BIOINFORMATICS

Vol. 18 no. 8 2002
Pages 1034–1045

# Improving gene recognition accuracy by combining predictions from two gene-finding programs

Sanja Rogic[1,*], B. F. Francis Ouellette[2] and Alan K. Mackworth[3]

[1]Computer Science Department, The University of California at Santa Cruz, Baskin Engineering, Santa Cruz, CA 95064, USA, [2]Centre for Molecular Medicine and Therapeutics, Children's and Women's Health Center of British Columbia, UBC, Vancouver, B.C., Canada V5Z 4H4 and [3]Computer Science Department, The University of British Columbia, 2366 Main Mall, Vancouver, B.C., Canada V6T 1Z4

Received on June 10, 2001; revised on February 16, 2002; accepted on February 22, 2002

## ABSTRACT

**Motivation:** Despite constant improvements in prediction accuracy, gene-finding programs are still unable to provide automatic gene discovery with desired correctness. The current programs can identify up to 75% of exons correctly and less than 50% of predicted gene structures correspond to actual genes. New approaches to computational gene-finding are clearly needed.

**Results:** In this paper we have explored the benefits of combining predictions from already existing gene prediction programs. We have introduced three novel methods for combining predictions from programs Genscan and HMMgene. The methods primarily aim to improve exon level accuracy of gene-finding by identifying more probable exon boundaries and by eliminating false positive exon predictions. This approach results in improved accuracy at both the nucleotide and exon level, especially the latter, where the average improvement on the newly assembled dataset is 7.9% compared to the best result obtained by Genscan and HMMgene. When tested on a long genomic multi-gene sequence, our method that maintains reading frame consistency improved nucleotide level specificity by 21.0% and exon level specificity by 32.5% compared to the best result obtained by either of the two programs individually.

**Availability:** The scripts implementing our methods are available from http://www.cs.ubc.ca/labs/beta/genefinding/

**Contact:** rogic@cse.ucsc.edu

## INTRODUCTION

In this era of intensive genomic sequencing, when millions of nucleotides of genomic DNA are sequenced daily, tools for interpreting the content of these genomes are more important than ever. The first step in deciphering the DNA sequence information is finding all the genes contained in a sequence and elucidating their structure. Although many gene-finding programs have been developed in the past 10 years and their prediction accuracy is constantly improving, we are still far away from completely automatic gene discovery with 100% accuracy. Current programs, although very good at discovering the majority of coding nucleotides (more than 90% predicted correctly) and moderately good in discovering exact exon boundaries (70–75% of exons predicted correctly) are still weak when it comes to predicting complete gene structures: less than 50% of predicted genes correspond exactly to the actual genes (Rogic *et al.*, 2001; Dunham *et al.*, 1999). Consequently, predictions given by these programs need to be verified by other evidence such as similarity to a cDNA sequence or similarity to a known protein or EST sequence. However, in many cases this additional evidence is not available: it has been shown that only a fraction of newly discovered genes have identifiable homologs in the current databases (Dunham *et al.*, 1999). *Ab initio* gene-finding remains the only available computational approach for identifying novel genes that do not have detectable similarities to known proteins and hence their predictions have significant effect on our understanding of the genomes and on future experimental directions. Therefore, improving the accuracy of these programs is essential and would lead to faster, cheaper and, above all, more accurate interpretation of sequenced genomes. In this paper we present a new approach to combining the prediction results of gene-finding programs in order to obtain better prediction accuracy.

Current gene prediction programs are sophisticated systems that integrate many different methods for identifying elements of the genes. *Content sensors* are coding statistics capable of distinguishing between coding and non-coding regions. The one proven to be the most

*To whom correspondence should be addressed.

effective, in-frame hexamer measure (Ficket and Tung, 1992), has been used predominantly by the developers of the programs. Other coding statistics incorporated in gene–finding systems are codon usage, GC content measure and position asymmetry measure. *Signal sensors* attempt to mimic closely processes occurring within the cell. They are intended to identify DNA sequence signals, usually just several-nucleotides-long subsequences, which are recognized by cell machinery and are initiators of certain processes. The signals that are usually modeled by gene-finding programs are promoter elements, start codons, splice sites, stop codons and polyA sites. Content and signal sensors are implemented by various statistical and pattern recognition methods and integrated in overall gene models usually using machine learning techniques (hidden Markov models (HMM), neural networks, decision trees) or discriminant functions (linear or quadratic).

The set of methods used and the way they are integrated differs among individual programs, as well as the sequence training sets used to build signal models and tune programs' parameters. Being distinct in their architecture and training, programs often give different gene structure predictions for the same DNA sequence. This characteristic of programs' predictions has motivated several authors to investigate the benefits of combining several gene-finding programs.

Burset and Guigo (1996) investigated the correlation between six *ab initio* gene-finding programs that they evaluated. The approximate correlation ($AC$) of the predictions at the nucleotide level varied from 49% to 68% and the average exon accuracy varied from 24% to 47%, when predictions from two programs were compared, indicating that the programs are not tightly correlated especially at the exon level. The exons predicted by all of the programs tested were correct in 99% of cases, suggesting that an exon prediction, which is unanimous among the programs, is almost certainly guaranteed to be correct. On the other hand, the proportion of exons completely missed by any of the programs was 1%, showing that each program can contribute to finding all annotated exons. These basic 'and' (intersection) and 'or' (union) approaches increase only one component of exon accuracy, either specificity or sensitivity, while the other one becomes significantly decreased. In order to improve overall accuracy sensitivity and specificity have to be increased simultaneously.

A more comprehensive study of methods for combining gene-finding programs was done by Murakami and Takagi (1998). They used five different methods to combine four gene-finding programs: FEXN (Solovyev *et al.*, 1994), GeneParser3 (Snyder and Stormo, 1995), Genscan (Burge and Karlin, 1997) and Grail2 (Uberbacher and Mural, 1991). The methods they tested were: the AND-based method, the OR-based method, the HIGHEST method, the RULE method and the BOUNDARY method. The first two methods are similar to Burset and Guigo's approach of accepting only regions predicted as coding by all of the programs (AND-based method) or accepting regions predicted by at least one of the programs (OR-based method). The other three methods use estimated exon probabilities to decide on the exon candidates. While approximate correlation was significantly improved when FEXN, GeneParser3 and Grail2 were combined by some of the methods (up to 10%), improvements were more marginal when Genscan was used because it was more difficult to outperform Genscan's high prediction accuracy. The best result of this analysis was a 4.7% increase of $AC$ and a 2.5% increase of average exon accuracy comparing to the best individual program (Genscan).

Here, we present three different methods for combining the predictions from gene-finding programs. The methods integrate previously described 'and' and 'or' approaches using the exon scores given by the programs. Rather then combining several gene-finding programs including those with generally low prediction accuracy we decided to combine only two programs with high accuracy using their prediction scores. Relying on the results of a comprehensive evaluation of recently developed programs (Rogic *et al.*, 2001), Genscan and HMMgene (Krogh, 1997) were chosen for their high prediction accuracy and their reliable estimate of the correctness of the exon prediction. The methods developed improve prediction accuracy at both nucleotide and exon levels, with some tradeoffs when tested on a long genomic sequence. The improvements are generally higher at the exon level, where a 7.9% increase of average exon accuracy was achieved when only the best results obtained by the combination methods and the two gene-finding programs were compared. Although our methods yielded improvements in both exon sensitivity and specificity, the latter was more significantly improved (11.7% compared to HMMgene and 22.9% compared to Genscan). When tested on a long genomic multi-gene sequence, our method that maintains reading frame consistency improved nucleotide level specificity by 21.0% and exon level specificity by 32.5% compared to the best result obtained by any program.

## SYSTEM AND METHODS

The previous experience with combining predictions from different gene-finding programs has shown that the simple OR- and AND-based methods can significantly improve one aspect of prediction accuracy. The OR-based method, which returns regions predicted by any of the programs used, will assure more sensitive prediction, i.e. will identify more exons than any single program. However, it will have decreased specificity since many low-quality exons without support from multiple programs will also be accepted. Analogously, the AND-based method, which

returns regions predicted by more than one program will increase specificity by eliminating low-quality, low-support predictions, but it will also miss many valid predictions made by only one program.

It seems that having a reliable estimate of predicted exon accuracy would help us improve these two approaches: doing the union only on the predicted regions of the higher quality and doing the intersection for low-quality regions. In order to test this new combined approach it is necessary to use gene-finding programs that have high prediction accuracy and, even more importantly, reliable accuracy estimates for their predictions. Relying on the results of our recent evaluation of gene-finding programs (Rogic *et al.*, 2001) we have chosen programs Genscan and HMMgene, which have the best overall prediction accuracy and are the only two that have reliable exon scores.

## Characteristics of the selected programs

Both Genscan and HMMgene model the structure of genomic sequence as an explicit state duration HMM (Rabiner, 1989), which is also known as a generalized HMM (Kulp *et al.*, 1996). In this type of probabilistic model each state of the model has an associated arbitrary length distribution. In the case of DNA sequences, the states of the HMM model functional elements of the genes or genomic regions: intergenic region, promoter, 5′ and 3′ untranslated regions, exons and introns and polyA site. The states in Genscan and HMMgene HMMs are probabilistic models themselves. Both programs use different types of HMMs to model coding and non-coding regions. For signal detection Genscan uses the weight matrix method, the weight array method (WAM), windowed WAM and maximal dependence decomposition. The details of signal modeling in HMMgene are not available.

Although both Genscan and HMMgene use the same underlying model, the training procedure they employ is different: while Genscan uses the traditional maximum likelihood approach to train the HMM, HMMgene uses a criterion called conditional maximum likelihood, which maximizes the probability of correct prediction.

Both programs were trained using human single- and multi-exon genes collected by D. Kulp and M. Reese (the dataset can be found at http://www.fruitfly.org/sequence/human-datasets.html) but Genscan's training set for coding region model was supplemented with a set of complete human cDNA sequences.

The programs are capable of predicting any number of single- or multi-exon genes, which can be complete or partial. The output of both programs contains information about exon location and type, their probabilistic score and reading frame.

In our analysis we used version 1.0 of Genscan with the HumanIso.smat parameter file and version 1.1d of HMMgene. Programs were run locally on a SUN Ultra 60 workstation, under the Solaris 5.6 operating system.

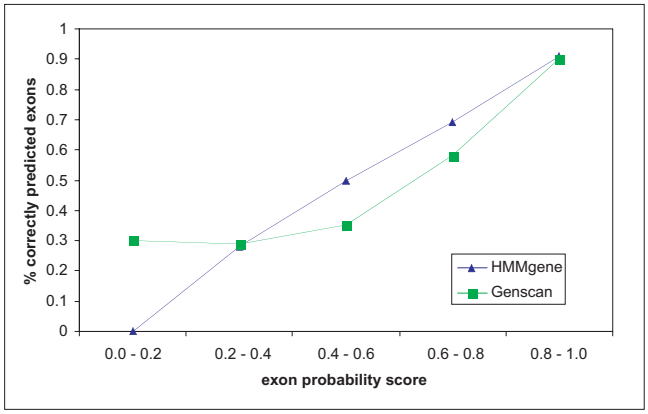


**Fig. 1.** Reliability of Genscan and HMMgene proportion of correctly predicted exons versus exon probability scores.

## Exon probability scores

In our earlier study of gene-finders (Rogic *et al.*, 2001), which involved seven recently developed programs: FGENES (Solovyev *et al.*, 1995), GeneMark.hmm (Lukashin and Borodovsky, 1998), Genie (Kulp *et al.*, 1996), Genscan, HMMgene, Morgan (Salzberg *et al.*, 1998) and MZEF (Zhang, 1997), we also examined the reliability of the prediction scores given by a program. Each of the programs evaluated in this study, except GeneMark.hmm, has a scoring scheme for its exon prediction. However, our analysis has shown that only the Genscan and HMMgene exon probability scores, which give the quantitative measure of the likelihood that the given exon is correct, are meaningful and reliable. Figure 1 shows the relationship between Genscan and HMMgene exon probability scores and the proportion of exactly predicted exons. We can see that there is an approximate linear dependence between these two variables for both programs and that the proportion of exactly predicted exons monotonically increases with the increase of exon probability score (disregarding a small anomaly for Genscan). This means that the exons with the higher scores are usually more accurate then the exons with lower scores and in the case of HMMgene the likelihood of correct prediction is almost perfectly estimated with the exon score. This characteristic of Genscan and HMMgene exon probability scores makes them very useful guides in deciding on the correctness of a predicted exon.

## Correlation between the programs

Even though it appears that Genscan and HMMgene are good candidates for the suggested combination approach

**Table 1.** Exon predictions by a pair of programs—For each pair of the programs we calculated the number of exons predicted correctly by at least one of the two programs. In parenthesis are the number correct predictions that are identical between two programs and the number of false positive exons that overlap. The numbers on the diagonal are the results for individual programs (the number of false positive exons is in parenthesis). The analysis was done for the HMR195 dataset, which has 948 annotated exons.

| | FGENES | GeneMark.hmm | Genie | Genscan | HMMgene |
|---|---|---|---|---|---|
| FGENES | 697 (154) | | | | |
| GeneMark.hmm | 799 (523; 38) | 625 (120) | | | |
| Genie | 824 (524; 33) | 773 (503; 31) | 651 (126) | | |
| Genscan | 840 (592; 33) | 796 (564; 42) | 807 (579; 34) | 735 (104) | |
| HMMgene | 825 (587; 29) | 811 (529; 34) | 793 (573; 31) | 826 (624; 32) | 715 (81) |

it is also necessary to investigate how correlated their predictions are. For our purposes correlation between predictions is good as long as they are correct predictions. However, in order for our approach to be successful it is necessary that each of the programs has a set of correctly predicted exons that were not identified by the other program. It is also important that their wrong exons, i.e. exons that do not overlap any real exon, do not coincide.

We have carried out this analysis for five of the seven programs evaluated in our previous study. For each pair of programs we calculated the number of exons predicted exactly (both exon boundaries predicted correctly) by at least one of the programs. The results are given in Table 1. The numbers on the diagonal are for individual programs. Comparing these numbers with the numbers off the diagonal it is apparent that any pair of programs can predict more exons correctly than any single program. The most successful pair in this respect is Genscan/FGENES, followed by Genscan/HMMgene.

The table also shows the number of correct predictions that are identical between two programs and the number of wrong exons that overlap. We can see that Genscan and HMMgene have the most coinciding correct predictions, but the number of their overlapping wrong exons is not any higher than for the other pairs. Thus, although Genscan and HMMgene exon predictions are highly correlated, as one could have conjectured from similarity of their architectures and training datasets, it appears that they agree when they are right and rarely when they are wrong. For the purposes of a proposed combination method this is very acceptable behavior, because according to Figure 1 wrong exons tend to have lower exon probability scores and thus would be accepted only if they were predicted by both programs.

The only pair of programs that surpasses Genscan/HMMgene in the number of correctly predicted exons is Genscan/FGENES. FGENES uses dynamic programming to find the optimal combination of exons, promoters, and polyA sites detected by a pattern recognition algorithm. Considering that their underlying models are different and the fact that different datasets were used for their training it can be suspected that Genscan and FGENES will tend to produce less correlated predictions. Thus, testing the accuracy of our methods using Genscan and FGENES appears very appealing. Unfortunately, as discussed in Rogic *et al.* (2001), FGENES's scores are not very informative and thus FGENES does not meet the most basic requirement for our prediction combination approach.

## Sequence datasets used

The HMR195 sequence dataset described in Rogic *et al.* (2001), which contains 195 human, mouse and rat sequences, was used to develop and test methods for combining Genscan and HMMgene predictions. To ensure independent testing of the methods' performance two additional control datasets were also used: the Burset/Guigo dataset and a *Drosophila melanogaster Adh* region used in the Genome Annotation Assessment Project (GASP) (Reese *et al.*, 2000).

The dataset assembled by Burset and Guigo (1996) consists of 570 vertebrate genomic sequences containing exactly one multi-exon gene. Similarly to the HMR195 dataset it has been filtered to exclude anomalous sequences.

The *Drosophila melanogaster Adh* region is 2.9 Mb long and has been extensively studied for the last 20 years (Ashburner, 2000). For the GASP experiment two different annotation sets were used to evaluate the gene-finding programs' predictions: st1 and st3. The first set, called st1, contained only highly accurate annotations, confirmed by aligning full-length cDNA sequences from this region with the high-quality genomic sequences. This approach left out many potential genes that did not have a matching cDNA. st1 originally contained annotation for 43 transcripts, but after some incorrect sequences were removed, the number of genes is 38. The second and more complete annotation set, st3, containing 222 gene structures, was compiled by biology experts using information from various sources: BLAST results, PFAM alignments, high scoring Genscan and Genefinder predictions, ORFFinder results, full-length cDNA alignments and alignments with genes from GenBank. Out of 222 annotated genes only 40 were based solely on strong Genscan and Genefinder predictions.

### Combining the programs' predictions

In this section we describe methods for integrating Genscan and HMMgene predictions. Our goal was to provide computationally straightforward techniques for combining the output from these two programs. On the assumption that they can be considered partially independent sources of evidence for gene structure, it should be possible to use output from the programs as follows: when either program is quite confident of its exon prediction use it regardless; in the cases where both programs are less certain of their exon prediction use it if they both agree.

Both Genscan and HMMgene produce output files for each DNA sequence submitted. The output files give the details of the gene structure predictions made by the programs. Each file contains enumerated exons with their location, type and probability score. Exons are labeled according to the gene they belong to. The methods that we describe below use this information to decide on the candidate exons. We present three different algorithms EUI (Exon Union–Intersection), GI (Gene Intersection) and EUI_frame (Exon Union–Intersection with Reading Frame Consistency):

#### Algorithm EUI (Exon Union–Intersection)

(1) Consider all the Genscan and HMMgene exons that have exon probability score greater or equal to a threshold $p_{\text{th}}$. The regions predicted by at least one of the programs are labeled as EUI exons (exon union—see Figure 2).

(2) Consider all the Genscan and HMMgene exons that have exon probability score less than $p_{\text{th}}$. The regions predicted by both programs are labeled as EUI exons (exon intersection—see Figure 2).

Consequently, a Genscan or HMMgene exon that does not overlap any exon predicted by the other program will be accepted if its exon probability is greater or equal to $p_{\text{th}}$ and refused otherwise.

There is one exception for step 1: if Genscan's internal exon has the same right boundary (donor site) as HMMgene's initial exon (both exons have the score $\geqslant p_{\text{th}}$) choose HMMgene's exon prediction as an EUI exon. This 'initial exon rule' was incorporated into the EUI method after our analysis showed that Genscan often predicts initial exons as internal, which have the correct donor site but false acceptor site preceding the true ATG codon. HMMgene's predictions of the initial exons are shown to be more accurate.

#### Algorithm GI (Gene Intersection)

(1) For each program's prediction select regions predicted as genes (genes are treated as continuous sequence from the beginning of the first predicted exon in the gene to the end of the last predicted exon). Regions predicted by both programs are labeled as GI genes (gene intersection—see Figure 2).

(2) Apply the EUI method to those exons that completely belong to GI genes (where both exon boundaries are within a GI gene).

This approach is primarily designed for identification of genes in long genomic regions where another level of constraints, namely considering only exons that belong to regions predicted as genes by both programs, helps further eliminate numerous wrong exons typical for *ab initio* predictions in the long sequences.

#### Algorithm EUI _frame (Exon Union–Intersection with Reading Frame Consistency)

This method applies the EUI method to Genscan and HMMgene predictions while maintaining reading frame consistency:

(1) For each program's prediction determine the gene boundaries and to each gene assign a gene probability calculated as the average of exon probability scores for all the exons contained in that gene. For each predicted exon determine the positions of acceptor and donor site in a reading frame of a gene it belongs to.

(2) If the gene predicted by Genscan overlaps the gene predicted by HMMgene, choose the one with the higher gene probability to impose the reading frame. Apply the EUI method to the exons belonging to the selected genes accepting EUI exons only if they are in the chosen reading frame.

The threshold value $p_{\text{th}}$ that is used in all three methods has been empirically derived using the HMR195 dataset. The optimal value is $p_{\text{th}} = 0.775$. However, the methods' accuracy results show very low sensitivity to the threshold variation, as can be observed in Figure 3. The average exon accuracy varies from 0.78 to 0.81 for EUI method and 0.79 to 0.82 for the GI method when the threshold value changes from 0.45 to 0.95. For both methods $(ESn + ESp)/2$ peaks when $p_{\text{th}}$ is between 0.75 and 0.80, and accordingly the average of these two values is chosen to be the threshold value.

The scripts implementing the EUI, GI and EUI_frame methods are available from http://www.cs.ubc.ca/labs/beta/genefinding/. We are also developing a web based annotation tool that will apply our methods, as well as Genscan and HMMgene programs, to a query sequence and return results in GFF (gene-finding format) and graphical formats.

## RESULTS

Accuracy measures for the three methods as well as for Genscan and HMMgene on the HMR195 dataset are given

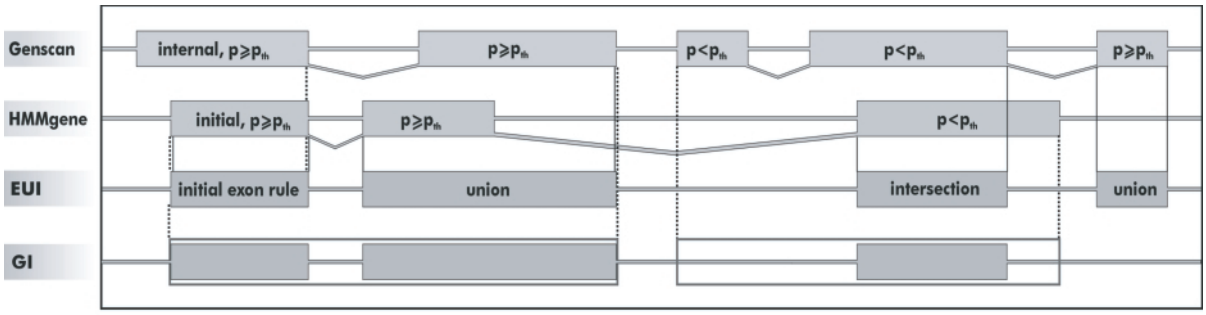


**Fig. 2.** Graphical representations of EUI and GI methods. The dotted lines mark the boundaries of the GI genes and the solid lines mark the boundaries of EUI exons. The labels on the EUI exons indicate which part of EUI algorithm was used to determine the exon.

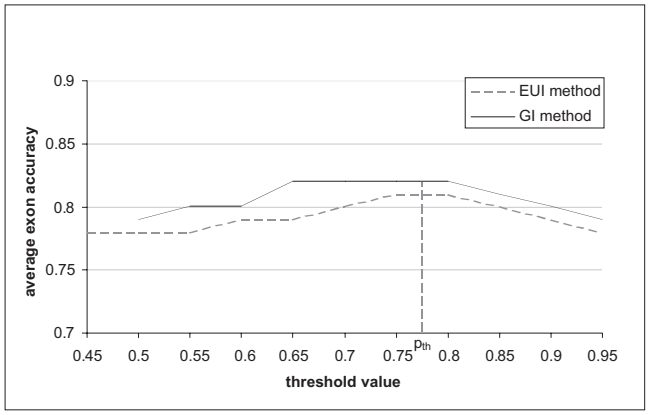


**Fig. 3.** Threshold sensitivity—Average exon accuracy versus threshold value. The optimal threshold value is $p_{\text{th}} = 0.775$.

in Table 2. The numbers in bold indicate an improvement when compared to either of the two programs. It can be observed that each of the methods outperforms Genscan and HMMgene in all categories except for nucleotide level sensitivity and proportion of missed exons. The results in Table 2 suggest that each of three methods improve specificity more than sensitivity at both the nucleotide and exon levels. While sensitivity is decreased at the nucleotide level from 0.95 for Genscan to 0.91–0.94 for the methods, specificity is increased from 0.93 for HMMgene to 0.96 for GI method (3.2% increase) and 0.95 for EUI and EUI_frame methods (2.2% increase). At the exon level, sensitivity increased from 0.76 for HMMgene to 0.78 for all the methods (2.6% increase), while specificity increased from 0.77 for HMMgene to 0.86 for GI (11.7% increase), 0.83 for EUI_frame (7.8% increase) and 0.82 for EUI method (6.5% increase). These numbers also imply that improvements are substantially better at the exon level than at the nucleotide level, which is also supported by an increase of 2.2% in $AC$ and an increase of 7.9% in $(ESn + ESp)/2$, when comparing only the highest accuracy values for the programs and the methods. While the number of missed exons was not improved by either of the methods, the number of wrong exons was substantially decreased: Genscan predicted 104 wrong exons, HMMgene 81 and the GI method only 44.

Results for the Burset/Guigo control set are summarized in Table 3. Bold numbers in Table 3 have the same pattern of appearance as in Table 2, which indicates that improvements are accomplished in the same categories. Similarly to the results in Table 2 improvements are better for specificity at both levels and generally better for exon level measures than for nucleotide level measures. The increases in accuracy values for this dataset were somewhat lower than for the HMR195 dataset.

The results on the 3 Mb *Adh Drosophila* region are shown in Table 4. The values for $Sn$, $ESn$ and $ME$ are calculated using annotation set st1 and the values for $Sp$, $ESp$ and $WE$ are calculated using st3. The rationale for this lies in the way these sets are built: st1 contains a subset of all genes in the *Adh* region that are correct in the details, while the st3 dataset is believed to be complete but the confidence in its correctness is not as high as for the st1 dataset. Thus, sensitivity, which is the measure of how well a program can predict the real coding features in a sequence, is more accurately estimated from st1 because we are sure that these annotations are correct. On the other hand, specificity, which is the measure of how well a program avoids false positive predictions, is better estimated from st3, which is thought to be complete. Similarly to the results for the previous two datasets, the three introduced methods have improved specificity more than sensitivity. At the nucleotide level specificity increased from 0.62 for Genscan to 0.75 for the GI and EUI_frame methods (21.0% increase) and 0.69 for the EUI method (11.3% increase), while the sensitivity values for the methods were less than or equal to the ones for the

**Table 2.** Results for HMR195—For each sequence in the HMR195 test set, the forward (+) strand exons in the default outputs of the programs tested were compared to the annotated exons. The standard measures of predictive accuracy on nucleotide and exon level were calculated for each sequence and averaged over all sequences for which they were defined: *Sn*—sensitivity on nucleotide level; *Sp*—specificity, *AC*—approximate correlation; *ESn*—exon level sensitivity; *ESp*—exon level specificity; *ME*—proportion of real exons that were not predicted by a program; *WE*—proportion of predicted exons that do not overlap with any of the actual exons. The second column gives the number of sequences where no prediction was made. The numbers in parenthesis in the last two columns are actual numbers of missed and wrong exons, respectively.

| *Methods* | *# no* | *Nucleotide accuracy* | | | *Exon accuracy* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *prediction* | Sn | Sp | *AC* | *ESn* | *ESp* | $(ESn + Esp)/2$ | *ME* | *WE* |
| Genscan | 3 | 0.95 | 0.90 | 0.91 | 0.70 | 0.70 | 0.70 | 0.08 (76) | 0.09 (104) |
| HMMgene | 5 | 0.93 | 0.93 | 0.91 | 0.76 | 0.77 | 0.76 | 0.12 (128) | 0.07 (81) |
| EUI | 3 | 0.94 | **0.95** | **0.93** | **0.78** | **0.82** | **0.80** | 0.10 (104) | **0.04** (55) |
| GI | 15 | 0.91 | **0.96** | **0.92** | **0.78** | **0.86** | **0.82** | 0.19 (149) | **0.03** (43) |
| EUI_frame | 3 | 0.93 | **0.95** | **0.93** | **0.78** | **0.83** | **0.80** | 0.11 (115) | **0.03** (46) |

**Table 3.** Results for Burset/Guigo dataset—For each sequence in the Burset/Guigo test set, the forward (+) strand exons in the default outputs of the programs tested were compared to the annotated exons and the standard measures of accuracy calculated.

| *Methods* | *# no* | *Nucleotide accuracy* | | | *Exon accuracy* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *prediction* | Sn | Sp | *AC* | *ESn* | *ESp* | $(ESn + Esp)/2$ | *ME* | *WE* |
| Genscan | 8 | 0.94 | 0.93 | 0.92 | 0.78 | 0.81 | 0.80 | 0.09 (203) | 0.05 (188) |
| HMMgene | 38 | 0.93 | 0.94 | 0.92 | 0.81 | 0.83 | 0.82 | 0.14 (308) | 0.04 (139) |
| EUI | 20 | 0.94 | **0.96** | **0.93** | **0.83** | **0.88** | **0.85** | 0.12 (250) | **0.03** (98) |
| GI | 43 | 0.91 | **0.97** | **0.93** | **0.82** | **0.90** | **0.86** | 0.18 (386) | **0.02** (67) |
| EUI_frame | 27 | 0.93 | **0.96** | **0.93** | **0.83** | **0.88** | **0.85** | 0.13 (286) | **0.03** (87) |

programs. At the exon level specificity increased from 0.40 for Genscan to 0.49 for GI (22.5% increase) and 0.53 for EUI_frame (32.5% increase), while sensitivity increased by 6.8%, (from 0.59 to 0.63) for EUI method and slightly decreased for the rest two methods when compared to the programs' best sensitivity result $ESn = 0.59$. The EUI method has the lowest $ME$ among the programs and the methods, while GI and EUI-frame have missed six and five more exons than Genscan, respectively. The last column in Table 4, showing the proportion of the wrong exons, illustrates the most important advantage of our methods over Genscan and HMMgene when used on a long genomic region: the number of false positive exons decreased from 873 for Genscan and 1379 for HMMgene to 631 for EUI, 366 for the GI and 318 for EUI_frame methods. The overall high numbers for $WE$ are the result of a known shortcoming of gene-finding programs: overpredicting exons and genes in long stretches of genomic sequences (Dunham *et al.*, 1999).

The results for HMMgene shown in Table 4 differ from those shown in Reese *et al.* (2000) and Krogh (2000) for two reasons: first, the results that we report are only for *ab initio* gene-finding without using any of the additional sources of evidence, which has been incorporated in

**Table 4.** Results for ***Drosophila Adh* region**—*Sn*, *ESn* and *ME* are reported for st1 annotation set and *Sp*, *ESp* and *WE* are reported for st3 annotation set. All the methods are tested on the both strands of *Adh* region.

| Methods | *# of predicted exons* | Nucleotide accuracy | | Exon accuracy | | | |
|---|---|---|---|---|---|---|---|
| | | Sn | Sp | *ESn* | *ESp* | *ME* | *WE* |
| Genscan | 1696 | 0.96 | 0.62 | 0.59 | 0.40 | 0.14 (15) | 0.51 (873) |
| HMMgene | 2101 | 0.95 | 0.61 | 0.49 | 0.19 | 0.14 (16) | 0.66 (1379) |
| EUI | 1376 | 0.96 | **0.69** | **0.62** | 0.40 | **0.13** (14) | **0.46** (632) |
| GI | 1043 | 0.92 | **0.75** | 0.56 | **0.49** | 0.19 (21) | **0.35** (366) |
| EUI_frame | 912 | 0.83 | **0.75** | 0.55 | **0.53** | 0.23 (25) | **0.35** (318) |

HMMgene for GASP purposes (Krogh, 2000) and second, the st1 standard set that we used is a refined version of the set used for the original GASP evaluation.

## DISCUSSION

Our analysis in Rogic *et al.* (2001) shows that the weakest component of the current gene-finding programs is signal detection, especially the detection of initiation and termination codons, which lowers the exon level prediction accuracy. Considering that the exon level sensitivity ($ESn$) is defined as a proportion of true exons (exactly predicted exons) to actual exons and specificity ($ESp$) as a proportion of true exons to predicted exons, it is obvious that the number of true exons is directly proportional to $ESn$ and $ESp$. Therefore, if the correct splice site is missed, even by just a couple of nucleotides, the predicted exon will not be counted as a 'true' exon, which simultaneously decreases $ESn$ and $ESp$. Thus, the exon prediction accuracy could be improved in two ways: identifying the correct exon boundaries would increase the number of 'true' exons, at the same time increasing both the exon sensitivity and specificity, and reducing the number of predicted exons (PE) would increase exon specificity. Of course, only the dismissal of the falsely predicted exons would be beneficial for the overall increase in $ESp$.

The EUI method, which is also incorporated in the other two methods introduced above, attempts to simultaneously find more probable exon boundaries and to discard the low-confidence exons. As shown in Burset and Guigo (1996) and Murakami and Takagi (1998), selecting the union of the exons predicted by two programs (OR-method) would result in increased sensitivity but decreased specificity and analogously, the intersection of the exons (AND-method) would increase specificity but decrease sensitivity. The EUI method integrates these two approaches by using them selectively depending on the confidence in exon correctness. When the probability scores for the two overlapping predicted exons are high (greater than or equal to $p_{\text{th}}$) the coding region predicted by either of the programs is chosen to be a resulting EUI exon. This potentially increases the sensitivity of the prediction, which is already supposed to be specific according to Figure 1 (proportion of the correctly predicted exons is almost equivalent to the specificity). When the exon scores for the two overlapping exons are low (less than $p_{\text{th}}$), the region predicted to be coding by both of the programs is selected to be the resulting exon, which potentially improves the specificity of the prediction. A 'stand-alone' exon that does not overlap with any exon predicted by the other program will be accepted only if it has an exon score greater or equal to $p_{\text{th}}$. This further improves exon specificity by eliminating low-probability exons that have a high chance of being wrong.

The relationship between Genscan and HMMgene prediction scores is shown in Figure 4. Each exon predicted by either of the two programs is represented by a data point in the graph. If two exons overlap they are represented by one dot whose coordinates correspond to the Genscan and HMMgene exon scores. The dots on the $x$- and $y$-axes represent exons predicted only by one program. We can distinguish three classes of exons from this scatter plot: the exons on the axes of the graph, which are 'stand-alone' exons, the exons predicted by both programs (they do not have to be identical exactly) with very high score, and the exons predicted by both programs whose scores from the two programs are not tightly correlated. This graph further emphasizes the non-correlation hypothesis for the two

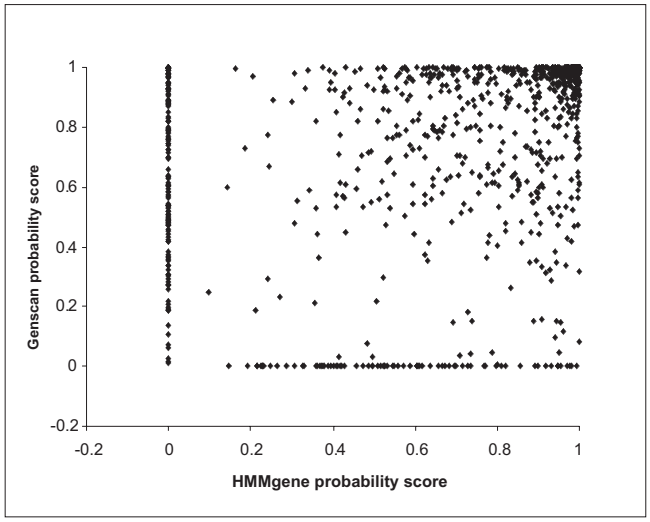


**Fig. 4.** Probability scores of all exons predicted by Genscan and HMMgene—Each exon predicted by either of the two programs is represented by a dot in the graph. If two exons overlap they are represented by one dot whose coordinates correspond to the Genscan and HMMgene exon scores. The dots on the $x$- and $y$-axes represent exons predicted only by one program.

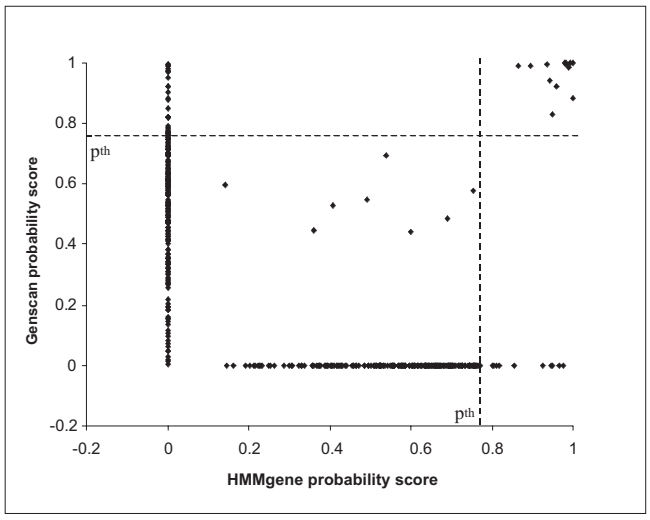


**Fig. 5.** Probability scores of all false positive exons predicted by Genscan and HMMgene.

programs: first, there are many exons predicted by only one program, as shown in Table 1, and also even if the two predictions overlap, very often their scores do not agree closely.

Figure 5 presents all the false positive exon predictions made by either program. The exons are represented in the same way as in Figure 4. Figure 5 clearly shows that most of the wrong exons predicted by one program were not predicted by the other—only 55 of 447 dots in the graph are not found on the axes. Comparing Figures 4 and 5, we can see that the false exons predicted by both programs are buried among numerous true predictions and it appears to be impossible to distinguish them using solely the exon scores. However, the exons plotted on the axes of the graph in Figure 5 can be easily excluded if we choose to keep only the exon predicted by both programs. This is exactly what the EUI algorithm is doing, except that it also retains all the 'stand-alone' exons with the probability greater than the threshold $p_{\text{th}}$. Figure 5 shows that dense clusters of dots on the axes of the graph are terminated around $p_{\text{th}}$ and there are fewer false positives with a score higher than $p_{\text{th}}$. The value for $p_{\text{th}}$ determines the trade-off between sensitivity and specificity and by choosing $p_{\text{th}} = 0.775$ we are making them as balanced as possible.

The EUI method was primarily designed to improve prediction accuracy on the relatively short sequences containing only one gene, which resemble the sequences used for training of gene prediction programs. Genscan and HMMgene do rather well when predicting genes in these sequences: the majority of the actual exons is identified, at least partially, and the fraction of false positive exons is only around 5%. Although this results in fairly high accuracy measures at the nucleotide level, the exon level accuracy is affected by weakness of the signal detection, which often misses exact exon boundaries. In order to improve the prediction accuracy EUI attempts to correct exon boundaries using the union and intersection of exons; only a very small number of exons get discarded due to low exon scores. This approach gives more correctly predicted exons than any other method resulting in the highest exon sensitivity for each of the test datasets.

On the other hand, GI was designed for longer genomic sequences containing more than one gene, for which gene-finding programs generally make more false positive predictions. To reduce the high rate of wrong exons GI first chooses gene candidates to be those regions predicted as genes by both programs. In this algorithmic step many genes that are predicted by just one program and many exons that do not belong completely to the newly selected GI gene get eliminated. In the next step the EUI method is applied to the resulting GI genes. These two rounds of exon elimination get rid of many falsely predicted exons resulting in considerably higher specificity than both programs and the EUI method.

As can be inferred from the definition of the methods, EUI or GI exons that belong to the same gene are not guaranteed to be in the same reading frame. Frame consistency is lost when exon boundaries are changed by applying the EUI algorithm. In order to investigate the effect of frame consistency on EUI method we designed the EUI_frame method that uses the EUI algorithm to

combine the predictions from Genscan and HMMgene, while maintaining a single reading frame. The program with the highest average exon score dictates the reading frame of the final prediction: exons whose boundaries are modified by the EUI method or high scoring 'stand-alone' exons (score $\geqslant p_{\text{th}}$) will be accepted in the final prediction only if they do not disrupt the chosen reading frame. Surprisingly, this method gave almost identical results to those of EUI on HMR195 and Burset/Guigo datasets. After the analysis of the results we found that EUI_frame missed some of the exons that were correctly predicted by EUI and at the same time eliminated some of the wrong exons predicted by EUI. These differences were proportionally too small to change the overall prediction accuracy except for the slight decrease in the sensitivity. Being trained on sequences similar to those from HMR195 and Burset/Guigo datasets, Genscan and HMMgene predictions on these two datasets are fairly accurate and similar: overlapping exons are usually in the same reading frame and there are not many false positive predictions that could disrupt the reading frame. This is why EUI and EUI_frame have almost identical results on the first two datasets. However, the *Adh* region sequence is much longer than any of the training sequences and contains a couple of hundred of genes, which presents a serious challenge for any gene-finding program. Genscan and HMMgene prediction accuracy for this region is substantially weaker than for the other two datasets: while most of the coding nucleotides have been identified correctly in many cases exact exon boundaries are missed resulting in much lower exon sensitivity and specificity. The major problem is the large number of wrong exons, which results in the drastic decrease in the specificity at both levels. These characteristics of Genscan and HMMgene predictions resulted in many reading frame disruptions in the EUI genes and thus caused elimination of more than 400 of exons when EUI_frame was applied. Most of the dismissed exons were false positives, but a few of the 'true' exons were also sacrificed. The discrepancy between the EUI and EUI_frame results is notable: due to the twofold decrease in the number of wrong exons EUI_frame has substantially higher specificity at both levels than EUI, but at the same time sensitivity was decreased, especially at the nucleotide level owing to the exceptionally large size of exons missed by EUI_frame method.

By selecting more probable exon boundaries exon level accuracy is directly improved. This does not have to affect the nucleotide level accuracy significantly since the correct splice site could have been missed by just a couple of nucleotides and the correction will just slightly change $Sn$, $Sp$ and $AC$. This explains why exon level accuracy is more improved than nucleotide level accuracy, as observed in Tables 2–4. Another phenomenon, observable for all three datasets, is that specificity is improved more than sensitivity at both levels. Since it is impossible for the EUI and GI methods to predict an exon that was initially missed by both programs, which would directly improve sensitivity of the prediction, our methods attempt to improve the accuracy of the predictions by correcting the exon boundaries and eliminating potentially wrong exons. The effect of this is that EUI and GI have approximately one half as many wrong exons as the individual programs, which primarily improves $Sp$ and $ESp$.

Although Tables 2-4 show that the methods introduced have improved accuracy measures for all three datasets they were tested on, the level of improvement varies among them. The results on the Burset/Guigo dataset show the lowest increase in the accuracy measures. This dataset has been available since 1996 and it contained the vast majority of available vertebrate genomic sequences at the time it was assembled. It is realistic to assume that, in many cases, the training sets of gene-finding programs developed afterwards overlap with the Burset/Guigo dataset and this is probably the case with the training datasets of Genscan and HMMgene. This assumption is supported by the programs' high accuracy results on this dataset, shown in Table 3. Since the programs have been trained on at least a subset of Burset/Guigo dataset, their predictions are often correct and identical. Consequently, the combination of their predictions does not improve prediction accuracy as much as for the new HMR195 dataset.

The highest increase in prediction specificity is achieved on the *Adh* region. In this region the GI and EUI_frame methods have 21% higher specificity at nucleotide level, while at the exon level GI has 22.5% and EUI_frame 32.5% higher specificity when compared to Genscan's accuracy results. This unusually high increase in specificity is a direct result of decreased number of false positive predictions. In long genomic sequences, such as the sequence of the *Adh* region, gene-finding programs make many false exon predictions, which lowers specificity at both levels. The effect of this shortcoming is also observable in our tables: the specificity values for Genscan and HMMgene at both levels are substantially lower for the *Adh* region than for the other two datasets. Each of our methods succeeded in eliminating many of the wrong exons predicted by Genscan and HMMgene, EUI_frame being the most successful by having approximately one quarter of the false positive exon predictions of HMMgene. However, this substantially increased specificity was also coupled with decreased sensitivity for the GI and EUI_frame methods. The decrease was marginal at the exon level since GI and EUI_frame had just a few correctly predicted exons less than Genscan, but more substantial at the nucleotide level due to the unusually large size of the exons completely missed by the methods.

Since Genscan was used to build the st3 annotation set, it is evident that the values in *Sp*, *ESp* and *WE* columns are not truly independent results of Genscan's and our methods' performance in the long genomic regions. Although only 40 of 222 annotated genes in st3 did not have any additional evidence except for strong Genscan and Genefinder prediction, it is very likely that the authors of st3 were also relying on Genscan's exon boundaries when other evidence were available. This can be inferred from the significantly higher *ESp* (and lower *WE*) for Genscan than for HMMgene, which cannot be observed for other datasets. However, our goal is to show the performance of our methods, rather than to give an independent evaluation of the programs on the *Adh* region and for that purpose the results in Table 4 are useful, showing that even though st3 was tailored using Genscan's predictions our methods have higher accuracy than Genscan.

## CONCLUSION

We have presented three methods, EUI, GI and EUI_frame, for combining exon predictions from two gene-finding programs, Genscan and HMMgene, which successfully improve prediction accuracy, especially on long genomic sequences. The improvements have been obtained at both the nucleotide and exon levels and for all three datasets used for testing. The major advantage of our methods is the elimination of many false positive exon predictions, which directly improves the specificity at both levels.

While other sources of evidence, such as database hits to known proteins or EST matches, are indispensable in the search for genes it is definitely worthwhile improving accuracy of *ab initio* gene prediction, which is essential when other evidence is not available. Our study demonstrates that the accuracy of computational gene-finding can be improved, exploiting only currently available methods. Using Genscan and HMMgene predictions as two partially independent sources of evidence we succeeded in correcting the exon boundaries, getting more exactly predicted exons, and in eliminating many false positive exons. The three methods that we developed have different strengths and are suitable for different purposes, depending whether sensitivity, specificity or reading frame consistency is the more valued characteristic of the predictions. Or on a practical application, EUI would be best applied to shorter sequences (where only one gene is expected), whereas the GI and EUI_frame methods are best applied to longer ones (where more than one gene is expected). The methods are not only limited to Genscan and HMMgene, but can be applied to other pairs of gene-finding programs, as long as they offer reliable exon scores.

## ACKNOWLEDGMENTS

We thank Holger Hoos (UBC, Vancouver) for his helpful comments on the manuscript. This work was founded by a Research Grant from the Natural Sciences and Engineering Research Council of Canada to Alan Mackworth who holds a Canada Research Chair in Artificial Intelligence.

## REFERENCES

Ashburner,M. (2000) A biologist's view of the *Drosophila* genome annotation assessment. *Genome Res.*, **10**, 391–393.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.

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

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

Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.

Ficket,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.

Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene-finding. In Gaasterland,T., Karp,P., Karplus,K., Ouzounis,C., Sander,C. and Valencia,A. (eds), *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 179–186.

Krogh,A. (2000) Using database matches with HMMgene for automated gene detection in *Drosophila*. *Genome Res.*, **10**, 523–528.

Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 134–142.

Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene-finding. *Nucleic Acids Res.*, **26**, 1107–1115.

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

Oliver,S.G., van der Aart,Q.J., Agostoni-Carbone,M.L., Aigle,M., Alberghina,L., Alexandraki,D., Antoine,G., Anwar,R., Ballesta,J.P., Benit,P. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.

Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

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.

Rogic,S., Mackworth,A.K. and Ouellette,B.F. F. (2001) Evaluation of gene finding programs on mammalian sequences. *Genome Res.*, **11**, 817–832.

Salzberg,S., Delcher,A., Fasman,K. and Henderson,J. (1998) A decision tree system for finding genes in DNA. *J. Comp. Biol.*, **5**, 667–680.

Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1–18.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) The prediction of human exons by oligonucleotide composition and discriminant analysis of splicable open reading frames. In Altman,R., Brultag,D., Karp,P., Lathrop,R. and Serls,D. (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 354–362.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. In Rawling,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 367–375.

Uberbacher,E.C. and Mural,R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.

Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Berks,M., Bonfield,J., Burton,J., Connell,M., Copsey,T., Cooper,J. *et al.* (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature*, **368**, 32–38.

Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.