Comparative Gene Prediction in Human and Mouse

Geni?s Parra, Pankaj Agarwal, Josep F. Abril, et al.

Genome Res. 2003 13: 108-117
Access the most recent version at doi:10.1101/gr.871403

Comparative Gene Prediction in Human and Mouse

Genís Parra,[1] Pankaj Agarwal,[2] Josep F. Abril,[1] Thomas Wiehe,[3] James W. Fickett,[4] and Roderic Guigó[1,5]

[1]*Grup de Recerca en Informàtica Biomèdica. Institut Municipal d'Investigació Medica / Universitat Pompeu Fabra / Centre de Regulació Genòmica 08003 Barcelona, Catalonia, Spain;* [2]*GlaxoSmithKline, King of Prussia, Pennsylvania 19406, USA;* [3]*Freie Universität Berlin and Berlin Center for Genome Based Bioinformatics (BCB), 14195 Berlin, Germany;* [4]*AstraZeneca R&D Boston, Waltham, Massachusetts 02451, USA*

The completion of the sequencing of the mouse genome promises to help predict human genes with greater accuracy. While current ab initio gene prediction programs are remarkably sensitive (i.e., they predict at least a fragment of most genes), their specificity is often low, predicting a large number of false-positive genes in the human genome. Sequence conservation at the protein level with the mouse genome can help eliminate some of those false positives. Here we describe SGP2, a gene prediction program that combines ab initio gene prediction with TBLASTX searches between two genome sequences to provide both sensitive and specific gene predictions. The accuracy of SGP2 when used to predict genes by comparing the human and mouse genomes is assessed on a number of data sets, including single-gene data sets, the highly curated human chromosome 22 predictions, and entire genome predictions from ENSEMBL. Results indicate that SGP2 outperforms purely ab initio gene prediction methods. Results also indicate that SGP2 works about as well with 3x shotgun data as it does with fully assembled genomes. SGP2 provides a high enough specificity that its predictions can be experimentally verified at a reasonable cost. SGP2 was used to generate a complete set of gene predictions on both the human and mouse by comparing the genomes of these two species. Our results suggest that another few thousand human and mouse genes currently not in ENSEMBL are worth verifying experimentally.

After the genome sequence of an organism has been obtained, the very first next step is to compile a complete and accurate catalog of the genes encoded in this sequence. For higher eukaryotic organisms, however, the accuracy of currently available gene prediction methods to perform such a task is limited (Guigó et al. 2000; Rogic et al. 2001; Guigó and Wiehe 2003). The increasing availability of genome sequences from different organisms, however, has lead to the development of new computational gene finding methods that use sequence conservation to help identifying coding exons, and improve the accuracy of the predictions (Fig. 1; Crollius et al. 2000; Wiehe et al. 2000; Miller 2001; Rinner and Morgenstern 2002). Indeed, three such comparative gene prediction programs, SLAM (Pachter et al. 2002), SGP2, and TWINSCAN (Korf et al. 2001) have been used for the comparative analysis of the human and mouse genomes. These analyses lead to more accurate gene predictions, and to the verification of previously unconfirmed genes. In this paper, we describe the program SGP2. Typical computational ab initio gene prediction methods rely on the identification of suitable splicing sites, start and stop codons along the query sequence, and the computation of some measure of coding likelihood to predict and score candidate exons, and delineate gene structures (see Claverie 1997; Burge and Karlin 1998; Haussler 1998; Zhang 2002 and references therein for reviews on computational gene finding).

Similarity between the query sequence and known coding sequences (amino acid or cDNA) can also be used to infer gene structures. When the query sequence encodes a protein for which a close homolog exists, a special type of alignment can be used between the DNA sequence and the target protein/cDNA sequence, in which gaps in the target sequence corresponding to introns in the query sequence must be compatible with potential splicing signals. This is the approach in GENEWISE (Birney and Durbin 1997) and PROCRUSTES (Gelfand et al. 1996). Alternatively, the results of searching the query sequence against a database of known coding sequences, using for instance BLASTX (Altschul et al. 1990, 1997; Gish and States 1993), can be incorporated more or less ad hoc into the scoring schema of an ab initio gene prediction method. The program GENOMESCAN (Yeh et al. 2001), which incorporates BLASTX search results into the predictions by the GENSCAN program (Burge and Karlin 1997), is an example of a recent development in that direction.

Recently developed comparative gene prediction programs further exploit sequence similarity. Instead of comparing anonymous genomic sequences to known coding sequences, anonymous genomic sequences are compared to anonymous genomic sequences from the same or different organisms, under the assumption that regions conserved in the sequence will tend to correspond to coding exons from homologous genes. The approach taken by the different programs to exploit this idea differs notably.

In one such approach (Blayo et al. 2002; Pedersen and Scharl 2002), the problem is stated as a generalization of pairwise sequence alignment: Given two genomic sequences coding for homologous genes, the goal is to obtain the predicted exonic structure in each sequence maximizing the score of the

[5]Corresponding author.
E-MAIL rguigo@imim.es; FAX 34 93 224-0875.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.871403.

13:108–117 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

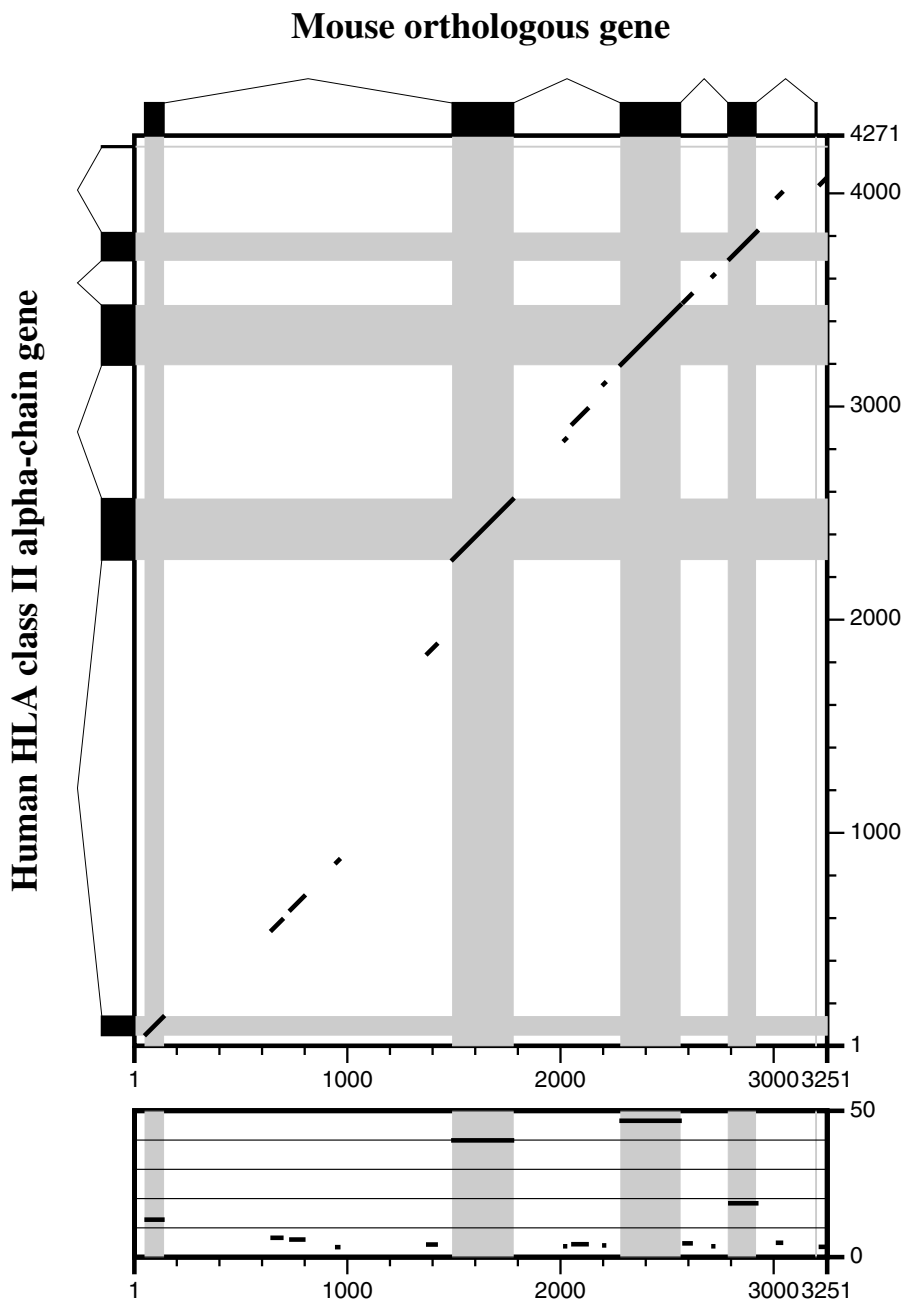


Figure 1 Pairwise comparison using TBLASTX of the human and mouse genomic sequences coding for the HLA class II alpha chain. Black boxes indicate the coding exons, while black diagonals indicate the conserved alignments. The score of the conserved alignments (divided by 10) is given in the lower panels. Although conserved regions between the human and mouse genomic sequences coding for these genes fully include the coding exons, a substantial fraction of intronic regions is also conserved. The TBLASTX outptut was post-processed to show a continuous non-overlapping alignment.

alignment of the resulting amino acid sequences. Both Blayo et al. (2002) and Pedersen and Scharl (2002) solve the problem through a complex extension of the classical dynamic programming algorithm for sequence alignment.

In a different approach, the programs SLAM (Pachter et al. 2002) and DOUBLESCAN (Meyer and Durbin 2002) combine sequence alignment pair hidden Markov Models (HMMs; Durbin et al. 1998) with gene prediction generalized HMMs (GHMMs; Burge and Karlin 1997) into the so-called generalized pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above; gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopt a more heuristic approach, and separate clearly gene prediction from sequence alignment. The programs ROSSETA (Batzoglou et al. 2000), SGP1 (from 'syntenic gene prediction'; Wiehe et al. 2001), and CEM (from 'conserved exon method'; Bafna and Huson 2000) are representative of this approach. All these programs start by aligning two syntenic sequences and then predict gene structures in which the exons are compatible with the alignment. The programs described thus far rely on the comparison of fully assembled (and when from different organisms, syntenic) genomic regions. This limits their utility when analyzing complete large eukaryotic genomes, and in particular when the informant genome is in nonassembled shotgun form. To overcome this limitation, the programs TWINSCAN (Korf et al. 2001) and SGP2 take still a different approach. The approach is reminiscent of that used in GENOMESCAN (Yeh et al. 2001) to incorporate similarity to known proteins to modify the GENSCAN scoring schema. Essentially, the query sequence from the target genome is compared against a collection of sequences from the informant genome (which can be a single homologous sequence to the query sequence, a whole assembled genome, or a collection of shotgun reads), and the results of the comparison are used to modify the scores of the exons produced by ab initio gene prediction programs. In TWINSCAN, the genome sequences are compared using BLASTN, and the results serve to modify the underlying probability of the potential exons predicted by GENSCAN. In SGP2, the genome sequences are compared using TBLASTX (W. Gish, 1996–2002, http://blast.wustl.edu), and the results are used to modify the scores of the potential scores predicted by GENEID. TWINSCAN and SGP2 have been successfully applied to the annotation of the mouse genome

(Mouse Genome Sequencing Consortium 2002), and have helped to identify previously unconfirmed genes (Guigó et al. 2003).

In the next section, we describe the algorithmic details of SGP2, and its implementation. We also describe the sequence sets used to benchmark SGP2 accuracy. Results based on these data sets indicate that SGP2 is an improvement over pure ab initio gene prediction programs, even when the informant genome is only in shotgun form. We have found that 3x coverage will generally suffice to achieve maximum accuracy. Finally, we describe the application of SGP2 to the comparative analysis of the human and mouse genomes.

METHODS

SGP2

SGP2 is a method to predict genes in a *target* genome sequence using the sequence of a second *informant* or *reference* genome. Essentially, SGP2 is a framework to integrate the ab initio gene prediction program GENEID (Guigó et al. 1992; Parra et al. 2000) with the sequence similarity search program TBLASTX. The approach is conceptually similar to that used in TWINSCAN to incorporate BLASTN searches into GENSCAN.

GENEID is a genefinder that predicts and scores all potential coding exons along a query sequence. Scores of exons are computed as log-likelihood ratios, which are a function of the splice sites defining the exon, and of the coding bias in composition of the exon sequence as measured by a Markov Model of order five (Borodovsky and McIninch 1993). From the set of predicted exons, GENEID assembles the gene structure (eventually multiple genes in both strands), maximizing the sum of the scores of the assembled exons, using a dynamic programming chaining algorithm (Guigó 1998).

When using an informant genome sequence to predict genes in a target genome sequence, ideally we would like to incorporate into the scores of the candidate exons predicted along the target sequence, the score of the optimal alignment at the amino acid level between the target exon sequence and the counterpart homologous exon in the informant genome sequence. If a substitution matrix, for instance from the BLOSUM family, is used to score the alignment, the resulting score can also be assumed to be a log-likelihood ratio: informally, the ratio between the likelihood of the alignment when the amino acid sequences code for functionally related proteins, and the likelihood of the alignment, otherwise. In principle, this score could be added to the GENEID score for the exon. TBLASTX provides an appropriate shortcut to often find a good enough approximation to such an optimal alignment, and infer the corresponding score: The optimal alignment can be assumed to correspond to the maximal scoring high-scoring segment pairs (HSP) overlapping the exon. However, when dealing in particular with the informant genome sequence in fragmentary shotgun form, often different regions of a candidate exon sequence will align optimally to different informant genome sequences. Thus, in the approach used here, we identify the optimal HSPs covering each fraction of the exon, and compute separately the contribution of each HSP into the score of the exon. In the next section, we describe in detail how this computation is performed.

Scoring of Candidate Exons

Let e be one of the candidate exons predicted by GENEID along the query DNA sequence S. In SGP2, the final score of e, $s(e)$, is computed as

$$s(e) = s_g(e) + ws_t(e)$$

where $s_g(e)$ is the score given by GENEID to the exon e, and $s_t(e)$ is the score derived from the HSPs found by a TBLASTX search overlapping the exon e. Both scores are log-likelihood ratios (and we compute both base two). Assuming that both components are independent, they can be summed up into a single score. However, the assumption of independence is not realistic, $s_g(e)$ depends on the probability of the sequence of e, assuming that e codes for a protein, while $s_t(e)$ depends on the probability of the optimal alignment of e with a sequence fragment of the mouse genome, assuming that both sequences code for related proteins. Obviously, these two probabilities are not independent. Their joint distribution could only be investigated—at least empirically—if the Markov Model of coding DNA used in GENEID, and the substitution matrix used by TBLASTX were inferred from the very same set of coding sequences. Since this is quite difficult, if not unfeasible, we use an "ad hoc" coefficient, w, to weight the contribution of TBLASTX search, $s_t(e)$ into the final exon score.

We compute $s_t(e)$ in the following way. Let $h_1 \cdots h_q$ be the set of HSPs found by TBLASTX after comparing the query sequence S against a database of DNA sequences (Fig. 2A).

First, we find the *maximum scoring projection* of the HSPs onto the query sequence. We simply register the maximum score among the scores of all HSPs covering each position, and then partition the query sequence in equally maximally scoring segments (bounded by dotted lines in Fig. 2A) $x_1 \cdots x_r$, with scores $s_p(x_1) \cdots s_p(x_r)$ (Fig. 2B).

Then, for each predicted exon e (Fig. 2C), we find X_e, the set of maximally scoring segments overlapping e

$$X_e = \{x_i : x_i \cap e \neq \emptyset\}$$

where $a \cap b$ denotes the overlap between sequence segments a and b, and $\emptyset$ means no overlap. We compute $s_t(e)$in the following way:

$$s_t(e) = \sum_{x \in X_e} s_p(x) \frac{|x \cap e|}{|x|}$$

where $|a|$ denotes the length of sequence segment a.

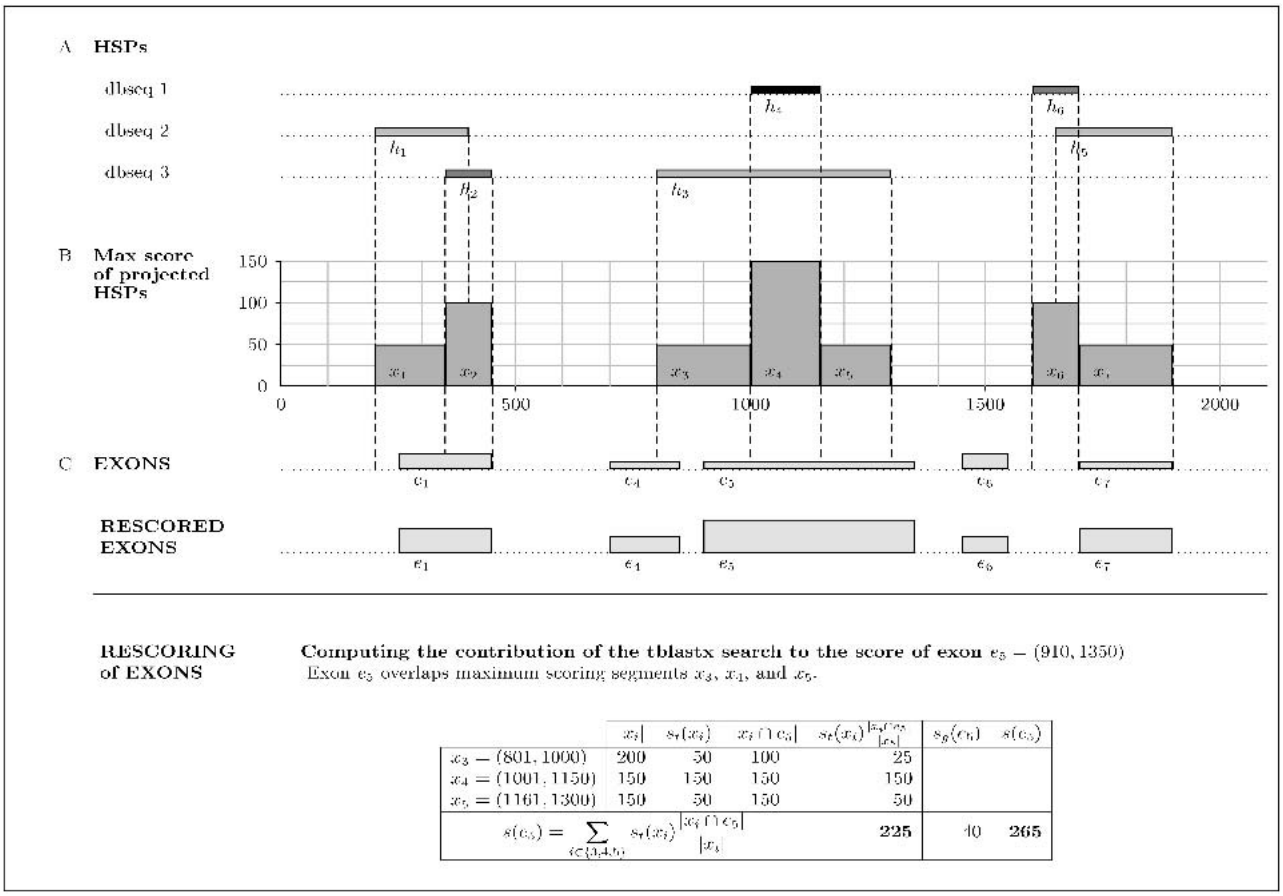
That is, each exon gets the score of the maximally scoring HSPs along the exon sequence proportional to the fraction of the HSP covering the exon. In other words, $s_t(e)$ is the integral of the maximum scoring projection function within the exon interval.

Once the scores s have been computed for all predicted exons in the sequence S, gene prediction proceeds as usual in GENEID: The gene structure is assembled maximizing the sum of scores of the assembled exons.

Running SGP2

In practice, we run SGP2 in the following way. Given a DNA query sequence and a collection of DNA sequences, we compare the query sequence against the collection using TBLASTX 2.0MP-WashU [23-Sep-2001]. The query sequence can be a genomic fragment of any size, including complete eukaryotic chromosomes, whereas the collection of sequences may be almost anything from just a homologous region or a partial collection of genomic sequences from the same or another species to the whole genome sequence of a second species, either completely assembled or in shotgun form at any degree of coverage. In particular, two different regions of the same genome coding for homologous genes can be used within SGP2; in this case the same genome acts as target and informant.

In all the analyses reported here, we used BLOSUM62 as the amino acid substitution matrix, but changed the penalty for aligning any residue to a stop codon to -500. This helps to get rid of a large fraction of HSPs in noncoding regions. Because of TBLASTX limitations, large query sequences may need to be split in fragments before the search, and the results reconstructed afterwards. Results of TBLASTX search are then



	$\lvert x_i\rvert$	$s_t(x_i)$	$\lvert x_i \cap e_3\rvert$	$s_t(x_i)\frac{\lvert x_i \cap e_3\rvert}{\lvert x_i\rvert}$	$s_g(e_3)$	$s(e_3)$
$x_3 = (801, 1000)$	200	50	100	25		
$x_4 = (1001, 1150)$	150	150	150	150		
$x_5 = (1161, 1300)$	150	50	150	50		
$s(e_3) = \sum_{i\in\{3,4,5\}} s_t(x_i)\frac{\lvert x_i \cap e_3\rvert}{\lvert x_i\rvert}$				**225**	40	**265**

Figure 2 Rescoring of the exons predicted by GENEID according to the results of a TBLASTX search. See the "SGP2" section for a detailed explanation of the figure.

parsed to obtain the *maximum scoring projection* of the HSPs onto the query sequence. The parsing includes discarding all HSPs below a given bit score cutoff, subtracting this value from the score of the remaining HSPs, weighting the resulting score by *w* (see above), and collapsing the HSPs in to the maximum scoring projections. In all analyses described here, the bit score cutoff was set to 50, and *w* to 0.20. These values were chosen to optimize the gene predictions in sequence sets of known homologous human and mouse genomic sequences (see the Results section).

The *maximum scoring projection* is given to GENEID in general feature format (GFF; R. Durbin and D. Haussler, http://www.sanger.ac.uk/Software/GFF/). GENEID uses it to rescore the exons predicted along the query sequence as explained, and assembles the corresponding optimal gene structure. GENEID was already designed to incorporate external information into the gene predictions, and no changes were required in the program to accommodate it into the SGP2 context, only a small adjustment in the parameter file to cope with the change in scale of the exon scores.

We have written a simple PERL script which, given a query DNA sequence and the results of the TBLASTX search, performs all the components of the SGP2 analysis transparently: the parsing of the TBLASTX search results, and the GENEID predictions. In the case wherein both the query and the informant sequence are single genomic fragments, the gene predictions can be obtained in both sequences (without the need for a second TBLASTX search). The script, as well as the individual components, can be found at http://www1.imim.es/software/sgp2/.

GENEID has essentially no limits to the length of the input sequence, and deals well with chromosome size sequences. Limits to the length of the input query sequence that can be analyzed by SGP2 are, thus, those imposed by TBLASTX. GENEID is quite fast; given the parsed TBLASTX results, it takes 6 h to reannotate the whole human genome in a MOSIX cluster containing four PCs (PentiumIII Dual 500 Mhz processors).

Accelerating TBLASTX Searches

TBLASTX searches, although efficient, are much slower. Its default usage may become computationally prohibitive when comparing complete eukaryotic genomes. In the context of SGP2, however, a number of TBLASTX options can be changed to speed up the search, without significant loss of sensitivity in the predictions (see the Results section). Thus, results in human chromosome 22 and whole-genome comparisons have been performed using the following set of parameters: W = 5, -nogap, -hspmax = 150,000, B = 200, V = 200, E = 0.01, E2 = 0.01, Z = 30,000,000, -filter = xnu + seg, and S2 = 80. In these cases, the query sequences have been broken up in 5 MB fragments, and the database sequences in 10 MB fragments. In all cases, stop codons are heavily penalized (-500) in the alignments. After the search is completed, locations of the resulting HSPs are recomputed in chromosomal coordinates. Results in the single-gene sequence benchmark data sets were obtained with default TBLASTX parameters.

Sequence Data Sets

Benchmark Sequence Sets

To optimize some of the parameters in SGP2 and to test its performance, we used a set of known pairs of genomic sequences coding for homologous human and rodent genes. The set is built after the set constructed by Jareborg et al. (1999). This is a set of 77 orthologous mouse and human gene pairs. We considered only the 33 pairs of sequences in this set

coding for single complete genes. In addition, we discarded six additional pairs, when we suspected that one of the members could be wrongly annotated. Orthology in the Jareborg et al. (1999) data set is based on sequence conservation. This could bias the set towards the more highly conserved human/mouse orthologous genes. To compensate for this bias, we obtained an additional set of pairs of human/rodent orthologous genes through an approach which does not involve sequence conservation: We obtained the set of pairs of human/mouse sequences from the SWISSPROT database sharing the prefix (indicating the gene) in their locus names. We kept only those pairs for which it was possible to find the corresponding annotated genomic sequence—including the mapping of the transcript, and not only of the coding regions—in the EMBL database. Fifteen additional genes were found this way. Three of them were discarded because we suspected wrong annotation in at least one of the members of the pair. We believe that orthology in the remaining cases is highly likely because of the absolute conservation of the exonic structure (number and length of exons, and intron phases) that we observed. We will call the resulting concatenated set of 39 pairs of human/mouse homologous genes the SCIMOG dataset (from Sanger Center IMim Orthologous Genes). The data set and the detailed protocol used to obtain it can be accessed at http://www1.imim.es/datasets/sgp2002/.

To test the accuracy of SGP2, we used the data set constructed by Batzoglou et al. (2000) of 117 orthologous human and mouse genes. We discarded those pairs in which in at least one of the sequences contained multiple genes, and those in which the coding region started in position 1 in one of the sequences of the pair. This resulted in 110 genes. We will call this set the MIT data set. There is some overlap between the SCIMOG and MIT data sets, and thus the latter cannot properly be called a test set. However, we decided not to eliminate the redundant entries, so that the results could be compared to those published for the ROSSETA program (Batzoglou et al. 2000).

Finally, we tested SGP2 in the complete sequence of human chromosome 22 (Dunham et al. 1999). The masked sequence was obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/. Chromosome 22 is probably the best annotated human chromosome. We used the gene annotations at http://www.cs.columbia.edu/~vic/sanger2gbd/. The CDS set contains 554 genes. This is a conservative set that only contains the coding region of genes and does not include pseudogenes. This may lead to an underestimation of the specificity of the predictions.

Mouse and Human Genome Sequences

We used versions MGSCv3 of the mouse genome (2,726,995,854 bp, http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/) and NCBI28 of the human genome (3,220,912,202 bp, http://genome.cse.ucsc.edu/goldenPath/22dec2001/). Both masked and unmasked sequences were obtained from these locations. ENSEMBL gene annotations for these genomes were obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz for the human genome, and from http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz for the mouse genome. ENSEMBL predicts 23,005 and 22,076 nonoverlapping transcripts genes on the human and mouse genome, respectively.

Evaluating Accuracy

The measures of accuracy used here are extensively discussed in Burset and Guigó (1996). We will restate them briefly. Accuracy is measured at three different levels: nucleotide, exon, and gene. At the nucleotide and exon levels, we compute essentially the proportion of actual coding nucleotides/exons that have been correctly predicted—which we call *sensitivity*—and the proportion of predicted coding nucleotides/exons that are actually coding nucleotides/exons—which we call *specificity*. To compute these measures at the exon level, we will assume that an exon has been correctly predicted only when both its boundaries have been correctly predicted. To summarize both *sensitivity* and *specificity*, we compute the *correlation coefficient* at the nucleotide level, and the average of *sensitivity* and *specificity* at the exon level. At the exon level, we also compute the *missing exons*, the proportion of actual exons that overlap no predicted exon, and the *wrong exons*, the proportion of predicted exons that overlap no real exons.

At the gene level, a gene is correctly predicted if all of the coding exons are identified, every intron–exon boundary is correct, and all of the exons are included in the proper gene. In addition, we compute the missed genes (MGs), real genes for which none of its exons are overlapped by a predicted gene, and the wrong genes (WGs), predictions for which none of the exons are overlapped by a real gene. In general, gene finders predict the initial and terminal exons very poorly. This often leads to so-called chimeric predictions—one predicted gene encompassing more than one real gene—or to split predictions—one real gene split in multiple predicted genes. Reese et al. (2000) developed two measures, split genes (SG) and joined genes (JG), to account for these tendencies. SG is the total number of predicted genes overlapping real genes divided by the number of genes that were split. Similarly, JG is the total number of real genes that overlap predicted genes divided by the number of predicted genes that were joined.

RESULTS

Benchmarking SGP2

We evaluated the accuracy of SGP2 using a number of different data sets. The lack of a gold standard of gene prediction makes it difficult to get accurate assessments from any single data set. We primarily used three data sets as described earlier.

To benchmark SGP2, we constructed BLAST databases from the mouse and human sections of SCIMOG and MIT, and each mouse/human sequence to the entire human/mouse database, respectively. This enabled us to predict genes in both the mouse and human databases. The results from

Table 1. Gene Prediction in the SCIMOG Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.86	0.92	0.84	0.75	0.79	0.04	0.14
TBLASTX default	0.89	0.76	0.81	0.81	—	—	0.19	0.11
SGP2 (single complete genes)	0.97	0.98	0.97	0.89	0.89	0.89	0.03	0.03
SGP2 (multiple genes)	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02

Table 2. Gene Prediction Accuracy in the MIT Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.89	0.93	0.82	0.75	0.78	0.06	0.13
ROSSETA	0.95	0.97	—	—	—	—	0.02	0.03
TBLASTX default	0.94	0.79	0.85	—	—	—	0.13	0.13
SGP2 (single complete genes)	0.97	0.98	0.97	0.84	0.85	0.84	0.05	0.03
SGP2 (multiple genes)	0.96	0.97	0.96	0.71	0.79	0.75	0.12	0.03

comparing SGP2, GENSCAN, and ROSSETA accuracy values in this case are taken from Batzoglou et al. (2000), and the results of a simple TBLASTX search on the MIT data set are in Table 2 (below). For the TBLASTX searches, the *maximum scoring projection* of the HSPs (see the above section titled "SGP2") was assumed to be the gene prediction. The score cutoff for the HSPs was chosen to maximize the correlation coefficient (CC) between the projected HSPs and the coding exons. In Table 1,2, we report the accuracy of GENSCAN, SGP2, and TBLASTX on the SCIMOG dataset. The accuracy values for SGP2 are reported under two scenarios: assuming a single complete gene and assuming multiple genes. Both GENEID and SGP2 allow the external specification of a *gene model* (i.e., a small number of rules specifying the legal assemblies of exons into gene structures). These rules can be used to force SGP2 to predict a single complete gene to make the results comparable to those of ROSSETA. Without such a restriction (i.e., making no assumptions about the number and completeness of the genes potentially encoded in the query sequence), the results are more directly comparable to those of GENSCAN (although GENSCAN also has a tendency to start a prediction in any sequence with an initial exon, and to terminate it with a terminal exon).

The accuracy of SGP2 is comparable to that of ROSSETA, and is significantly higher than that of GENSCAN. SGP2 also improves substantially over a simple TBLASTX search. The relative low specificity of the TBLASTX search—even after the large penalties for stop codons—reflects the fact that a substantial fraction of the conservation between the human and mouse genomes extends into the noncoding regions (Mouse Genome Sequencing Consortium 2002). At the nucleotide level, SGP2 accuracy is almost equal in the MIT data set and the SCIMOG data set (even though the SGP2 was trained on SCIMOG). The accuracy at the exact exon level, however, decreases, in particular when prediction of multiple genes is allowed. This is a problem inherited from GENEID, which tends to replace short initial and terminal exons with longer internal exons.

Accuracy of SGP2 as a Function of the Coverage of the Mouse Genome

To investigate the utility of partial shotgun data as informant sequence in our approach based on TBLASTX, we simulated shotgun mouse sequence data at different levels of coverage (1.5x, 3x, and 6x) from the mouse genes in the SCIMOG data set, and used them to compare the human sequences in SCIMOG using TBLASTX. The mouse genomic sequences was shredded with uniformly distributed length between 500 and 600 bp with random starting points. No sequencing errors were introduced. At each coverage, we measured the CC between the TBLASTX hits projected along the human genome sequence, and the coding exons (choosing the TBLASTX score cutoff resulting in the optimal CC). With 1.5x coverage, a substantial fraction of the human coding region is not identified by TBLASTX, whereas with 3x, the results are quite similar to those obtained with 6x, which are identical to those obtained with the fully assembled syntenic regions (Table 3). This indicates that even with 3x coverage of the informant genome, our method will produce results nearly identical to those obtained with fully assembled regions. Assembled genomes, however, result in faster TBLASTX searches.

Accuracy of SGP2 in Human Chromosome 22

Human chromosome 22 was the first human chromosome fully sequenced (Dunham et al. 1999), and it is quite the best annotated thus far, due to a number of experimental followups (Das et al. 2001; Shoemaker et al. 2001). Therefore, it provides an excellent data set to validate any gene prediction technology. Human chromosome 22 was searched using TBLASTX against the masked whole-genome assembly from the mouse genome (MGSCv3). The HSPs in chromosomal coordinates resulting from the TBLASTX search were used in GENEID to perform SGP2 gene prediction. Although the HSPs had been computed on the masked sequence, in this case the SGP2 predictions were obtained on the unmasked one. SGP2 predicted 729 genes on human chromosome 22. Table 4 shows the comparative accuracy of the SGP2, GENSCAN, GENOMESCAN, and pure ab initio GENEID predictions (without TBLASTX data). GENSCAN predictions on the masked sequence were taken from the USCS genome browser http://genome.cse.ucsc.edu/. GENOMESCAN predictions were obtained from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz. Pure ab initio GENEID predictions were obtained on the masked sequence, and can also be downloaded from http://www1.imim.es/genepredictions/.

Although SGP2 is not more sensitive than GENSCAN, it appears to be more specific (as it utilizes the mouse genome).

Table 3. Accuracy of TBLASTX Predictions as a Function of the Degree of Coverage in the SCIMOG Data Set

	Nucleotide			Exon	
Coverage	Sn	Sp	CC	ME	WE
Simulated 1.5x	0.79	0.78	0.77	0.25	0.10
Simulated 3x	0.86	0.76	0.80	0.21	0.11
Simulated 6x	0.89	0.76	0.81	0.19	0.11
Fully assembled	0.89	0.76	0.81	0.19	0.11

Table 4. Accuracy of Gene-finding Programs on Human Chromosome 22

	Nucleotide			Exon					Gene						
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE	Sn	Sp	(Sn+Sp)/2	MG	WG	JG	SG
GENSCAN	0.86	0.50	0.64	0.70	0.40	0.55	0.13	0.50	0.06	0.04	0.05	0.11	0.45	1.24	1.07
GENOMESCAN	0.87	0.44	0.59	0.72	0.36	0.54	0.10	0.55	0.11	0.06	0.08	0.12	0.52	1.07	1.14
GENEID	0.80	0.63	0.69	0.66	0.53	0.59	0.19	0.35	0.09	0.07	0.08	0.14	0.39	1.20	1.08
TBLASTX	0.84	0.39	0.54	—	—	—	0.12	0.74	—	—	—	0.11	—	—	—
SGP2	0.83	0.67	0.73	0.68	0.56	0.62	0.16	0.31	0.13	0.10	0.11	0.14	0.36	1.14	1.13

Fifty percent of the GENSCAN-predicted exons do not overlap annotated chromosome 22 exons; this number is only 31% for SGP2. Overall, SGP2 appears to be more accurate than GENSCAN in human chromosome 22: GENSCAN's CC at the nucleotide level is 0.64, whereas that of SGP2 is 0.73. Although accuracy decreases for both programs when going from single-gene sequences (Tables 1, 2) to an entire chromosome, SGP2 retains more accuracy. GENSCAN overall shows higher sensitivity than SGP2, but there were 45 real genes not predicted by GENSCAN on human chromosome 22, and SGP2 was able to predict, at least partially, 15 of them. This suggests that SGP2 and GENSCAN may play complementary roles. GENOMESCAN, on the other hand, did not appear to be superior to GENSCAN in human chromosome 22.

Mouse matches (TBLASTX HSPs) covered 11% of the human chromosome 22. Though they covered 85% of the coding nucleotides, 74% of the HSPs fell outside annotated coding regions. This illustrates the difficulties of using genome sequence conservation even at the protein level between human and mouse genomes to infer coding genes.

Prediction of Genes in the Human and Mouse Genomes

We used SGP2 to predict the entire complement of human (NCBI28) and mouse (MGSCv3) genes. The masked sequences of these two genomes were compared using TBLASTX. The TBLASTX HSPs were used within SGP2. SGP2 predicted 44,242 genes in the human genome, and 44,777 genes in the mouse genome. Obviously, it is difficult to accurately assess these predictions. We used ENSEMBL genes as the set of reference annotations and compared both GENSCAN and SGP2 predictions to it. Figure 3 shows summaries of the accuracy of SGP2 at the chromosome level in the human and mouse genomes. When compared against ENSEMBL, SGP2 is more accurate than GENSCAN.GENSCAN. It is more specific at the nucleotide level: the average SGP2 specificity is 0.60 for human and 0.61 for mouse, whereas these values for GENSCAN are 0.43 and 0.44. SGP2 is also equally sensitive at the nucleotide level: The average SGP2 sensitivity is 0.82 for human and 0.85 for mouse; these values for GENSCAN are 0.82 and 0.84. Overall, the average SGP2 CCs are 0.70 for human and 0.72 for mouse, and for GENSCAN, the respective averages are 0.59 and 0.61. The accuracy of the SGP2 predictions, moreover, appears to be more consistent across chromosomes than that of the GENSCAN predictions. Interestingly, human chromosome Y is an outlier, with genes in this chromosome being poorly predicted. Genes in chromosome Y appear to be more difficult to predict than genes in other chromosomes for pure ab initio gene prediction programs, because chromosome Y is also an outlier for GENSCAN. SGP2 suffers, in addition, on human chromosome Y because the mouse chromosome Y has yet to be sequenced, and thus there was no comparative information available.

Overall, 23,913 of the human predictions and 24,203 of the mouse predictions overlapped ENSEMBL genes, whereas 95% of the mouse and 93% of the human ENSEMBL genes were among the genes predicted by SGP2. Of the remaining putative novel 20,570 mouse SGP2 genes and 20,193 human SGP2 genes, 10,456 mouse and 9,006 human predictions were found to be similar at $P < 10^{-6}$ to a prediction in the counterpart genome. Of these, 5,960 and 4,909 have multiple exons and are longer than 300 bp. A significant fraction of these putative homologous predictions are likely to correspond to real genes (Guigó et al. 2003). The predictions are interactively accessible through the USCS genome browser (http://genome.cse.ucsc.edu/) and through the DAS server at ENSEMBL (http://www.ensembl.org, under "DAS sources"). The complete set of prediction files is available at http://www1.imim.es/genepredictions/.

Speeding Up TBLASTX Searches

Using TBLASTX to compare human and mouse whole-genome sequences, even in masked form, is quite expensive computationally because of the 6-frame translation on both query and target. To substantially reduce the search time, we used a word size of 5 and sacrificed some sensitivity (see the section above titled "Accelerating TBLASTX Searches" for details). We also penalized stop codons heavily and did not permit gaps. The computation took an estimated 500 CPU days on a farm of Compaq Alphas.

Accuracy in Tables 1 and 2 was computed using default TBLASTX parameters. Table 5 shows the comparative accuracy of TBLASTX and SGP2 predictions, under the default and the speed-up configuration of TBLASTX parameters on the SCIMOG data set. The sensitivity of speed-up TBLASTX searches drops from 0.89 to 0.72, but specificity increases slightly. SGP2 is more robust, and it compensates for some of the sensitivity lost in the TBLASTX search. Overall accuracy for SGP2, as measured by the CC, drops only from 0.95 to 0.93.

Predictions on human chromosome 22 and the whole human and mouse genomes have been obtained with this speed-up configuration of parameters.

DISCUSSION

We have described the program SGP2 for comparative gene finding, and presented the results of its application to the human and mouse genome sequences. Results in controlled benchmark sequence data sets indicate that, by including in-

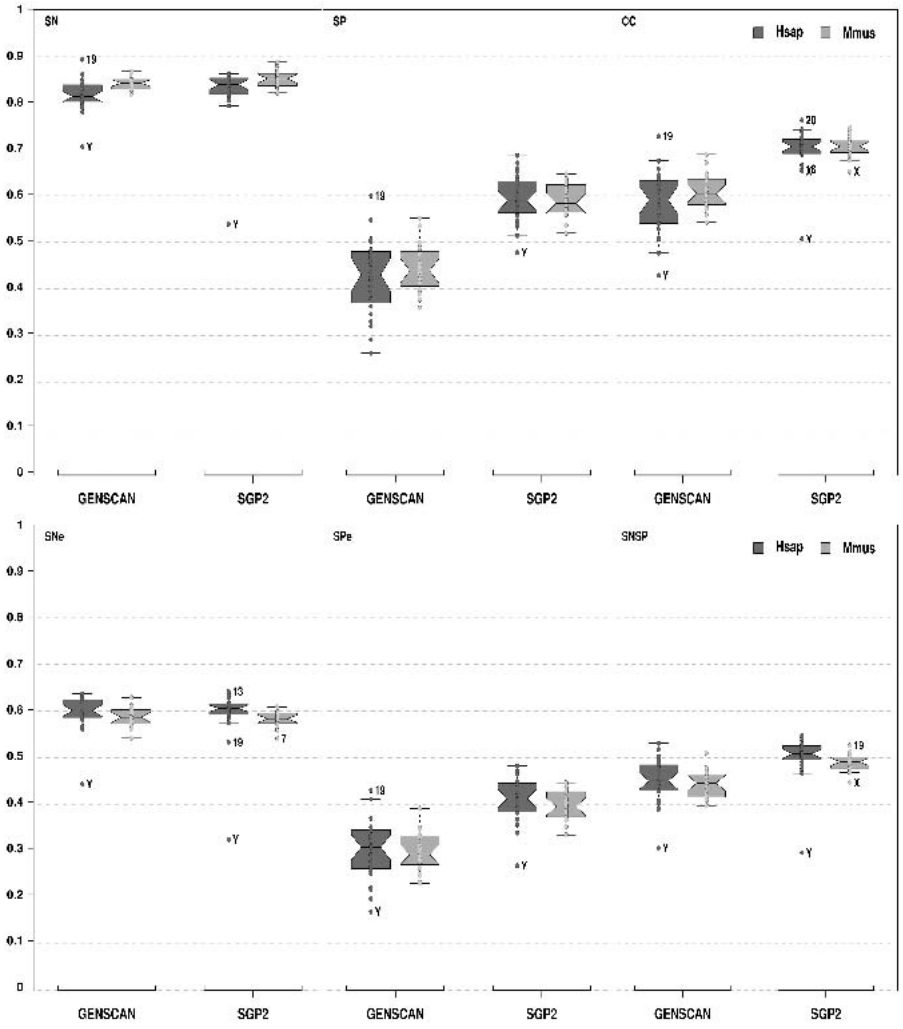


Figure 3 Accuracy of the human and mouse SGP2 and GENSCAN predictions. The accuracy was measured in the entire chromosome sequences using the standard accuracy measures: SN, (sensitivity); SP, (specificity); CC, (correlation coefficient); SNe, (exon sensitivity); SPe, (exon specificity); and SNSP, (average of sensitivity and specificity at exon level). Predictions from both programs were compared against the human and mouse ENSEMBL annotations. Each dot corresponds to the accuracy measure of one chromosome. Chromosome labels are shown for outlier values. The boxplots (Tukey 1977) were obtained using the R-package (http://cran.r-project.org/).

formation from genome sequence conservation, predictions by SGP2 appear to be more accurate than those obtained by pure ab initio programs, exemplified here by GENSCAN and GENEID. Although there is not a significant gain in sensitivity, the specificity of the predictions appears to increase substantially, and a smaller number of false positive exons are predicted.

Indeed, one the major obstacles towards the completion of the catalog of human (mammalian) genes is our inability to assess the reliability of the large number of computational gene predictions that have not been verified experimentally. Whereas the ENSEMBL pipeline produces about 25,000 human and mouse genes, the NCBI annotation pipeline predicts almost 50,000 genes in mouse, and the program GENOMESCAN predicts close to 55,000 genes in this species. Although a large fraction of the ENSEMBL genes correspond to computational predictions without experimental verification, the method is quite conservative, and recent experiments suggest that essentially all ENSEMBL genes are indeed real (Guigó et al. 2003). The problem remains with the tens of thousands of additional computational predictions that are not included in ENSEMBL. A fraction of them are likely to be real, but the question is how large this fraction is. The results obtained here in human chromosome 22 seem to indicate that it may not be very large. Although the existence of hundreds of unidentified genes in this chromosome cannot be completely ruled out, the results strongly suggest that a substantial fraction of these additional computational gene predictions are false positives.

In this regard, the results presented here demonstrate that through the comparison of the human and mouse genomes using SGP2 (or another available comparative gene prediction tool), the false-positive rate can be reduced significantly, and the catalog of mammalian genes better defined. SGP2 predicts a few thousand candidate genes not in ENSEMBL that we believe are worth verifying experimentally. Indeed, the experimental verification of a subset of these provides evidence of at least 1000 previously nonconfirmed genes (Guigó et al. 2003).

The predictions by SGP2 obtained here are, of course, still far from definitively setting this catalog. For one thing, the mouse may be too close a species to human: A large fraction of the sequence has been conserved between the genomes of these two species. Indeed, most sequence conservation between human and mouse does not correspond to coding exons (Mouse Genome Sequencing Consortium 2002), compounding gene prediction. This suggests that the genome of another vertebrate species evolutionarily located between fish and mammals could be of great utility towards closing in the vertebrate (and mammalian) gene catalog.

SGP2 is flexible enough so that it can be easily accommodated to analyze species other than human and mouse. The fact that it can deal with shotgun data at any level of coverage means that as the sequence of a new genome starts becoming available, it can be used to improve the annotation of other already existing genomes. Particularly relevant in this context is a feature of SGP2 (and GENEID) that we have not explored here. SGP2 can produce predictions on top of preexisting annotations. For instance, we could have given to SGP2 the location and exonic coordinates (in GFF format) of known REFSEQ genes (or ENSEMBL), and SGP2 would have predicted genes only outside the boundaries of these genes of

Table 5. Accuracy of TBLASTX and SGP2 Predictions Using "Default" versus Speed-Up Parameters

		Nucleotide			Exon				
		Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
Default	TBLASTX	0.89	0.76	0.81	—	—	—	0.19	0.11
	SGP2	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02
Speed-up	TBLASTX	0.72	0.80	0.75	—	—	—	0.22	0.10
	SGP2	0.88	0.98	0.93	0.77	0.85	0.81	0.12	0.02

already well known exonic structure. Preliminary results indicate that this approach improves gene prediction outside of the preassumed genes, and reduces the rate of chimeric predictions (i.e., predictions encompassing multiple genes). Moreover, we believe that SGP2 can be substantially improved. The flexibility of the SGP2/GENEID framework makes it quite easy to integrate additional information that can contribute to the accuracy of the predictions: synonymous versus nonsynonymous substitution rates in the alignments by TBLASTX, conservation of the splice signals in the informant genome, amino acid substitution matrices specific to the phylogenetic distance between the species compared, etc.

In this regard, the reasons to use the default BLOSUM62 matrix are not obvious. Given the expected sequence similarity between mouse–human orthologs, BLOSUM80 appears to be a better choice. However, we intended to also detect divergent families. Towards that end, the superiority of BLOSUM80 is less clear. We have compared TBLASTX search results on human chromosome 22 against the whole mouse genome. Whereas the HSPs resulting from the BLOSUM62 search cover 84% of the chromosome 22 coding nucleotides, BLOSUM80 HSPs cover 88% of them. However, BLOSUM80 is much less specific than BLOSUM62: 60% of the nucleotides in the BLOSUM62 HSPs fall outside coding regions, compared to 88% for BLOSUM80. It is thus clear that the optimal matrix or combination of matrices for comparative gene-finding using TBLASTX requires further investigation.

Although a large fraction of the human genome sequence has been known for more than a year, the exact number of human genes and their precise definition remain unknown. Gene specification in higher eukaryotic sequences is the result of the complex interplay of sequence signals encoded in the primary DNA sequence, which is only partially understood. Without an exhaustive catalog of human genes, however, the promises of genome research in medicine and technology cannot be completely fulfilled. The work presented here, in which it is shown that human–mouse comparisons can contribute to the completion of the mammalian (human) gene catalog, underscores the importance of the comparisons of the genomes of different organisms to fully understand the phenomenon of life, and in particular to deciphering the mechanism, central to life, by means of which the genome DNA sequence specifies the amino acid sequence of the proteins.

ACKNOWLEDGMENTS

We thank the Mouse Genome Sequencing Consortium for providing the mouse genome sequence as well as support throughout the analysis process. We especially thank Francisco Câmara for arranging the data listed in the gene-prediction page on our group Web site, and for setting up and taking care of our DAS server. We also thank Ian Korf for inspiring discussions regarding the parameters to use in the TBLASTX search. We thank Enrique Blanco, Sergi Castellano, and Moisés Burset for helpful discussions and constant encouragement. This work was supported by a grant from Plan Nacional de I+D (BIO2000-1358-C02-02), Ministerio de Ciencia y Tecnologia (Spain), and from a fellowship to J.F.A. from the Instituto de Salud Carlos III (99/9345).

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

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

Bafna, V. and Huson, D.H. 2000. The conserved exon method. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 3–12.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Birney, E. and Durbin, R. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 56–64.

Blayo, P., Rouzé, P., and Sagot, M.-F. 2002. Orphan gene finding—An exon assembly approach. *Theoretical Computer Science* (in press).

Borodovsky, M. and McIninch, J. 1993. GenMark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17:** 123–134.

Burge, C.B. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–357.

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

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25:** 235–238.

Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. 2001. Assessment of the total number of human transcription units. *Genomics* **77:** 71–78.

Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Durbin, R., Eddy, S., Crogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, Cambridge.

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

Gish, W. and States, D. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Guigó, R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comp. Biol.* **5:** 681–702.
Guigó, R. and Wiehe, T. 2003. Gene prediction accuracy in large DNA sequences. In *Frontiers in computational genomic* (eds. M.Y. Galperin and E.V. Koonin), Caister Academic Press, Norfolk, UK.
Guigó, R., Knudsen, S., Drake, N., and Smith, T.F. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.
Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. Gene prediction accuracy in large DNA sequences. *Genome Res.* **10:** 1631–1642.
Guigó, R., Dermitzakis, E.T., Agarwal, P., Pontig, C.P., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* (in press).
Haussler, D. 1998. Computational genefinding. *Trends in biochemical sciences, supplementary guide to bioinformatics,* pages 12–15.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.
Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** 140–148.
Meyer, I.M. and Durbin, R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18:** 1309–1318.
Miller, W. 2001. Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17:** 391–397.
Mouse Genome Sequencing Consortium 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Pachter, L., Alexandersson, M., and Cawley, S. 2002. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.* **9:** 389–400.
Parra, G., Blanco, E., and Guigó, R. 2000. Geneid in *Drosophila*. *Genome Res.* **10:** 511–515.
Pedersen, C. and Scharl, T. 2002. Comparative methods for gene structure prediction in homologous sequences. In *Algorithms in Bioinformatics* (eds. R. Guigó, and D. Gusfield), Springer-Verlag, Berlin, Germany.
Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.
Rinner, O. and Morgenstern, B. 2002. Agenda: Gene prediction by comparative sequence analysis. *In Silico Biol.* **2:** 0018.
Rogic, S., Mackworth, A.K., and Ouellette, F. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.
Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922–927.
Tukey, J.W. 1977. *Exploratory data analysis*. pp. 39–41. Addison-Wesley, Boston, MA.
Wiehe, T., Guigó, R., and Miller, W. 2000. Genome sequence comparisons: Hurdles in the fast lane to functional genomics. *Brief. Bioinform.* **1:** 381–388.
Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigó, R. 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11:** 1574–1583.
Yeh, R., Lim, L., and Burge, C. 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.sanger.ac.uk/Software/formats/GFF/; GFF format description page.
http://genome.cse.ucsc.edu/goldenPath/22dec2001/; Human genome sequence goldenpath from Dec. 22, 2001 (hg10) equivalent to NCBI28 build.
http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/; Mouse genome sequence goldenpath from Feb. 2002 (mm2) equivalent to MGSCv3.
http://www.cs.columbia.edu/~vic/sanger2gbd; Victoria Haghighi, Human chromosome 22 curated annotations.
ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz; Genomescan predictions from NCBI.
http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz; Mouse ENSEMBL annotations file.
http://blast.wustl.edu; Washington University BLAST Archives
http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz; Human ENSEMBL annotations file.
http://genome.cse.ucsc.edu; UCSC genome browser.
http://www.ensembl.org; ENSEMBL genome browser.
http://www1.imim.es/genepredictions/; GENEID and SGP2 full data predictions.
http://www1.imim.es/software/sgp2/; SGP2 home page.
http://www1.imim.es/datasets/sgp2002/; SGP2 training data sets page.

Received November 4, 2002; accepted in revised form November 15, 2002.