BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 12 2007, pages 1468–1475
doi:10.1093/bioinformatics/btm133

Genome analysis

Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron–exon structure

Avril Coghlan* and Richard Durbin

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1HH, UK

Received on January 29, 2007; revised on March 28, 2007; accepted on March 30, 2007

Advance Access publication May 5, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Correct gene predictions are crucial for most analyses of genomes. However, in the absence of transcript data, gene prediction is still challenging. One way to improve gene-finding accuracy in such genomes is to combine the exons predicted by several gene-finders, so that gene-finders that make uncorrelated errors can correct each other.

Results: We present a method for combining gene-finders called Genomix. Genomix selects the predicted exons that are best conserved within and/or between species in terms of sequence and intron–exon structure, and combines them into a gene structure. Genomix was used to combine predictions from four gene-finders for *Caenorhabditis elegans*, by selecting the predicted exons that are best conserved with *C.briggsae* and *C.remanei*. On a set of ~1500 confirmed *C.elegans* genes, Genomix increased the exon-level specificity by 10.1% and sensitivity by 2.7% compared to the best input gene-finder.

Availability: Scripts and Supplementary Material can be found at http://www.sanger.ac.uk/Software/analysis/genomix

Contact: alc@sanger.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The number of genomes being sequenced is increasing, making automatic prediction of genes all the more important (Brent, 2005). At present, the most accurate gene-finders are those that use transcript data to predict genes (Guigó *et al.*, 2006). However, many genomes being sequenced lack extensive, or in some cases any, transcript data. For example, whole-genome sequencing projects for at least a dozen nematode species are under way (Liolios *et al.*, 2006), and none of these will have nearly as much transcript data as the model organism *Caenorhabditis elegans*.

A straightforward way to improve gene-finding accuracy is to combine the results of several gene-finders, to take advantage of the fact that different gene-finders are good at predicting different types of genes. In this way, gene-finders that make uncorrelated errors can correct each other (Dietterich, 1997). Burset and Guigó (1996) realized this when they ran nine different gene-finders, and noticed that although the predictions were poorly correlated at the nucleotide level, only 1% of exons were missed by all the programs. A combined gene set is most likely to improve accuracy if the input gene-finders predict different sets of genes correctly and incorrectly (Ali and Pazzani, 1996).

Several different combiners have been inspired by Burset and Guigó's observations (Allen *et al.*, 2006; Howe *et al.*, 2002; Murakami and Takagi, 1998; Pavlović *et al.*, 2002; Schiex *et al.*, 2001; Shah *et al.*, 2003; Yada *et al.*, 2003; Zhang *et al.*, 2003; Supplementary Table 1). The combiner software JIGSAW (Allen *et al.*, 2006) performed as well or better than any of the other entries in the EGASP competition (Guigó *et al.*, 2006). In that competition, JIGSAW's predictions for the human ENCODE regions had 81% sensitivity and 89% specificity at the exon level.

Combiners for gene-finders generally have two elements: a method of scoring predicted features such as exons or splice sites, and a method of combining the highest-scoring features into a gene structure. Most combiners score predicted exons using the confidence scores provided by the input gene-finders for the predicted exons. However, such confidence scores are often poorly correlated with exon-level accuracy (Rogic *et al.*, 2001).

Several recently published combiners also use information on the conservation level of predicted features. For example, Zhang *et al.* (2003)'s combiner takes gene predictions, and predicted conserved splice sites, start and stop codons, and bases as input. Similarly, JIGSAW can use predicted conserved regions as an input (Allen *et al.*, 2006). The advantage of using predicted conserved features as input to a combiner is that sequence conservation is a good indicator of whether a predicted feature is likely to be real (Parra *et al.*, 2003; Ureta-Vidal *et al.*, 2003).

Here we present a new method for combining the results of several gene-finders, called Genomix. Genomix selects the predicted exons that are best conserved within and/or between species in terms of sequence and intron–exon structure, and combines them into a gene structure. Genomix differs from previous combiners that use conservation information in that it uses conservation as its primary score, and uses conservation of exon boundaries as well as of exon sequence.

*To whom correspondence should be addressed.

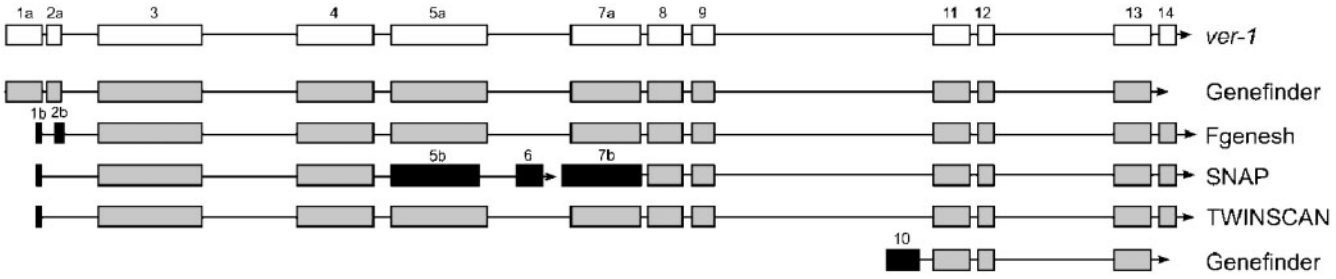

Fig. 1. The predictions for the *C.elegans ver-1* gene from different gene-finders form one 'exon cluster'. The gene model shown at the top has been experimentally confirmed. Genomix aims to select the subset of predicted exons that are most likely to be correct, and join them into a frame-consistent gene structure. For *ver-1*, although none of the input gene-finders predict the correct gene structure, Genomix predicts the correct structure by selecting all the correct predicted exons (grey) and no incorrect predicted exons (black).

Here we describe how Genomix was used to combine exons predicted in the *C.elegans* genome by four different gene-finders. We measured the accuracy of Genomix's predictions on a test set of ~1500 confirmed *C.elegans* genes, and found that Genomix increases the exon-level specificity by 10.1% and sensitivity by 2.7% over the best input gene-finder. Compared to JIGSAW, a state-of-the-art combiner (Allen *et al.*, 2006), Genomix has higher specificity by 3.5% (Fisher's test: $P<10^{-9}$) but slightly lower sensitivity (by 0.7%; Fisher's test: $P=0.1$) when the same core set of input predictions are used.

2 METHODS

Exon and gene predictions that have sequence similarity matches in related species are more likely to overlap real genes than those that lack similarity matches (Parra *et al.*, 2003). We surmised that predicted exons that have intron–exon boundaries that are conserved with homologous exons are more likely to be correct than predicted exons that have non-conserved intron–exon boundaries. Thus, we have developed a method for combining the outputs of several gene-finders, by assuming that the most plausible predicted exons are those that are best conserved within or between species in terms of their sequence and intron-exon boundaries.

2.1 Input gene sets

We refer to the species for which we want to make gene predictions as the 'query species'. To predict genes in a piece of genomic DNA from the query species (e.g. from *C.elegans*) or for its whole genome, our program requires gene predictions from multiple gene-finders for the DNA sequence. It takes GFF (Gene Feature Format; R. Durbin and D. Haussler; http://www.sanger.ac.uk/Software/GFF) files of gene predictions as input. It also requires predictions from multiple gene-finders for one or more homologous genomic regions, either from the same species and/or from one or more related species (e.g. *C.briggsae* and *C.remanei*). In the absence of genomic sequence from related species, it is possible to run Genomix by using input predictions from the query species alone. (In this case, Genomix selects the predicted exons that are best conserved between paralogous genes in the query species.)

2.2 Overall strategy of Genomix

The first step in our method is to divide the input DNA sequences from the query species and related species (if any) into regions that are likely to contain just one or a few genes. Then, for each genomic region defined in the query species:

(i) we identify its top homologous region, which can be either an orthologous region in a related species, or a paralogous region in the query species;
(ii) for each predicted exon in the genomic region, we calculate a score that reflects its conservation relative to the exons in the top homologous region;
(iii) we use dynamic programming to select the best conserved (top-scoring) predicted exons in the query region, and combine them into a gene structure.

We explain each of these steps in more detail below.

2.3 Exon clusters

Different gene-finders often split one gene into several predictions (for example, the SNAP predictions for *ver-1* in Fig. 1), or merge several genes into one prediction. To divide a piece of input DNA into regions that are likely to contain one or just a few genes each, we followed the approach used by Murakami and Takagi (1998) to identify 'clusters' of predicted exons along the input DNA. That is, two or more exons on the same strand are put in the same 'exon cluster' if they overlap, or if more than one gene prediction program placed them together in a gene prediction. For example, an exon cluster may consist of overlapping predictions for the *C.elegans ver-1* gene (Fig. 1). In practice each exon cluster usually contains 1–3 genes. Exon clusters are identified along each contig or chromosome in each species. We will refer to the exons in a query species exon cluster X as x_1, $x_2 \ldots x_n$, and the exons in an exon cluster Y in a related species as y_1, $y_2 \ldots y_m$.

Gene-finders sometimes predict the incorrect reading frame for an exon. Thus, all of the exons in each exon cluster are translated in each of their three possible reading frames. Any exon translations that contain internal stop codons are discarded.

2.4 Finding matching exon clusters

For each exon cluster X in the query species, we identify matching exon clusters within and between species. In other words, we identify exon clusters corresponding to genes that are paralogs or orthologs of the gene(s) in the query exon cluster. To do this, we run BLASTP (Altschul *et al.*, 1997) to compare the exon translations from the query species (e.g. *C.elegans*) to a database of all translated exons from the query species and related species (e.g. *C.briggsae* and *C.remanei*). BLASTP is run using an *E*-value cut-off of 0.1.

In general, the exons in a query exon cluster have BLASTP hits to exons from several different exon clusters from the query species and related

species. A score is assigned to the match between a query exon cluster and each of its matching exon clusters, by summing the bit scores for the highest scoring BLASTP hits (HSPs) between their exons. For each query exon cluster, the top-scoring matching cluster is identified, and the others are discarded. The top-scoring matching cluster could correspond to a gene that is either a paralog or an ortholog of the gene in the query exon cluster.

2.5 Exon sequence conservation

We can rapidly identify matching exon clusters by using BLASTP to search for matching exons. The BLASTP bit score gives a rough measure of the sequence conservation between a predicted exon x_i in a query exon cluster X and an exon y_j in the top matching exon cluster Y. However, BLASTP does not give a very accurate measure of sequence conservation, especially for short or poorly conserved exons. Furthermore, as Zhang *et al.* (2003) pointed out, BLAST does not identify the limits of the region of similarity between two exons very well.

To obtain a more accurate measure of the sequence conservation between each pair of exons x_i and y_j, we use PRSS (version 3.4t25; Pearson, 1996) to calculate a *P*-value P_{ij} for the significance of the Smith-Waterman alignment between their amino acid sequences (Fig. 2). We use the BLOSUM50 scoring matrix and 200 shuffles in PRSS, and since indels are relatively rare within coding exons, high gap-open and gap-extension penalties (−15 and −10) are used. The PRSS *P*-value is transformed into a sequence similarity score $S1_{ij}$ for exons x_i and y_j using the equation:

$$S1_{ij} = -\log_{10}(P_{ij}) - b_1$$

where b_1 is a constant that determines the threshold used for the PRSS *P*-value. For example, a value of 2 for b_1 corresponds to a PRSS *P*-value threshold of 10^{-2}. Our method uses parameters b_1–b_{10} and we discuss how the values for these were set below.

2.6 Intron–exon boundary conservation

If the sequence similarity score for exons x_i and y_j is significant ($S1_{ij} > 0$), we also calculate an 'exon-boundary similarity score' ($S2_{ij}$), which reflects how well the exon boundaries of exons x_i and y_j are conserved:

$$S2ij = I_t{*}b_2 + (I_s + I_e){*}b_3 + 30 - ((O_s + O_e){*}b_4)$$

where:

(a) I_t is the total number of positions in the PRSS alignment of exons x_i and y_j that are identical and are flanked by two identical positions and b_2 is a constant;

(b) I_s and I_e are the numbers of identities in the first ten and last ten PRSS alignment positions, and b_3 is a constant;

(c) O_s and O_e are the number of overhanging residues at the start and end of the PRSS alignment and b_4 is a constant. We only penalize up to 15 overhanging residues at the start or end: if $O_s > 15$ then we set $O_s = 15$, and if $O_e > 15$ we set $O_e = 15$.

If exons x_i and y_j have high sequence identity, and if there are no overhanging residues at the start (or end) of the PRSS alignment, then the intron–exon boundary at the start (or end) of exons x_i and y_j is inferred to be highly conserved.

2.7 Exon phases and length conservation

If exons x_i and y_j have significant sequence similarity ($S1_{ij} > 0$), we calculate a 'phase similarity score' $S3_{ij}$ that reflects to what extent their phases are conserved. We will use the term 'exon start phase' to refer to the phase of an exon's 5′-flanking intron, and the term 'exon end phase' to refer to the phase of the exon's 3′-flanking intron. $S3_{ij}$ is initialized to zero, and is increased by a constant b_5 if the start phases of exons x_i and y_j are the same and/or by b_5 if their end phases are the same.

Even if an exon's sequence is not conserved, its length may still be conserved. Thus, if the sequence similarity score for exons x_i and y_j is not significant ($S1_{ij} \leq 0$), a score $S4_{ij}$ is calculated that reflects whether their lengths are conserved. If the difference between their lengths is ≤3 bp, we set $S4_{ij}$ equal to a constant b_6 (otherwise $S4_{ij} = 0$).

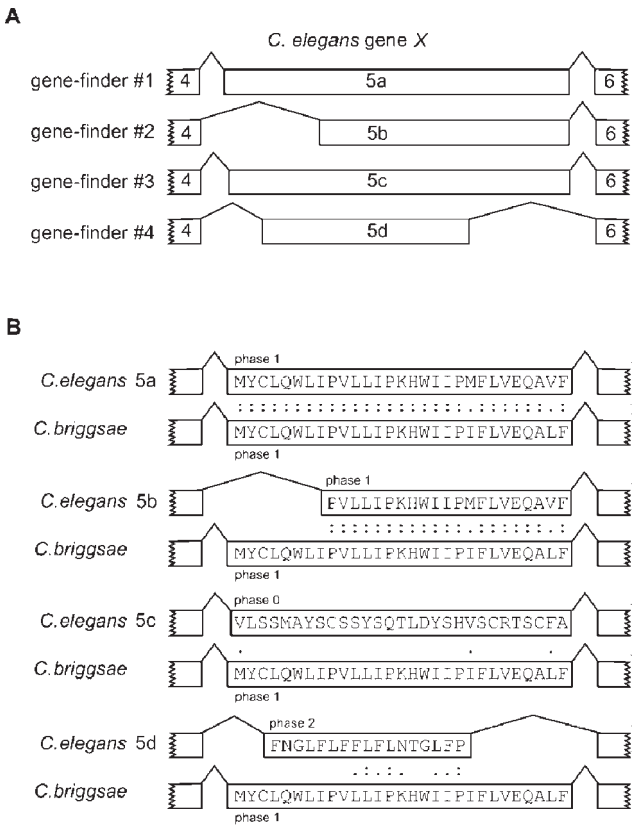

Fig. 2. Measuring exon conservation. (**A**) Four different gene-finders predict different coordinates for the fifth exon of *C.elegans* gene *X*. (**B**) Genomix calculates a score for each of the four predicted *C.elegans* exons, which reflects how conserved is its sequence, intron–exon boundaries, phases and length relative to the *C.briggsae* exons in a matching exon cluster. *Caenorhabditis elegans* predicted exon 5a has the highest conservation score, followed by 5b, then 5c and 5d.

2.8 Total conservation score

We calculate a total conservation score S_{ij} for exons x_i and y_j as:

$$S_{ij} = S1_{ij} + S2_{ij} + S3_{ij} + S4_{ij}$$

The matrix of scores S describes the similarities between each exon in a query exon cluster and each exon in a matching exon cluster. Matrix S is used as a scoring matrix in the dynamic programming step of Genomix, as described below.

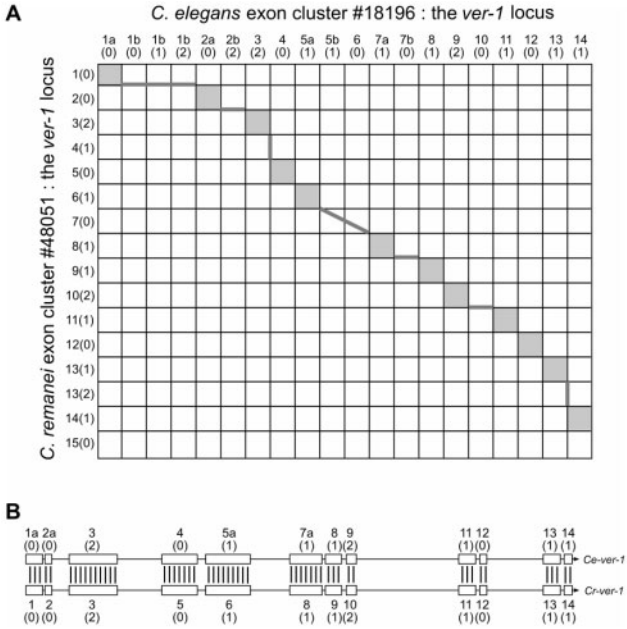

Fig. 3. Using dynamic programming to select predicted exons. (**A**) To select the best subset of exons predicted in the *C.elegans* query exon cluster (here exon cluster 18196, which corresponds to the *C.elegans ver-1* locus), dynamic programming is used to find the optimal alignment between the *C.elegans* exons and the exons in the top matching exon cluster (here *C.remanei* exon cluster 48 051, which corresponds to the *C.remanei ver-1* locus). (**B**) The solution of the dynamic programming algorithm is the optimal alignment between the predicted exons from the query *C.elegans* exon cluster and the predicted exons from the matching *C.remanei* exon cluster.

2.9 Dynamic programming

For each query exon cluster, dynamic programming is used to search for the set of predicted query exons whose added conservation scores are maximized under their reading frame constraints (Fig. 3). The exons in the query exon cluster and in its top matching exon cluster are first sorted from 5′ to 3′. We then proceed to fill the dynamic programming matrix D from top left to bottom right, by calculating D_{ij} as:

$$D_{ij} = S_{ij} + \max(D_{kl} + B_7 + B_8 + B_9 + B_{10})$$

where:

(i) $k \leq i$ and $l \leq j$, but $!(k = i, l = j)$.

(ii) exon x_k must be in the same reading frame as exon x_i, and exon y_l must in the same reading frame as exon y_j.

(iii) if $k < i$, then query exon x_k must end >30 bp before query exon x_i begins in the chromosomal DNA. This prevents us from predicting introns of <30 bp, which would probably be shorter than the minimum length for an intron to be spliced correctly (Deutsch and Long, 1999). Likewise, if $l < j$, then exon y_l must end >30 bp before exon y_j begins in the chromosomal DNA.

(iv) if $(k = i, l! = j)$, then the PRSS alignment between the amino acid sequences of exons x_i and y_l, and the PRSS alignment between exons x_i and y_j, must not overlap with respect to the amino acid sequence of exon x_i.

(v) B_7 is a score designed to favour exons that were predicted by the same input gene-finder. If $k! = i$ and if exons x_k and x_i were predicted as adjacent exons in a gene by at least one of the input gene-finders, then $B_7 = b_7$, where b_7 is a constant. $B_7 = 0$ otherwise.

(vi) B_8 is a score designed to favour predicted conserved initial exons. If exons x_k and y_l have significant sequence conservation ($S1_{kl} > 0$), and if they are both possible initial exons (start in phase 0 with Met), then $B_8 = b_8$, where b_8 is a constant. Otherwise $B_8 = 0$.

(vii) B_9 is a score designed to favour predicted conserved terminal exons. If exons x_i and y_j have significant sequence conservation ($S1_{ij} > 0$), and if they are both terminal exons (end in phase 0 with a stop codon), then $B_9 = b_9$, where b_9 is a constant. $B_9 = 0$ otherwise.

(viii) B_{10} is a penalty designed to disfavour a very long predicted intron between exons x_k and x_i that is much longer than the median *C.elegans* intron length (65 bp). $B_{10} = 0$ if the intron length is <250 bp; and $B_{10} = 0.75 * b_{10}$ if the intron is 250–500 bp, where b_{10} is a constant. For introns larger than 500 bp, B_{10} increases in steps of b_{10} every 250 bp, from $B_{10} = b_{10}$ for introns of 500–750 bp to $B_{10} = 5 * b_{10}$ for introns of ≥1500 bp.

After calculating matrix D, the optimal alignment between the exons in the query exon cluster and the exons in the top matching exon cluster is found by tracing back through matrix D. We start at the cell (i,j) that has the $\max(D_{ij})$. This is the best score for an alignment between predicted exons $x_1 \ldots x_i$ from the query exon cluster and predicted exons $y_1 \ldots y_j$ from the top matching exon cluster. We trace back through the matrix until a cell for which $D_{ij} \leq 0$ is reached. The exons selected by the dynamic programming algorithm form Genomix's final gene prediction(s) for the query exon cluster. Note that this process can, and frequently does, generate multiple genes and can also generate partial genes, although these are rarer. To make gene predictions for a whole genome, for example, for *C.elegans*, dynamic programming is used to select the best predicted exons in each query exon cluster, by comparison to the top matching exon cluster, for example from *C.briggsae*, *C.remanei* or *C.elegans*.

Dynamic programming can also be used to select the best predicted exons in a query exon cluster by comparison to two, three or more top matching exon clusters. For example, we can select the best conserved predicted exons in the *C.elegans ver-1* locus by comparison to the predicted exons in the *C.remanei ver-1* locus using 2D dynamic programming (Fig. 3), or by comparison to the predicted exons in both the *C.remanei ver-1* and *C.briggsae ver-1* loci by using 3D dynamic programming.

In a post-processing step, Genomix discards gene predictions that lack any high-scoring exon (having exon score $S \geq 1$). This is designed to discard gene predictions that correspond to pseudogenes or gene fragments.

2.10 Output gene set

The output gene set is in GFF format. A score S is given to each exon predicted by Genomix: this is the maximum conservation score S_{ij} observed for exon x_i when compared to all other exons y_j in the top matching exon cluster.

2.11 Optimizing Genomix parameters

The parameters b_1–b_{10} were hand-tuned by starting with prior estimates and iteratively adjusting these parameters to improve Genomix's prediction accuracy on the training set of 381 genes (described below). Using this manual tuning strategy, we chose parameter values of $b_1 = 1$, $b_2 = 3$, $b_3 = 0.3$, $b_4 = 1.5$, $b_5 = 5$, $b_6 = 2$, $b_7 = 1$, $b_8 = 3$, $b_9 = 10$, and $b_{10} = -0.2$.

2.12 Training set and test set

A total of 4025 *C.elegans* genes that each have one coding transcript that has been confirmed by mRNA or EST alignment (and are not known to be alternatively spliced) were downloaded from WormBase (version WS147; Schwarz *et al*., 2006). Out of these genes, 381 were randomly chosen as a training set for tuning parameters during development of our algorithm. Excluding the training set genes, 1534 of the remaining genes were randomly selected as a test set for assessing the accuracy of Genomix. Note that not all of the 4025 confirmed genes were used for the training and test sets; some were kept back in case an extra test set was needed. To measure Genomix's specificity in intergenic regions, we randomly selected 1179 intergenic regions from the ~20 000 intergenic regions in the whole genome. Any predicted exons in gene predictions that lie completely in these intergenic regions were counted as false positives. The test set genes and intergenic regions span ~6% of genic DNA and ~6% of intergenic DNA in *C.elegans*.

2.13 Input gene sets

We analysed *C.elegans* gene predictions for the WormBase WS147 release of the genome (Schwarz *et al*., 2006). TWINSCAN predictions (Korf *et al*., 2001) were downloaded from http://mblab.wustl.edu. Several other gene-finders were also run on the *C.elegans* genome: Genefinder (release 980504; P. Green, unpublished data), FGENESH (Salamov and Solovyev, 2000), and SNAP (version 2005-10-24; Korf, 2004). For FGENESH the nematode-specific trans-splicing (-n) option was used.

We analysed gene predictions for the cb25.agp8 *C.briggsae* genome assembly (Stein *et al*., 2003). *C.briggsae* gene sets made using FGENESH and Genefinder during the *C.briggsae* genome project (Stein *et al*., 2003) were downloaded from WormBase. TWINSCAN predictions were downloaded from http://mblab.wustl.edu.

Gene predictions based on version 041227 of the *C.remanei* genome were analysed (GenBank accession AAGD01000000). Gene sets that were made as part of the *C.remanei* genome project using FGENESH, Genefinder and SNAP were downloaded from WormBase. TWINSCAN predictions were downloaded from http://mblab.wustl.edu.

2.14 JIGSAW predictions

Version 3.2.5 of JIGSAW (Allen *et al*., 2006) was downloaded from http://www.cbcb.umd.edu/. We made two different gene sets using JIGSAW, each with different input data:

(i) FGENESH, Genefinder, SNAP and TWINSCAN predictions for *C.elegans*;
(ii) The four gene sets, plus BLAT alignments of *C.elegans* mRNAs to the *C.elegans* genome, and predicted splice sites.

We will refer to these gene sets as JIGSAW_1 and JIGSAW_2. The positions of the best BLAT alignments of *C.elegans* mRNAs to the genome were downloaded from WormBase (release WS147). To predict the positions of splice sites in the *C.elegans* genome, the Genefeatures program by R. Durbin was used. Genefeatures is based on part of the Genefinder software (P. Green, unpublished data) and is available as part of the AceDB database software (Durbin and Thierry-Mieg, 1994 and http://www.acedb.org). Predicted splice sites that were assigned confidence scores of $\geq$3.0 by Genefeatures were used as input for JIGSAW. For each of the two JIGSAW gene sets, JIGSAW was trained using the 'oblique splits' option on our training set of 381 fully confirmed *C.elegans* genes and then was run on the whole *C.elegans* genome.

Table 1. Comparison of the accuracy of the input gene sets to that of the combined gene sets made by Genomix and JIGSAW, for a test set of 1534 confirmed *C.elegans* genes and 1179 intergenic regions

Method	Base Sn	Base Sp	Exon Sn	Exon Sp	ME	WE	Gene Sn	Gene Sp	MG	WG
Genefinder	97.5	81.5	86.7	72.7	3.7	19.4	53.8	51.4	2.6	3.6
TWINSCAN	96.1	83.6	88.8	77.2	4.8	17.2	62.8	57.7	2.8	9.1
SNAP	96.2	85.4	84.2	73.3	5.3	17.2	53.3	42.1	2.0	15.9
FGENESH	97.1	80.8	86.4	72.3	4.0	20.0	55.3	50.0	1.4	7.0
Genomix	97.2	91.9	91.5	87.3	3.7	8.1	69.2	67.4	3.5	2.7
JIGSAW_1	98.2	87.2	91.8	82.1	2.9	13.2	70.5	64.1	1.3	8.4
JIGSAW_2	98.2	88.7	92.2	83.8	2.7	11.6	72.1	67.0	1.5	6.7

Standard measures of predictive accuracy are given: Sn–sensitivity; Sp–specificity; ME–proportion of true exons for which there is no overlapping predicted exon; WE–proportion of predicted exons that do not overlap any true exon; MG–proportion of true genes for which there is no overlapping predicted gene; WG–proportion of predicted genes that do not overlap any true gene.

3 RESULTS

The accuracy of Genomix was assessed on a test set of 1534 randomly chosen *C.elegans* genes, and 1179 randomly chosen intergenic regions (see Methods Section). These cover 6% of the total genic DNA and 6% of the intergenic DNA in the *C.elegans* genome. Since we do not expect any genes in the intergenic regions, predictions within these regions are counted as false positives.

3.1 Combining four gene-finders using Genomix

Genomix was used to combine predictions from FGENESH (Salamov and Solovyev, 2000), Genefinder (P. Green, unpublished data), SNAP (Korf, 2004) and TWINSCAN (Korf *et al*., 2001) for *C.elegans*. Conservation with *C.elegans* paralogs and homologs from *C.briggsae* and *C.remanei*, was used to identify the most conserved *C.elegans* predicted exons. That is, each *C.elegans* exon cluster was compared to its top matching exon cluster from either *C.elegans*, *C.briggsae* or *C.remanei*.

Gene prediction accuracy was measured in terms of sensitivity and specificity at the nucleotide, exon and gene level (Burset and Guigó, 1996; Table 1). The exon-level sensitivity is the fraction of real exons predicted correctly by a gene prediction program. For an exon prediction to be considered correct, both the 5′ and 3′ boundaries must match the true exon exactly. The exon level sensitivities of the input gene-finders were 86.4% for FGENESH, 86.7% for Genefinder, 84.2% for SNAP and 88.8% for TWINSCAN (Table 1). Only 96.4% of the real exons in the test set of 1534 *C.elegans* genes were predicted correctly by at least one of FGENESH, Genefinder, SNAP or TWINSCAN. As for other combiners that focus on exons as the indivisible elements to be combined, it is impossible for Genomix to predict an exon that was missing from all the input predictions. As a result, Genomix's maximum possible sensitivity is 96.4%. Genomix has 91.5% sensitivity at the exon level, i.e. it has 2.7% higher sensitivity than the most sensitive of the four input gene-finders. Furthermore, Genomix outperforms the individual gene-finders in terms of gene-level sensitivity by 6.4% (increasing gene-level

sensitivity from ≤62.8 to 69.2%). Here a test set gene is considered to have been predicted correctly if the prediction made for the gene contains all of its real exons and no false-positive exons.

The exon-level specificity is the fraction of the predicted exons that are real. Taking into account false-positive whole-gene predictions completely contained in the 1179 test set intergenic regions, the exon level specificities for the input gene-finders were 72.3% for FGENESH, 72.7% for Genefinder, 73.3% for SNAP and 77.2% for TWINSCAN (Table 1). Genomix has 87.3% exon-level specificity, which is 10.1% higher than that of TWINSCAN, the input gene-finder with the highest specificity. In addition, in terms of gene-level specificity Genomix performs better than the best of the input gene-finders, TWINSCAN, by 9.7%.

3.2 Comparison to JIGSAW

The accuracy of Genomix was compared to that of JIGSAW (Allen *et al*., 2006), a combiner software that performed as well or better than any of the other entries in the EGASP competition (Guigó *et al*., 2006). For EGASP, Allen *et al*. (2006) used JIGSAW to combine six different gene-finders, mRNA and EST alignments, curated genes and predicted conserved elements.

To see how well JIGSAW performs using just raw gene predictions as input, a JIGSAW gene set was made by combining the FGENESH, Genefinder, SNAP and TWINSCAN predictions ('JIGSAW_1'). JIGSAW had 92% exon-level sensitivity and 82% exon-level specificity. This is roughly the same as the exon-level sensitivity of Genomix (92%), but slightly less than Genomix's exon-level specificity (87%; Table 1).

However, while JIGSAW's input can be restricted to the outputs of several gene-finders, it performs better if its input also includes other data sources, especially alignments of mRNAs (Allen *et al*., 2006). Thus, to make a fair comparison of JIGSAW's and Genomix's accuracies when supplied with their optimal input data, a second JIGSAW gene set was made by combining the four input gene-finders, plus predicted splice sites and mRNA alignments. At the exon level, this second JIGSAW gene set ('JIGSAW_2') had 92.2% sensitivity and 83.8% specificity. Compared to Genomix, the JIGSAW_2 gene set has 0.7% higher exon-level sensitivity (Fisher's test: $P = 0.1$; McNemar's test: $P = 0.02$), but 3.5% lower exon-level specificity (Fisher's test: $P < 10^{-9}$).

3.3 Suggested changes to WormBase curated genes

Genomix sometimes assigns a higher score to a predicted exon than to an overlapping confirmed exon in the test set. We surmised that some of these high-scoring predicted exons could be extensions to existing confirmed exons, or alternative transcripts. To investigate this, we examined exons to which Genomix assigns a score that is >100 higher than the score that it assigns to the overlapping test set exon. For example, Genomix assigns a score of 215 to the confirmed initial exon of *T20D3.5*, but assigns a score of 401 to an overlapping exon predicted by FGENESH and TWINSCAN. In its final prediction for that locus, Genomix selected the FGENESH/TWINSCAN exon and also selected an additional upstream exon. We realized

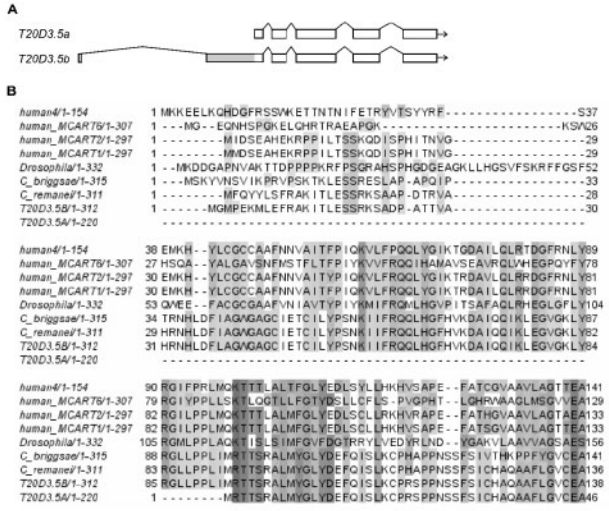

Fig. 4. An alternative isoform of a WormBase curated gene that was suggested by Genomix. (**A**) WormBase release WS147 contained one confirmed transcript for gene *T20D3.5*, now known as *T20D3.5a*. Genomix suggested an alternative isoform *T20D3.5b*. The grey box indicates the extra upstream coding region of *T20D3.5b* that is missing from *T20D3.5a*. (**B**) A multiple alignment of *T20D3.5a*, *T20D3.5b* and their *C.briggsae*, *C.remanei*, *Drosophila melanogaster* and human orthologs. The alignment is truncated after the start of *T20D3.5a*, since *T20D3.5a* and *T20D3.5b* are identical after this point. The human and *D.melanogaster* orthologs were identified from TreeFam (Li *et al*., 2006). Here 'human4' is Ensembl gene *ENSG00000189332*. Gene predictions for the *C.briggsae* and *C.remanei* orthologs were made using Genomix.

that, compared to the confirmed transcript from WormBase WS147, the FGENESH/TWINSCAN exon and extra initial exon align further upstream to the orthologs in human and fly. This information was sent to WormBase, who curated our suggested gene structure as alternative isoform *T20D3.5b* and renamed the original confirmed transcript as *T20D3.5a* (Fig. 4).

We also examined exons predicted by Genomix that did not overlap any curated WormBase WS147 exon, but that had high Genomix conservation scores of >100. There were 4373 such conserved exons, belonging to 2418 gene predictions that do not overlap any curated WormBase gene or annotated pseudogene. Of these 2418 putative new genes, 106 have ≥3 exons with conservation scores of >100, have valid start and stop codons, and contain coding DNA that is <10% repetitive. Two examples that we examined had already been added to WormBase since release WS147 (independently of our predictions): *Y46E12A.4* and *T07C4.11*. A third example was a novel *C.elegans* gene prediction that does not have similarity to any curated *C.elegans* gene but shows high sequence conservation (61% identity) with the *C.briggsae* and *C.remanei* orthologs predicted by Genomix. We informed the WormBase curators about this putative gene, and it has been accepted into the next WormBase release (WS169) as *R01H2.8*. We are compiling a list of the most convincing new genes and exons suggested by Genomix to give to WormBase.

4 DISCUSSION

Predicting genes in eukaryotic DNA is still a challenge for the best gene-finders. We have presented a method for combining predictions from two or more gene-finders, which successfully improves prediction accuracy. Genomix should prove useful in providing accurate gene predictions for the increasing number of genomes that are being sequenced, and especially where transcript data is not available to aid gene prediction, for example in the many groups of related invertebrate species currently being sequenced. These include arthropods (12 *Drosophila* species), platyhelminths (*Schistosoma mansoni* and *S.japonicum*), cnidarians (the sea anemone *Nematostella* and coral *Acropora*), mollusks (bivalves *Spisula* and *Mytilus* and gastropods *Lottia*, *Aplysia* and *Biomphalaria*), and annelids (*Helobdella* and *Capitella*) (Liolios *et al*., 2006).

In a large test set of *C.elegans* genes, Genomix improves the exon-level sensitivity by 2.7% and the exon-level specificity by 10.1% compared with the input gene-finders. Genomix's high exon-level specificity is due to its approach of scoring predicted exons according to their conservation, which allows it to discard many false positive non-conserved exons that were predicted by the input gene-finders. Genomix also correctly predicts some exons that the most accurate gene-finder misses. At the exon level TWINSCAN has 2.1% higher sensitivity and 3.9% higher specificity than FGENESH, Genefinder or SNAP, but Genomix can still improve on TWINSCAN's accuracy by combining it with the other three gene-finders (Table 1). In fact, despite the fact that Genomix does not make use of EST or mRNA data, its exon-level accuracy is close to the 92% sensitivity and 87% specificity achieved by TWINSCAN_EST (on a different data set), which uses EST data (Wei and Brent, 2006).

The recent combiner GLEAN (Elsik *et al*., 2007) is the first combiner for gene-finders that does not require any training set. This is a huge advantage for annotating newly sequenced genomes that lack EST or cDNA resources. In contrast, in order for Genomix to achieve its best performance on a new genome, its parameters b_1, $b_2 \ldots b_{10}$ should be re-tuned for the new genome. However, we found that Genomix performs quite well on an unseen genome even if parameters b_1, $b_2 \ldots b_{10}$ are not re-tuned. We used the version of Genomix that was tuned for *C.elegans* to make gene predictions for *D.melanogaster* by combining the output of three different gene-finders, and found that Genomix increased exon-level sensitivity by 2.3% and exon-level specificity by 3.5% over the best input gene-finder (Supplementary Table 2). That is, while it is advisable to re-tune Genomix parameters, in the absence of any training data it is safe to assume that Genomix will produce reasonably good predictions without re-tuning. This is probably due to the fact that Genomix tries to avoid the need for a large training set by using PRSS (Pearson, 1996) to directly estimate the probability that a predicted exon overlaps a real coding exon. That is, we use PRSS to calculate the *P*-value for the significance of the alignment between the amino acid sequence of a predicted exon and a homologous exon from the same or a related species. The *P*-values calculated by PRSS for amino acid sequences are very accurate (Brenner *et al*., 1998; Pearson, 2000), and so give us a reliable estimate of the probability that a predicted exon has conserved amino acid sequence. As a result, predicted exons and genes that are assigned non-significant *P*-values by PRSS are often wrong (do not overlap any real exon). This probably underlies Genomix's good performance with respect to completely wrong exon predictions (WE = 8.1%) and completely wrong gene predictions (WG = 2.7%; Table 1).

The fact that Genomix assigns a conservation score to each predicted exon separately makes it possible to use input gene sets that are correlated without biasing the results. This is because an exon predicted in correlated gene sets will only be assigned a high score if it is highly conserved. As a result, it is possible, for example, to combine predictions from a single gene-finder using two different parameter settings.

Genomix only works for genes that have sequenced homologs in the same species or in a closely related species. However, this does not affect sensitivity too badly. While Genomix does completely miss some whole genes that are not conserved, these are only ~3.2% of test set genes (Genomix's total MG = 3.5%, of which 0.3% is due to genes completely missed by the four input gene-finders). Likewise, Genomix completely misses some non-conserved exons in otherwise conserved genes, but this only occurs for ~1% of test set exons.

Many genes in animal genomes are thought to undergo alternative splicing (Kan *et al*., 2001). Ten percent of *C.elegans* genes are annotated as having alternative splicing in the coding region, and this is probably an underestimate. Predicting alternative splicing is a very important challenge for the gene-finding field (Guigó *et al*., 2006). So far, only a few combiner programs such as EuGène (Foissac and Schiex, 2005) predict alternative transcripts. In its current implementation, Genomix predicts only a single isoform for each gene. The test set analysed here consists of genes that each has a single isoform. However, for genes that do undergo alternative splicing, Genomix will miss many real alternative isoforms. An important future direction is to develop Genomix so that it can predict alternative isoforms, for example by making use of mRNA/transcript data.

ACKNOWLEDGEMENTS

This project was supported by The Wellcome Trust and an EMBO long-term fellowship to A.C. We thank John Spieth (Washington University GSC) for kindly allowing us to use *C.remanei* sequence data and gene predictions, Bill Pearson for advice on running PRSS, Jonathan Allen for advice on running JIGSAW, Venky Iyer for *Drosophila* gene sets, and Des Higgins (University College Dublin) for hosting A.C. in his group during part of this work. We are grateful to members of the Durbin research group and WormBase for useful discussions, and to two anonymous reviewers, as well as to Noel O'Boyle, Alan Moses, Jean-Karim Hériché, Li Heng and David Carter for helpful comments on the manuscript.

Conflict of Interest: none declared.

REFERENCES

Ali,K.M. and Pazzani,M.J. (1996) Error reduction through learning multiple descriptions. *Machine Learning*, **24**, 173–206.

Allen,J.E. *et al*. (2006) JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol*., **7** (Suppl. 1), S9.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Brenner,S.E. *et al*. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*., **95**, 6073–6078.

Brent,M.R. (2005) Genome annotation past, present and future: how to define an ORF at each locus. *Genome Res*., **15**, 1777–1786.

Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Deutsch,M. and Long,M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*., **27**, 3219–3228.

Dietterich,T.G. (1997) Machine-learning research: four current directions. *The AI Magazine*, **18**, 97–136.

Durbin,R. and Thierry-Mieg,J. (1994) The ACeDB Genome Database. In: Suhai,S. (ed.) *Computational Methods in Genome Research*. Plenum Press, New York, pp. 45–56.

Elsik,C.G. *et al*. (2007) Creating a honey bee consensus gene set. *Genome Biol*., **8**, R13.

Foissac,S. and Schiex,T. (2005) Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics*, **6**, 25.

Guigó,R. *et al*. (2006) EGASP: the human ENCODE genome annotation assessment project. *Genome Biol*., **7** (Suppl. 1), S2.

Howe,K.L. *et al*. (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res*., **12**, 1418–1427.

Kan,Z *et al*. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*., **11**, 889–900.

Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Korf,I. *et al*. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17** (Suppl. 1), S140–S148.

Li,H. *et al*. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*., **34**, D572–D580.

Liolios,K. *et al*. (2006) The genomes on line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*., **34**, D332–D334.

Murakami,K. and Takagi,T. (1998) Gene recognition by combination of several gene-finding programs. *Bioinformatics*, **14**, 665–675.

Parra,G. *et al*. (2003) Comparative gene prediction in human and mouse. *Genome Res*., **13**, 108–117.

Pavlović,V. *et al*. (2002) A bayesian framework for combining gene predictions. *Bioinformatics*, **18**, 19–27.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol*., **266**, 227–258.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol*., **132**, 185–219.

Rogic,S. *et al*. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res*., **11**, 817–832.

Salamov,A.A. and Solovyev,V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res*., **10**, 516–522.

Schiex,T. *et al*. (2001) EUGENE: An eukaryotic gene finder that combines several sources of evidence. *Lecture Notes in Computer Science*, **2066**, 111–125.

Schwarz,E.M. *et al*. (2006) WormBase: better software, richer content. *Nucleic Acids Res*., **34**, D475–D478.

Shah,S.P. *et al*. (2003) Genecomber: combining outputs of gene prediction programs for improved results. *Bioinformatics*, **19**, 1296–1297.

Stein,L.D. *et al*. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol*., **1**, E45.

Ureta-Vidal,A. *et al*. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet*., **4**, 251–262.

Wei,C. and Brent,M.R. (2006) Using ESTs to improve the accuracy of gene prediction. *BMC Bioinformatics*, **7**, 327.

Yada,T. *et al*. (2003) DIGIT: a novel gene finding program by combining gene-finders. *Pac. Symp. Biocomput*., **8**, 375–387.

Zhang,L. *et al*. (2003) Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res*., **13**, 1190–1202.