of sequence similarity approaches on independent data sets. Initially the performance of similarity search approaches was tested on the SWISS-PROT and RefSeq databases. The specificity of BLASTX was found to be better on the RefSeq database (see Supplemental material). The reason for this is that RefSeq is a representative database, unlike SWISS-PROT, which is an all-inclusive, nonspecific database. Many false hits are thus avoided by using the RefSeq database. However, the results shown in Table 3 suggest that the present RefSeq size may not be sufficient for all sequences, and additional proteins from more organisms are required to make the database complete.

As shown in Table 1, the performance of the initial BLASTX was quite low but improved significantly when a second BLASTX search against hits of the first BLASTX was performed. These observations agree with results reported by Yeh et al. (2001). The performance of our BLASTX approach is comparable to that obtained by Guigo et al. (2000) on using default BLAST. In our study, we introduce the concept of detecting introns by searching query sequence against an intron database (Sakharkar et al. 2002) using BLASTN. These potential introns are compared with probable exons obtained from a similarity search in order to filter/remove wrongly predicted exons. Surprisingly, the incorporation of intron information reduces the WE significantly. Intron sequences are thought to have little conservation among them. However, it is also clear that coding exon regions will have very negligible or no similarity with intron sequences. This may be due to a variety of factors including codon bias, di-nucleotide frequency bias, hexamer bias, etc. We attempted to use these differences between coding exon and intron regions to filter out spurious exons. However, the results suggest that any advantage gained from intron information will be highly dependent on related sequences being present in the intron database.

We observed that the performance of ab initio methods improved at the nucleotide level but decreased at the exon level when we incorporated exons obtained from a similarity search (data not shown). This is because exons from similarity search approaches do not predict correct exon-intron and intron-exon boundaries. As shown in Table 1, the performance of similarity-based methods (with protein and introns) improved significantly (EAVG increases from 0.16 to 0.57) at the exon level when splice sites were predicted using the NNSPLICE program (Reese et al. 1997). Thus in our strategy we used NNSPLICE and BLAST search against a protein database and intron database.

The accuracy values achieved by the original ab initio programs shown in Table 2 are lower than that reported by Rogic et al. (2001). This is because we computed the average of accuracy measures over all sequences in the data sets instead of over only those sequences from which predictions are made by these individual programs. For example, in the case of Genscan we computed the performance for the HMR195 data set, on 195 where Rogic et al. measured performance on 192 (195 − 3 missed genes). Rogic et al. (2001) also accept that their accuracy measures may lead to overestimation of performance. These results suggest that the increase in gene prediction performance is not uniform for all programs. The accuracy on incorporation of similarity information is dependent more on the accuracy achieved by the ab initio programs themselves (Tables 2, 3). Overall the accuracy of all ab initio programs improved significantly, particularly at the exon level, where performance increases from 4% to 10%. It must be noted that there were no missed genes when similarity information was integrated with these ab initio methods, whereas the original programs missed from three to 15 genes in the HMR195 data set.

An important observation is that the rules derived for merging conflicting evidence are generalized for sequences with the one-gene-per-sequence property. Our results from the Burset/Guigo data set prove that the rules work effectively on sequences independent of the HMR195 data set (Table 3). For deriving maximum benefit for using the rules, an important requirement is that overlapping predictions from multiple sources should be available. Experimentally proven sequences with one-gene-per-sequence like those in the HMR195 and Burset/Guigo data sets have protein sequences in the databases that are similar to their protein products. Moreover, programs such

as NNSPLICE use available sequence information for training. This would suggest an evident bias of performance of EGPred toward sequences that have similar sequences available in the database.

The reliability of EGPred for predicting genes in large DNA sequences is also proved by the demonstration of two different EGPred strategies on a region of ~95 Mbp-long human chromosome 13. Initial results from the experiment suggest a much greater number of genes than what is currently available in the public domain. This difference may be due to a variety of reasons, including the fact that the manual annotation is the product of all curated predictions based on alignment to all publicly available expressed sequences and application of gene-prediction algorithms (Dunham et al. 2004). The manual annotation is therefore a conservative estimate of the actual number of genes that may be present in chromosome 13. Although the actual expression of the EGPred-predicted genes would be proven only through much experimental work, a fraction of these predicted genes are supported by known protein sequence data and therefore are more likely to be protein-coding genes. However, we emphasize that users should consider the predictions on human chromosome 13 hypothetical until experimentally proven.

The data shown in Tables 4 and 5 suggest that more than half of the manually annotated exons are missed by the EGPred method, whereas it predicts a large number of unannotated or novel exons. Our initial survey of novel exons suggests a high false-positive rate, integral mainly to the ab initio methods—Genscan and HMMgene—used in EGPred. However, removal of the number of genes proportional to the fraction of wrong exons (WEs) obtained by these individual ab initio methods still results in a high number of unaccounted exons. Because the predictions in EGPred from ab initio methods are filtered based on high confidence scores, it is more than likely that these novel exons are protein-coding. One of the reasons that EGPred detects only half of the manually annotated exons may be based on the use of only protein sequence data for combination-based predictions, whereas manual annotation uses evidence from multiple sources including ESTs and cDNA apart from protein sequences. This would suggest the need for integrating evidence from these additional sources and other gene-prediction programs as well. The incorporation of recently published methods such as the Combiner program (Allen et al. 2004) that integrate multiple gene-prediction programs and evidence from cDNA, ESTs, proteins, and splice site predictors may further improve EGPred predictions. Combiner makes use of three different strategies for effective combination of predictions from more than one gene-prediction program. Two of the Combiner methods are linear methods (LC1 and LC2) that use an equal or unequal voting function with a subsequent dynamic programming (DP) algorithm to construct gene models from different inputs. The third method, a decision tree-based nonlinear statistical combiner (SC) model, uses confidence scores for combinations of predictions from more than two gene finders. Incorporating one or more of these methods into the EGPred algorithm would provide a strategy for combining predictions from more than one gene finder.

In conclusion, we can say that the strategy incorporated in this study is very effective to improve the performance of gene-prediction methods. Although we have demonstrated the use of EGPred only for combinations of similarity-based approaches to five different ab initio methods, EGPred can be extended to other existing/new programs. The EGPred program is written in Perl. As a service to the interested community, we developed a Web server, EGPred (http://www.imtech.res.in/raghava/egpred/) for predicting genes in nucleotide sequences.

Limitations

Information flow during transcription and translation events in a cell is not static or determined by a simple, single method as the predictions from gene-finding programs suggest. There are several variants regarding how the information is passed from DNA to RNA to protein. Some of the possible events include alternative splicing, use of nonconsensus splice sites, exon skipping (where a possible second transcript from a gene does not include one or more exons that are included in the first transcript from the same gene), and nested genes (a gene that is present inside an intron of another protein-coding gene). Similarity evidence from a protein database could possibly provide valuable information regarding alternative splice sites and use of nonconsensus splice sites in a gene. One logical method to solve these problems would be to keep the complete 'global transcript' information of all protein hits from a database similarity search. From such information, it is possible to derive overlapping hits to probable exon regions. The differences in the length and signal content of overlapping exon regions from different transcripts would easily provide information regarding complex events such as use of alternate translation start sites (TSSs) and alternate splicing sites. Use of nonconsensus splice sites is also easily identifiable with the similarity-based approach against a protein database. Similarly, nested genes could possibly be identified using a negative evidence (intron) approach. Any such gene that will be completely masked by similarity to intron sequences can be considered a probable nested gene. Because protein-coding regions are usually conserved, all exons from such a smaller nested gene should show similarity to a single intron sequence (if available) in a database. A similar strategy can also be used to identify events such as exon skipping. Multiple transcripts obtained from similarity searches against protein sequences could provide information regarding whether a particular exon is not being included in one or more transcripts. Alternatively, complete masking of a probable exon (with strong supporting evidence) from a protein similarity-based transcript by an intron sequence would confirm the probability of the occurrence of an exon-skipping event. We are presently continuing work on many of the topics mentioned here to further improve the prediction through logical uses of multiple transcript information from the similarity searches against protein and intron sequences, and also to improve the information content of the output in this respect.

ACKNOWLEDGMENTS

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Allen, J.E., Pertea, M., and Salzberg, S.L. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res.* **14:** 142–148.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Ashurst, J.L. and Collins, J.E. 2003. Gene annotation: Prediction and testing. *Annu. Rev. Genomics Hum. Genet.* **4:** 69–88.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31:** 365–370.

Birney, E. and Durbin, R. 2000. Using GeneWise in *Drosophila* annotation experiment. *Genome Res.* **10:** 547–548.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and GenomeWise. *Genome Res.* **14:** 988–995.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.
Burset, M. and Guigo, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–367.
Claverie, J-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.
Dunham, A., Matthews, L.H., Burton, J., Ashurst, J.L., Howe, K.L., Ashcroft, K.J., Beare, D.M., Burford, D.C., Hunt, S.E., Griffiths-Jones, S., et al. 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428:** 522–528.
Fickett, J.W. 1996. Finding genes by computer: The state of the art. *Trends Genet.* **12:** 316–320.
Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.
Foissac, S., Bardou, P., Moisan, A., Cros, M-J., and Schiex, T. 2003. EuGene'Hom: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **31:** 3742–3745.
Gelfand, M.S., Mironov, M.A., and Pevzner, P.A. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.* **93:** 9061–9066.
Gish, W. and States, D.J. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.
Guigo, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10:** 1631–1642.
Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* (Suppl.) **17:** S140–S148.
Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene-finding. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. (eds. T. Gaasterland et al.), pp 179–186. AAAI Press, Menlo Park, CA.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitz Hugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Mathè, C., Sagot, M-F., Schiex, T., and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30:** 4103–4117.
Murakami, K. and Takagi, T. 1998. Gene recognition by combination of several gene-finding programs. *Bioinformatics* **14:** 665–675.
Pavlovic, V., Garg, A., and Kasif, S. 2002. A Bayesian framework for combining gene predictions. *Bioinformatics* **18:** 19–27.
Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.
Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. 1997. Improved splice site detection in Genie. *J. Comput. Biol.* **4:** 311–323.
Rogic, S., Mackworth, A.K., and Ouellette, B.F.F. 2001. Evaluation of gene finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.
Rogic, S., Ouellette, B.F.F., and Mackworth, A.K. 2002. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* **18:** 1034–1045.
Sakharkar, M., Passetti, F., de Souza, J.E., Long, M., and de Souza, S.J. 2002. ExInt: An exon intron database. *Nucleic Acids Res.* **30:** 191–194.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Yeh, R-F., Lim, L.P., and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.
Zhang, M.Q., 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3:** 698–709.

WEB SITE REFERENCES

http://www.imtech.res.in/raghava/egpred/; The EGPred server.
http://genes.mit.edu/GENSCAN.html; Genscan.
http://bioinformatics.ubc.ca/genecomber/; GeneComber.
http://www.cbs.dtu.dk/services/HMMgene/; HMMgene.
http://www.fruitfly.org/seq_tools/splice.html; NNSPLICE program.
http://genes.mit.edu/genomescan.html; GenomeScan.
http://www.sanger.ac.uk/HGP/Chr13/; The Human Chromosome 13 Project overview.
http://vega.sanger.ac.uk/Homo_sapiens/exportview; Download site for chromosome 13 annotation.
http://www.imtech.res.in/raghava/egpred/supl/HChr13/; Supplemental material on EGPred analysis of human chromosome 13.

Received February 27, 2004; accepted in revised form July 7, 2004.