Computer programs for eukaryotic gene prediction

Abstract
Seven popular programs for gene prediction in eukaryotic organisms are described and evaluated on the basis of availability for in-house and on-line use and prediction accuracy. This report outlines generally applicable approaches to computational gene prediction and known limitations in this field.

Computational determination of coding regions in eukaryotic genomes is an important problem\(^1\),\(^2\) that has been brought into the spotlight by advances in genomic sequencing.\(^3\)–\(^5\) Since the publication of the human genome, public interest in gene finding has somewhat declined, but this change in fashion does not make the problem of computational gene prediction any less valid or useful. While it is clear that at present no computer algorithm would be able to annotate a newly sequenced genome in a fully automated fashion, gene-finding programs have substantial utility, in particular when they are combined with large volumes of experimental data.\(^6\) The goal of this review is to describe and compare the computer programs for gene finding in eukaryotic organisms that are available to the average user, and to describe limitations and considerations that arise in practice.

The following seven programs are discussed: FgeneS,\(^6\) Genie,\(^7\) GenScan\(^8\) and derived from it GenomeScan,\(^9\) HMMGene,\(^10\) geneid\(^11\) and GlimmerM.\(^12\) These are not all the gene-finding programs developed to date. Rather these are the most modern and advanced methods, able to predict multiple, complete eukaryotic gene structures in an unannotated DNA sequence, and in some cases able to include homology information in the theoretical prediction. The last five methods are also suited for prediction of so-called ‘suboptimal’ exons, paving a way for computational annotation of alternatively spliced transcripts.

In this paper, we will compare these programs on availability and accuracy as the primary criteria. Availability is evaluated in terms of:

- possibility to submit a prediction request through the web interface;
- ability to install the program locally if desired;
- ability to train and test the program independently;
- availability of source code; and
- freedom from undue licensing restrictions.

In this review, we stress availability over other possible criteria, for instance, ease of use or user-friendliness, because ultimately availability matters the most to the practising scientist. The list above may be viewed as a minimal set of guidelines for the development of research software needed to preserve one’s ability to reproduce and verify results of others. Such ability has been a cornerstone of all modern science and is absolutely crucial if the results are to be trusted.

By accuracy we mean consistent
performance of a program on a set of standard test cases. It may be assessed by two parameters: sensitivity and specificity. Definitions of these parameters in application to gene prediction traditionally follow those of Burset and Guigo\(^1\) and are illustrated in Figure 1. Sensitivity in gene prediction is a measure of the fraction of coding region that the method is able to predict correctly. On the other hand, specificity is a measure of how much of the prediction is indeed true. In a successful prediction both sensitivity and specificity should be as close to 100 per cent as possible. In a case where sensitivity is high but specificity is low, over-prediction should be suspected.

In the gene-finding application this would

![Figure 1: Sensitivity and specificity (adapted from Burset and Guigo\(^1\))]  
In the upper pane in Figure 1 it is assumed that all events are binary with values of true and false. When reality and results of some predictive method agree, we obtain either a true positive (both true) or a true negative (both false) case. These cases are designated by the TP and TN abbreviations in the picture. When the real value is true but is predicted as false, it is called a false negative (FN). The opposite case when prediction evaluates to true, but should have been false, is a false positive (FP). The lower pane shows an application of this formalism to gene prediction. Thick bars correspond to coding regions, and thin lines to non-coding ones. In a sequence every nucleotide is either coding (part of an exon, true) or non-coding (introns, regulatory elements and inter-genic regions, false). When a nucleotide in an exon is correctly predicted as coding, it counts as a true positive. It is easy to codify all other cases as shown in the figure. Sensitivity is then defined as a ratio of nucleotides correctly predicted as coding (true positives) to all truly coding nucleotides:

\[
Sn = \frac{TP}{all \ true \ in \ reality} = \frac{TP}{TP + FN}
\]

This is a measure of how much of the reality the method predicts correctly. Specificity is a ratio of nucleotides correctly predicted as coding (TP) to all nucleotides which are simply predicted to be coding, irrespective of whether this prediction is correct or not:

\[
Sp = \frac{TP}{all \ true \ in \ prediction} = \frac{TP}{TP + FP}
\]
correspond to a pathological case of just one endless exon. On the other hand, if sensitivity is low while specificity is high, the prediction is overly conservative and may lack the power to discover anything new. Sensitivity and specificity can be defined on the levels of individual nucleotides, whole exons and entire genes. In the last two cases, it is assumed that in order to count as correct the feature should have all of its components predicted correctly. For an exon it means having all of its nucleotides correctly predicted as coding. An error in just one nucleotide would negate the entire predicted exon. In turn, for a gene to be correct, all of its exons must be predicted correctly. Because of this the accuracy values reported on the levels of whole exons and genes are typically much lower than those for the nucleotide counts. A possibility of alternative splicing presents an added problem for computational gene assembly from individual exons. Assuming that the ultimate goal of gene prediction is to determine the protein sequence(s) encoded by the gene, one has to admit that it is definitely not easy.

Our findings are summarised in Tables 1, 2 and 3. In all tables the programs are ordered from best to worst in terms of the respective criterion, with the best one being on top.

AVAILABILITY
We note that availability of many gene-finding programs is severely restricted. Of the seven programs listed, only three – geneid, GlimmerM and GenScan – are available for download for local use. Only one program, geneid, is completely available according to the listed criteria. These restrictions, stemming from commercialisation of some gene-finding programs, limit their utility and acceptance by scientists. Other limitations arise due to availability of parameters for various organisms. In making a prediction all gene-finding methods rely on a large number of numeric parameters that can only be obtained in a training process. This training typically involves running another suite of programs, separately from the predictor itself, on a set of sequences for which the gene structure is known. Parameters obtained in training vary if training is performed on a different model organism. Consequently, it is only possible to use a gene prediction method on those organisms for which parameter sets have been derived, for this specific method. Most gene prediction programs are trained on human sequences, and some offer parameter sets for fruit fly and Arabidopsis thaliana. Although it is fairly usual to extend the applicability of these parameter sets to all vertebrates, all invertebrates and all green plants, respectively, there is no clear justification for doing so, and the user is advised against it. An exception among the programs listed is GlimmerM, which does not provide parameters for human genes. GlimmerM is optimised for small eukaryotes such as malaria parasite, and for some plant species (Arabidopsis thaliana and rice). When working with an organism other than those mentioned, developing a parameter set that is based on the known gene sequences for that particular organism may be worthwhile. In this respect, only two programs, geneid and GlimmerM, explicitly disclose their training data sets and procedures. As the number of organisms with known genome sequences grows, the need for re-training of gene-finding methods is expected to increase as well. Thus providing access to the training software and methodology to the end user is an important feature of gene-finding software.

ACCURACY
The accuracy of the reviewed programs is typically in the range of 90 per cent on the nucleotide level, as measured by both sensitivity and specificity. This apparent high prediction accuracy translates into somewhat lower rate of absolutely correct exons that is usually between 50 and 70 per cent and rarely exceeds 80 per cent (data not shown). Accuracy on the level of whole genes is lower. In practice this means that although most of the coding
Table 1: Availability of gene prediction software

<table>
<thead>
<tr>
<th>Program</th>
<th>URL</th>
<th>Web interface</th>
<th>Download</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geneid</td>
<td><a href="http://www1.imim.es/software/geneid/index.html">http://www1.imim.es/software/geneid/index.html</a></td>
<td>yes</td>
<td>yesa</td>
<td>yesa</td>
</tr>
<tr>
<td>GlimmerM</td>
<td><a href="http://www.tigr.org/softlab/glimmerm/">http://www.tigr.org/softlab/glimmerm/</a></td>
<td>yes</td>
<td>yesb</td>
<td>yesb</td>
</tr>
<tr>
<td>GenScan</td>
<td><a href="http://genes.mit.edu/GENSCAN.html">http://genes.mit.edu/GENSCAN.html</a></td>
<td>yes</td>
<td>yesb</td>
<td>no</td>
</tr>
<tr>
<td>GenomeScan</td>
<td><a href="http://genes.mit.edu.genomescan/">http://genes.mit.edu.genomescan/</a></td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>MHHgene</td>
<td><a href="http://www.cbs.dtu.dk/services/HMMgene/">http://www.cbs.dtu.dk/services/HMMgene/</a></td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>FGENSES</td>
<td><a href="http://genomic.sanger.ac.uk/glf/gf.shtml">http://genomic.sanger.ac.uk/glf/gf.shtml</a></td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Genie</td>
<td><a href="http://www.cse.ucsc.edu/~dikulc/cgi-bin.genie">http://www.cse.ucsc.edu/~dikulc/cgi-bin.genie</a></td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

aFree to all under GNU licence.
bFree only to academic users. Commercial users must purchase a licence.
cWeb interface to Genie uses an older version of this program, not that one referred to in publications of fruit fly and human genomes or the GASV.

Table 2: Applicability of gene prediction software by organism

<table>
<thead>
<tr>
<th>Program</th>
<th>Parameters available for</th>
<th>May train and test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geneid</td>
<td>Human and fruit fly (Drosophila melanogaster)</td>
<td>Yes</td>
</tr>
<tr>
<td>GlimmerM</td>
<td>Plasmodium falciparum (the malaria parasite), Arabidopsis thaliana, Oryza sativa (rice) and Aspergillus</td>
<td>Yes</td>
</tr>
<tr>
<td>GenScan</td>
<td>General vertebrate parameter set, Arabidopsis thaliana, maize</td>
<td>No</td>
</tr>
<tr>
<td>GenomeScan</td>
<td>General vertebrate parameter set, Arabidopsis thaliana, maize</td>
<td>No</td>
</tr>
<tr>
<td>HMMGene</td>
<td>Humans and Caenorhabditis elegans</td>
<td>No</td>
</tr>
<tr>
<td>FGENSES</td>
<td>Humans, fruit fly (D. melanogaster), nematode, yeast and plant</td>
<td>No</td>
</tr>
<tr>
<td>Genie</td>
<td>Humans, fruit fly (D. melanogaster)</td>
<td>No</td>
</tr>
</tbody>
</table>

aThe general vertebrate parameter set is derived from the training and testing set of 570 sequences described in Burset and Guigo. It is not limited to human sequences.
bThere is no explanation on the web site for what exact species of nematodes, yeast and plants the parameter sets are offered.

cWeb interface to Genie uses an older version of this program, not that one referred to in publications of fruit fly and human genomes or the GASV.

dTest results not available

eSame as (d), except tested on the Adh region in Drosophila.

Table 3: Comparative accuracy of gene prediction on nucleotide level

<table>
<thead>
<tr>
<th>Program</th>
<th>Sn1</th>
<th>Sp1</th>
<th>Sna</th>
<th>Spn</th>
<th>Sn2</th>
<th>Sp2</th>
<th>Snd</th>
<th>Spd</th>
<th>Sn3</th>
<th>Sp3</th>
<th>Snd</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMGene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geneid</td>
<td>97</td>
<td>91</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>GlimmerM</td>
<td>86</td>
<td>83</td>
<td>85</td>
<td>92</td>
<td>94</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genie</td>
<td>96</td>
<td>92</td>
<td>91</td>
<td>90</td>
<td>78</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGENSES</td>
<td>89</td>
<td>77</td>
<td>86</td>
<td>88</td>
<td>77</td>
<td>85</td>
<td>92</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenomeScan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test results not available</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GlimmerM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test results not available</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aTested on the Adh region in Drosophila; data from Pavlovic et al. and Reese et al.
bTested on a set of 195 high-quality mammalian sequences (human, mouse and rat), which has been experimentally validated and includes both multiple and single gene sequences.
cTested on a set of 570 single-gene vertebrate sequences; data from Burset and Guigo.
dGeneid tested by the author in-house (previously unpublished) on a set of 570 sequences.
eOther performance figures collected by the author from the web sites for respective programs.

Taken together these observations lead nucleotides are identified correctly by these methods, the precise gene structure is not, and as a result the predicted peptide may be inaccurate. In addition none of the present gene finders can explicitly account for the possible alternative splicing of the mRNA, opening a possibility for further degradation of the quality of the predicted protein product. However, the situation is not entirely hopeless, as it has been observed that the so-called ‘suboptimal’ exons, as predicted by GenScan, GenomeScan, HMMGene and geneid, may be indicative of the alternately spliced variants of a gene (V. Makarov, unpublished, and Baxevanis). We also note that when measured on the same set of data, performance figures for different programs are often similar, varying by a few percentage points. On the other hand both sensitivity and specificity may vary between the test sets. Consistency of prediction accuracy, or the lack of it, indicates the extreme importance of careful training and testing procedures and accumulation of adequate statistics. Also, it has been observed that combinations of different gene prediction methods may perform substantially better than any of the programs used by itself, at least in case of one test set. A number of computational methods for automatically combining multiple independent gene predictions into a single consensus prediction have been described.
to the following strategy. When choosing a gene-finding program one should consider not only the raw accuracy estimates but also the consistency of the program’s performance on various test sets, as well as the availability of parameters for the organism studied. Whenever possible multiple methods should be used, and results of these multiple predictions should be examined for consistency. In addition to generation of the optimal gene structure one may attempt to predict a number of alternative exon assemblies, or possibly a list of unassembled exons that should include suboptimal exons. The candidate exons should be compared with available experimental data by means of alignment to protein or cDNA sequences that have been obtained independently of the prediction. Be aware that some protein sequences contained in public data banks are predictions themselves and therefore do not satisfy this requirement. Unfortunately this approach involves substantial degree of expert intervention and cannot be completely automated at present. It is to be hoped that advances in data management and standardisation would enable it in the near future.

Acknowledgement
The author is grateful to Dr D. Gilbert (Indiana University, Bloomington) for many inspiring discussions and helpful suggestions regarding this paper.

Disclaimer
The views presented in this manuscript are those of the author and do not necessarily represent the views of BionomiX, LLC.

References