BIOINFORMATICS

Vol. 15 no. 11 1999
Pages 887–899

Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana *sequences*

Nathalie Pavy[1], *Stephane Rombauts*[2], *Patrice Déhais*[2], *Catherine Mathé*[2], *Davuluri V. V. Ramana*[2,3], *Philippe Leroy*[1,4] *and Pierre Rouzé*[1,*]

[1] *Laboratoire associé de l'INRA(France) and* [2] *Department of Plant Genetics, Flanders Interuniversity Institute of Biotechnology, Universiteit Gent, K.L. Ledeganckstraat, 35, B-9000 Ghent, Belgium*

Abstract

Motivation: *The annotation of the* Arabidopsis thaliana *genome remains a problem in terms of time and quality. To improve the annotation process, we want to choose the most appropriate tools to use inside a computer-assisted annotation platform. We therefore need evaluation of prediction programs with Arabidopsis sequences containing multiple genes.*

Results: *We have developed AraSet, a data set of contigs of validated genes, enabling the evaluation of multi-gene models for the Arabidopsis genome. Besides conventional metrics to evaluate gene prediction at the site and the exon levels, new measures were introduced for the prediction at the protein sequence level as well as for the evaluation of gene models. This evaluation method is of general interest and could apply to any new gene prediction software and to any eukaryotic genome. The GeneMark.hmm program appears to be the most accurate software at all three levels for the Arabidopsis genomic sequences. Gene modeling could be further improved by combination of prediction software.*

Availability: *The AraSet sequence set, the Perl programs and complementary results and notes are available at http://sphinx.rug.ac.be:8080/biocomp/napav/.*

Contact: *Pierre.Rouze@gengenp.rug.ac.be*

Introduction

For the genomic sequence annotations, several prediction programs are used to deduce a gene model and similarity searches help to assign protein functions (Rouzé *et al.*, 1999). Although the annotation step has been considered essential by some authors to gain benefit from functional genomics, two main criticisms have already risen up: its deceiving quality and the time required to achieve it. Consequently, at this stage of genomic sequencing, several authors have already pointed out the need for reviewing the annotation process and for curing databases (Korning *et al.*, 1996; Smith, 1998).

Some discrepancies in annotations performed by different teams are documented (Galperin and Koonin, 1998; Brenner, 1999; Terryn *et al.*, 1999). Reannotation became therefore a preoccupation common to communities of biologists whatever the organism. The manual annotation of a 400 kb contig in the *Arabidopsis thaliana* genome was performed by our team and several errors in gene models were outlined as they are built automatically by prediction programs (Terryn *et al.*, 1999). Moreover, as stated by Bork and Koonin (1998), there is a need for a 'widely accepted, robust and continuously updated suite of sequence analysis methods integrated into a coherent and efficient prediction system' to perform computer-assisted annotation of genomic sequences. The issue is consistency, updating, speed and cost of annotation. Such platforms with various ambitions have recently been developed (Harris, 1997; Bailey *et al.*, 1998; Kleffe *et al.*, 1998; Médigue *et al.*, 1999). Our aim is to adapt to the *Arabidopsis* genome a task launching platform with software chaining capacity. We want to classify the prediction tools to choose which individual programs to use to build their chaining into annotation scenarios.

Although many prediction tools are available for *Arabidopsis* sequence analysis and annotation, they have not been evaluated altogether. For vertebrate sequences, several comparative studies of the power of gene prediction programs were reported (Fickett and Tung, 1992; Snyder and Stormo, 1995; Burset and Guigó, 1996; Claverie, 1997; Guigó, 1997; Burge and Karlin, 1998). The accu-

[*] To whom correspondence should be addressed.

[3] On leave from Avesthagen Graine Technologies, Plant Genome Biology Laboratory, P.O. Box 5091, Cubbon Park GPO, Bangalore-560001, India. Present address: CSHL, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

[4] Present address: Station INRA d'Amélioration des Plantes - Domaine de Crouelle 234 avenue du Brézet, 63039 Clermont-Ferrand cedex 2, France.

racy of gene prediction programs can be studied at three levels: site, exon and gene model. Until now, prediction programs have been evaluated with sequence test sets consisting of single genes (Lopez *et al.*, 1994; Burset and Guigó, 1996). Small sets of genes in their actual genomic context, or annotated contigs have nevertheless recently been used (Jiang and Jacob, 1998). Indeed, *in silico* gene prediction has to deal with DNA sequences with genes occurring on both strands and with intergenic regions, and the ability of the programs to build multi-gene models has to be evaluated. We describe here an improved strategy, which allows a precise evaluation of the gene models and of the deduced proteins. This has been made possible through the building of an original data set consisting of authentic small contigs of validated genes. This allowed us to evaluate the performance of gene prediction software for gene modeling and gene product finding, besides the sites and exons. Database mining uses mainly information on the protein encoded by genes and we thus devise a way to evaluate the gene prediction programs also in that respect. The metrics of Burset and Guigó (1996) have been used and extended to give measurements for the additional evaluations.

Because of genome style variations, the prediction software have to be developed for, or adapted to, a given (group of) species. Therefore, the validation of these programs must be performed for each species. Although implemented here for the *Arabidopsis* genome, the evaluation tool that is presented is of general value. Indeed, it could apply to any new gene prediction software and to any eukaryotic genome, provided a suitable sequence test set has been built.

Methods

The sequence data set

The sequences and their annotation were retrieved from GenBank. We searched the annotated *Arabidopsis* BAC sequences by eyes for the occurrence of two or more well documented genes in a row ending with 240 entries. The proper location of all gene and exon borders was checked and sometimes corrected through sequence homology searches and literature reading. All coding sequences were translated and compared with the protein sequence of at least one close homologue. We discarded all sequences presenting similarities with *Arabidopsis* gene sequences deposited in public databases before January 1997 (about 20%), to exclude sequences that may have been used to train the programs. To evaluate the influence of sequencing errors on prediction results, insertions and deletions of one nucleotide were introduced randomly in the sequence test thanks to the ‘corrupt’ program from the GCG package. Altogether this task took 5 man-months for completion.

In order to test the prediction quality of first and last exons, we have considered 300 nt upstream of the first exon of the first gene and 300 nt downstream of the last exon of the last gene for our analysis purpose. The choice of $L = 300$ nt is a compromise between the need of extra sequence for validation of false positive exon sequence and the risk to encompass a real exon from the next gene. Indeed, according to our data set, only one intergenic sequence out of 94 was shorter than 300 nt (Figure 1). We end up with this choice of 300 nt on each border with a density of about one gene every 3.6 kb, only slightly lower that the actual one in the genome, as observed by our own statistics, i.e. one gene per 4.4 kb. Nevertheless, the sequences in the input files used to run the prediction programs were longer on both sides, in order to avoid border effects in predictions, being 2000 nt before the translation start of the first gene and 2000 nt after the stop codon of the last gene. This size of 2000 nt is close to the average length of the intergenic sequences in our data set (2446 nt). This data set with the 2000 nt-long borders was called AraSet.

Prediction programs

We compared prediction programs that have been specifically developed (or adapted) for annotation of the *Arabidopsis* genome sequence. Several analysis levels were considered depending on the purpose of each of the programs listed in Table 1. Whenever possible, programs were run by using different values of the parameters.

Measures of performance

The output files from the prediction programs were standardized and the computation of the various measures and their comparisons were all done through programs written in Perl. To evaluate prediction accuracy at the nucleotide and at the exon level, the main metrics were as in Burset and Guigó (1996).

Evaluation of site predictions. The percentage of correctly predicted sites, percentage of missing sites and percentage of over-predicted sites was calculated. We also reported the location of the over-predicted sites in relation to the actual gene structure (exon, intron or intergenic sequence). The usual performance measures were used: sensitivity (Sn) and specificity (Sp), as defined by Snyder and Stormo (1995).

Evaluation of exon predictions. The accuracy was measured at both nucleotide and exon levels. At the nucleotide level, sensitivity and specificity are defined as follows, according to the notations of Burset and Guigó (1996):

$$\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \qquad \mathrm{Sp} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$$

where TP stands for true positives, FN for false negatives and FP for false positives.

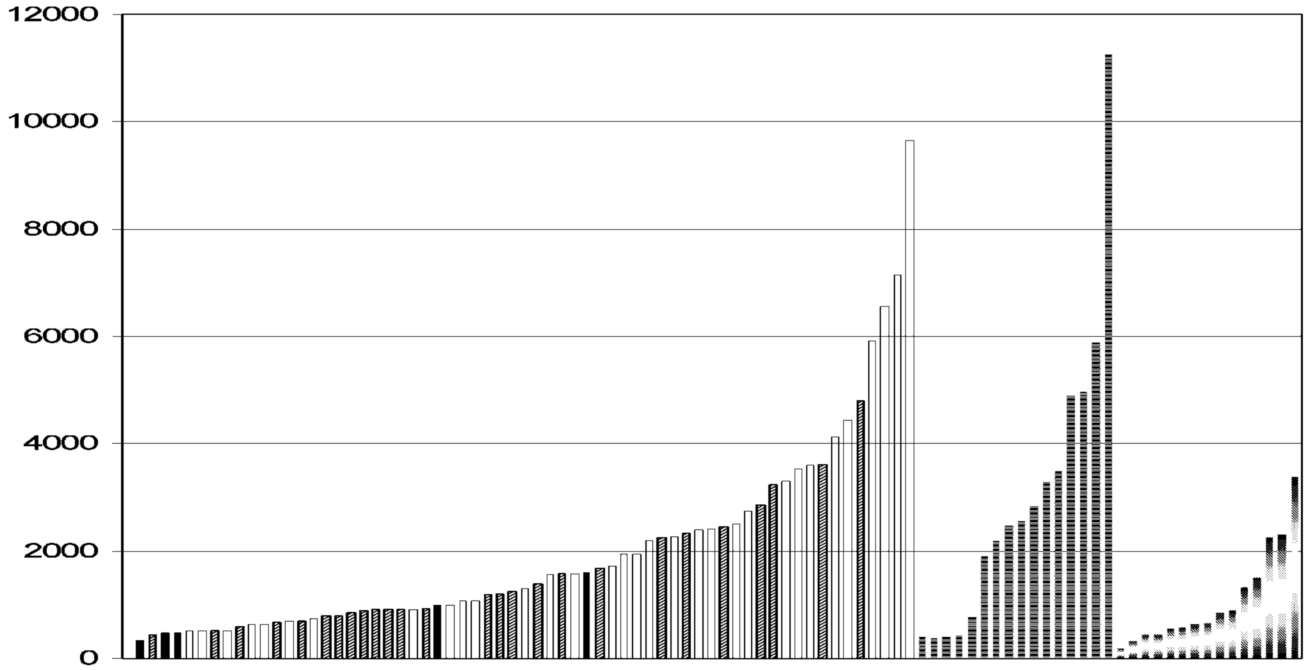

Fig. 1. The intergenic sequences and their predictions by gene modelers.

□ Intergenic sequence composed of one promoter and one terminator – not fused by the gene modelers

■ Intergenic sequence composed of one promoter and one terminator – fused by GeneMark.hmm and GENSCAN

▨ Intergenic sequence composed of one promoter and one terminator – fused by GENSCAN

▤ Intergenic sequence composed of two promoters

☰ Intergenic sequence composed of two terminators

Table 1. The analyzed programs

Program	Web site	Type of prediction and analysis level					Reference
		ATG	Splice sites	Exons	Gene model	Protein or coding region	
NetStart	http://www.cbs.dtu.dk/services/NetStart/	X					Pedersen and Nielsen, 1997
NetPlantGene	http://www.cbs.dtu.dk/NetPlantGene.html		X				Hebsgaard *et al.*, 1996
NetGene2	http://www.cbs.dtu.dk/services/NetGene2/		X				Tolstrup *et al.*, 1997
SplicePredictor	http://gremlin1.zool.iastate.edu/cgi-bin/sp.cgi/		X				Brendel and Kleffe, 1998
SPL			X				
FEX			X	X		X	Solovyev *et al.*, 1994, Sobvyev and Salamov, 1997
FGENE	http://genomic.sanger.ac.uk/gf/gf.html		X	X	X	X	
FGENESP		X	X	X	X	X	
MZEF	http://sciclio.cshl.org/genefinder/		X	X			Zhang, 1998
GRAIL v1.3	http://compbio.ornl.gov/Grail-1.3/		X	X			Xu and Uberbacher, 1997
GeneMark	http://genemark.biology.gatech.edu/GeneMark/hum.cgi					X	Borodovsky and McIninch, 1993
GeneMark.hmm		X	X	X	X	X	Lukashin and Borodovsky, 1998
GENSCAN	http://CCR-081.mit.edu/GENSCAN.html	X	X	X	X	X	Burge and Karlin, 1997

To evaluate gene modeling, we distinguished correct exons, overlapping exons, missing exons and wrong exons (Figure 2). We defined ce, oe, me and we as the number of exons in each category, ae the number of actual exons and pe the total number of predicted exons. At the exon level, we computed the specificity Spe, sensitivity Sne, and ME

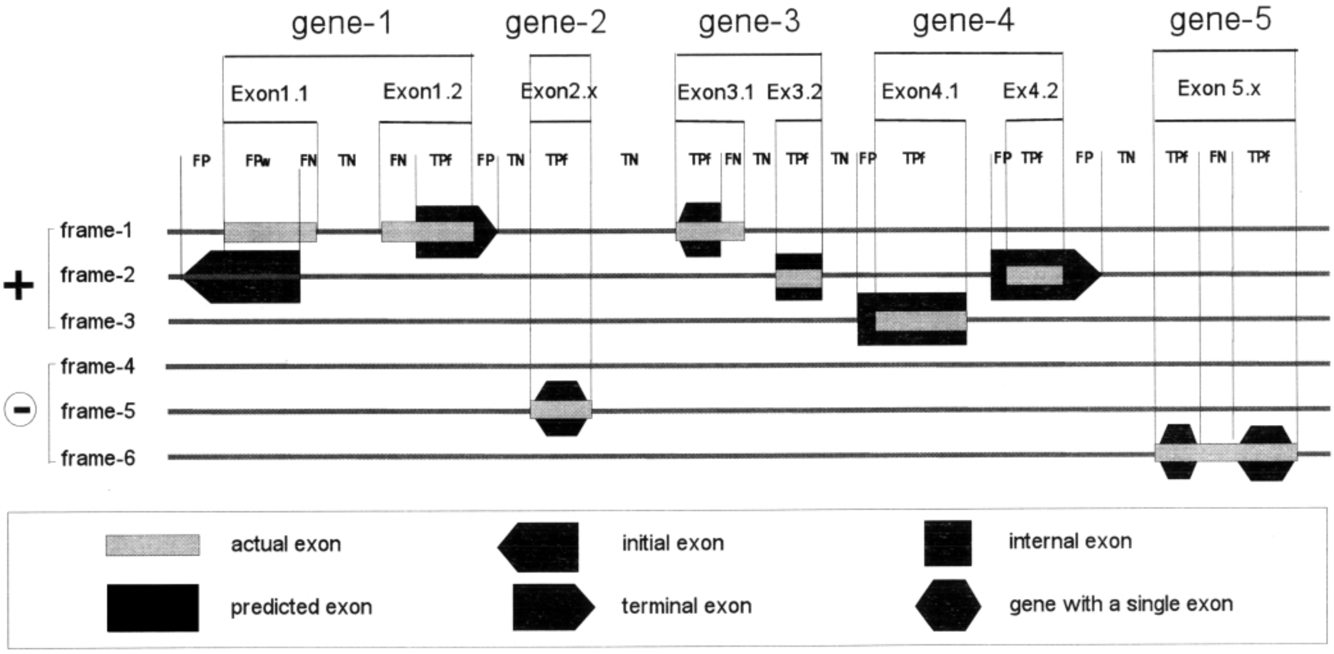

Fig. 2. Theoretical example showing a contig composed of five actual genes and predicted as containing three genes. In this case, the exon 1.1 is missing(mef), exons 1.2, 3.1, 4.1, 4.2 and the unique exon of gene 5 are predicted as overlapping exons (oef), the single exon of gene 2 and exon 3.2 are correct (cef). The gene models are faulty for genes 1, 3 and 4 and correct for gene 2. The actual genes 3 and 4 are fused. The actual gene 5 is split.

and WE, where

$$\mathrm{Sne} = \mathrm{ce}/\mathrm{ae} \quad \mathrm{Spe} = \mathrm{ce}/\mathrm{pe} \quad \mathrm{ME} = \mathrm{me}/\mathrm{ae} \quad \mathrm{WE} = \mathrm{we}/\mathrm{pe}$$

Evaluation of the prediction of protein sequences. For a biologist, the significance of a prediction takes into account the appropriate location of the predicted coding regions and also their reading frame. In order to evaluate the quality of the programs to predict the frame of each exon, we introduced the following measures (Figure 2):

TPf: the number of nucleotides predicted as exonic that are correctly located inside the limits of a coding exon AND in the proper frame;

FPw: number of nucleotides predicted as exonic that are located inside the limits of a coding exon, BUT in a wrong frame;

FPf = FP + FPw: total of false positives according to the frame.

At the exon level, we end up with the cef (exon with the proper borders AND in the correct frame), oef (predicted exon overlapping an actual coding exon in the proper coding frame), mef (an actual exon with either no overlapping predicted exon OR with an exon predicted in a wrong frame) and wef (an exon predicted outside the limits of an actual coding exon OR inside BUT in a wrong frame). These definitions allow the calculation of sensitivity and specificity at the nucleotide level (Snf, Spf) and at the exon level (Snef, Spef) as well as the exons which are really missing or wrong, MEf and WEf, using the same formulas as above.

In practice also, the SIZE of PROPER PROTEIN SEQUENCE is a major criterion of quality of gene prediction, since this will directly influence the efficiency of database searching and the chance to retrieve experimentally the expressed gene (e.g. by a Polymerase Chain Reaction). We thus computed the longest correctly predicted protein sequence (lgs) by assembling the consecutive predicted exons, as long as their content and borders were correct

$$\mathrm{lgs} = \mathrm{oel}_i + \mathrm{ce}_{i+1} + \cdots + \mathrm{ce}_n + \mathrm{oer}_{n+1}$$

where oel is the overlapping fraction of the left-most predicted exon overlapping an actual coding exon in the proper coding frame with only the right (3′) border being correct; and oer is the overlapping fraction of the right-most predicted exon overlapping an actual coding exon in the proper coding frame with only the left (5′) border being correct.

Table 2. Prediction of the splice sites

	Donor sites (860 actual sites)					Acceptor sites (860 actual sites)				
Program	Sn	Sp	% of overpredicted donors in			Sn	Sp	% of overpredicted acceptors in		
			Exons	Introns	Intergenic sequences			Exons	Introns	Intergenic sequences
NetPlantGene	**0.91**	**0.33**	11.64	6.48	81.89	**0.89**	**0.19**	10.88	10.04	79.08
NetGene2 all sites	**0.95**	**0.31**	12.45	7.17	80.38	**0.85**	**0.40**	10.52	7.31	82.17
NetGene2 score $\geq$ 0.90	**0.91**	**0.47**	8.25	6.10	85.65	**0.67**	**0.59**	8.68	7.20	84.12
NetGene2 score $\geq$ 0.95	**0.81**	**0.57**	9.00	4.60	86.40	**0.49**	**0.71**	8.52	5.11	86.36
NetGene2 score $\geq$ 0.98	**0.61**	**0.65**	9.03	5.56	85.42	**0.23**	**0.77**	8.77	7.02	84.21
NetGene2 score=1	**0.55**	**0.69**	7.44	6.05	86.51	**0.22**	**0.76**	8.77	7.02	84.21
NetGene2H-all scores	**0.51**	**0.61**	5.99	4.23	89.79	**0.22**	**0.76**	9.84	6.56	83.61
SPL	**0.84**	**0.30**	13.91	10.37	75.72	**0.76**	**0.23**	13.30	11.89	74.81
SplicePredictor 100% learning set	**0.96**	**0.07**	18.55	11.19	70.26	**1.00**	**0.04**	19.11	10.55	70.33
SplicePredictor/tau maximal/star-value 14	**0.40**	**0.68**	10.00	6.88	83.13	**0.41**	**0.68**	8.43	10.24	81.33
SplicePredictor/tau maximal/star-value 11	**0.63**	**0.60**	11.91	6.65	81.44	**0.57**	**0.54**	14.08	10.80	75.12
SplicePredictor/tau maximal/star-value 8	**0.73**	**0.47**	12.29	8.43	79.29	**0.63**	**0.44**	14.57	11.26	74.17
SplicePredictor/tau maximal/star-value 5	**0.83**	**0.35**	14.16	8.48	77.37	**0.68**	**0.36**	16.06	10.33	73.62
GRAIL	**0.59**	**0.46**	16.18	47.85	35.97	**0.69**	**0.55**	28.03	25.52	46.44
MZEF ($p = 0.03$)	**0.63**	**0.69**	25.10	10.12	64.78	**0.60**	**0.66**	17.34	16.24	66.42
FEX	**0.75**	**0.58**	17.83	7.22	74.95	**0.69**	**0.41**	6.72	8.23	85.05
FGENE	**0.75**	**0.63**	27.01	8.56	64.44	**0.70**	**0.61**	21.65	10.31	68.04
GENSCAN	**0.77**	**0.82**	47.89	16.20	35.92	**0.73**	**0.78**	29.48	25.43	45.09
GeneMark.hmm	**0.93**	**0.81**	37.50	19.57	42.93	**0.90**	**0.84**	38.03	23.94	38.03
FGENESP	**0.58**	**0.72**	41.45	15.54	43.01	**0.55**	**0.70**	28.93	23.86	47.21

Evaluation of multi-gene modeling. We parsed the predicted genes as either entirely correct, faulty (predicted genes sharing some part of an actual gene, i.e. having at least one 'oef' or 'cef' exon), missing (actual genes for which NONE of the exons are predicted, i.e. all of them are in the 'mef' category) or wrong (over-predicted genes for which NONE of the exons are actual exons, even partially, i.e. all of them are in the 'wef' category). From this parsing, gene modeling sensitivity and specificity were calculated. Among the faulty genes the ones that are fused (predicted genes sharing exons – 'cef' or 'oef' – from consecutive actual genes) or split (cases where exon(s) – 'cef' or 'oef' – from a single actual gene are predicted in separate consecutive genes) were identified.

Results and discussion

The gene contig collection

In AraSet, each sequence in the initial set was carefully checked based on the alignment with the corresponding *Arabidopsis* coding sequences or with homologues from other species, and several were removed for lack of evidence or major errors. The eventuality of alternative splicing has been considered, but, as of now there is no cDNA data suggesting occurrence of alternative splicing in AraSet. This gene collection as well as documentation is available on the Internet at the URL http://sphinx.rug.ac.be:8080/biocomp/napav.

Statistics on the data set

The AraSet multi-gene data set contains 566 014 nucleotides divided in 74 loci. It is composed of 57, 14 and 3 sequences containing respectively two, three and four genes in a row. These loci are coming from different regions of the *Arabidopsis* genome: 14 from chromosome I, 24 from ch-II, two from ch-III, nine from ch-IV and 25 from ch-V, reflecting the progress of *Arabidopsis* genome sequencing when they were retrieved. There are 168 genes in AraSet, with a total of 1028 exons and 860 introns, and 94 entire intergenic sequences. The average size of the genes, exons and introns calculated with all data are respectively: 2010 nt/gene (from ATG translation initiation to stop), 197 nt/exon and 154 nt/intron. Intergenic sequences and genes represent about 40 and 60% of the AraSet sequences, respectively, and introns and exons represent respectively 40 and 60% of the gene sequences;

about half of the genome would be intergenic sequence, one-third coding sequence and one-sixth introns. This fits with the more recent statistics on the *Arabidopsis* genome (Sato *et al.*, 1999; Terryn *et al.*, 1999) indicating that our data set is not biased.

Length of the intergenic sequences

In AraSet, we observe a length polymorphism of the intergenic sequences (Figure 1). Surprisingly some intergenic sequences are very short. Among the 94 intergenic sequences, 12 are shorter than 500 nucleotides. The shortest intergenic sequence is only 179 nt long and is composed of two terminator regions. About two-thirds of the genes occurred next to each other on the same strand, the intergenic sequence containing a promoter and a terminator region (63 data). The remaining third of the consecutive gene pairs occurred in opposite orientation, the intergenic sequences containing either two promoter regions (16 data) or two terminator regions (15 data). It would be interesting to see if this bias towards co-oriented genes will be confirmed for the whole genome.

We performed regular database searches to check if a new gene could be found inside intergenic sequences and only found one suspicious case, recently. We are aware that the intergenic sequences, especially the longest ones, may contain orphan genes, which would escape database similarity searches. Consequently, the gene prediction results in sequences annotated as intergenic should be assessed, even if parsed as wrong.

Evaluation of splice site predictions

The results of validation of splice site predictions (Table 2, Figure 3) showed that GeneMark.hmm was the program having both the highest sensitivity and specificity. GENSCAN was almost as specific as GeneMark.hmm, but was less sensitive. Among the splice predictors, the programs, Netgene2, NetPlantGene and SplicePredictor were as sensitive as GeneMark.hmm, the most successful splice prediction software being NetGene2. But these programs were much less specific than GeneMark.hmm.

We analyzed the splice prediction program results with different probability or score threshold to guide their usage and to possibly use them in combination. SplicePredictor and NetGene2 were analyzed by filtering the predicted sites according to the computed star-value or confidence score, respectively (Figure 3). We are aware that the score given by one program and the value by another have not the same meaning, our analysis of the sensitivity and specificity according to this criterion being only an attempt to guide the user choice. With NetGene2 at low cutoff stringency, the sensitivity and specificity of the donor predictions are respectively equal to 0.95 and 0.31. When increasing the score, the specificity hardly increased whereas the sensitivity decreased deeply. Using Netgene2 at high cutoff stringency (‘H’ sites) did not offer a significant increase in specificity while decreasing sensitivity greatly (Table 2).

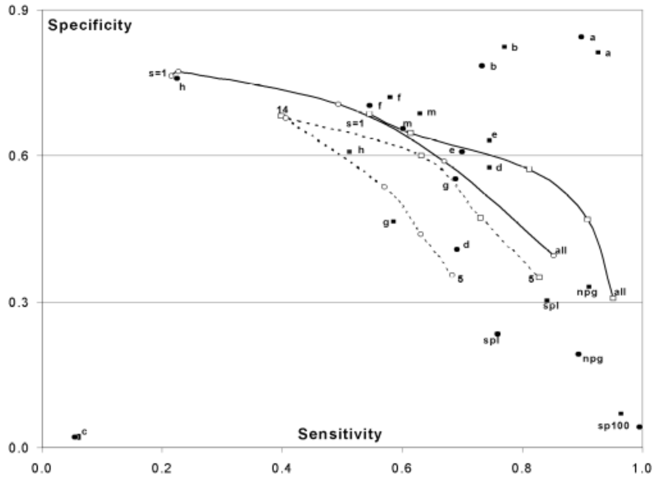

Fig. 3. Evaluation of splice site prediction. –□–, NetGene2 datapoints with scores: $s = 1$, ⩾0.98, ⩾0.95, ⩾0.90, all; - - □ - -, SplicePredictor (with parameter tau maximal) star-value 14, star-value 11, star-value 8, star-value 5; —○—, NetGene2 datapoints with scores: $s = 1$, ⩾0.98, ⩾0.95, ⩾0.90, all; – – ○ – –, SplicePredictor (with parameter tau maximal) star-value 14, star-value 11, star-value 8, star-value 5. With NetGene2, the score is the confidence score at low cutoff (90% sensitivity), except for the point h (prediction at high cutoff ‘H’). With SplicePredictor, the filtering of the prediction results was performed according to the star-value. For NetGene2 and SplicePredictor, the number besides each point indicated the score chosen as threshold. ■, Donors; ● acceptors. a, GeneMark.hmm; b, GENSCAN; c, GeneMark/CUall; f, FGENESP; m, MZEF ($p = 0.03$); e, FGENE; d, FEX; g, GRAIL; npg, NetPlantGene; sp100, SplicePredictor (with the parameter 100% learning set); spl, SPL.

GENSCAN, GeneMark.hmm, FGENESP, MZEF and GRAIL had almost the same quality for acceptor and donor sites. On the contrary, both SplicePredictor and NetGene2 were more sensitive and more specific for donors than for acceptors, especially at low star-values and confidence scores.

Location of the over-predicted splice sites. With all the programs except GRAIL, the over-predicted donor sites are mostly located in the intergenic sequences (Table 2). This is especially true for the three splice site prediction programs. Since intergenic sequences represent 40% of the data set, the difference observed in the locations of the over-predicted sites are not due to an over-representation of the intergenic sequences in the data set, but to a strong propensity of the programs to wrongly detect splice sites in intergenic sequences. For all the programs except

GeneMark.hmm, the distribution of the over-predicted sites differs significantly from a uniform distribution (for the donor sites, the difference is significant with $\alpha = 0.001$ for NetPlantGene, SPL, SplicePredictor, NetGene2, GRAIL, MZEF and FGENE, and $\alpha = 0.1$ with GENSCAN and FGENESP). GRAIL is unique in making more false predictions in introns than in intergenic regions.

As for the donor sites, the distribution observed for the location of the acceptor sites, which are over-predicted by GeneMark.hmm, does not differ significantly from the expected distribution according to the size of each genome category. But, contrary to the results observed for the donors, the difference is not significant either for the distribution of the acceptors over-predicted by GENSCAN, FGENESP and GRAIL. For all the other programs, the over-predicted acceptors are preferentially located in intergenic sequences, the difference between the observed and the theoretical distribution being significant (threshold $\alpha = 0.001$).

The software specifically developed for splice sites prediction were the most sensitive to the over-prediction problem. Interestingly, almost all their over-predicted sites were located in intergenic sequences, indicating that the distinction intergenic/genic sequences would allow a considerable improvement of the specificity of these programs. Especially, constraining NetGene2 to the relevant genic sequences makes it the most specific splice site prediction program.

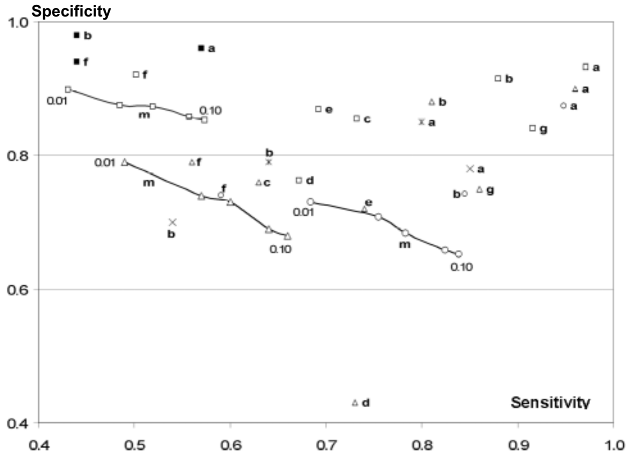

Fig. 4. Evaluation of the exon prediction at the nucleotide level. a, GeneMark.hmm; b, GENSCAN; c, GeneMark/Cuall; f, FGENESP; m, MZEF; e, FGENE; d, FEX; g, GRAIL. □, All the predicted exons; ×; initial exons; ○, internal exons; ∗; terminal exons, △, all the exons predicted in the correct frame; ■, single exons. ———, MZEF with prior probability = 0.01, 0.03, 0.04, 0.08, 0.10.

Evaluation of exon prediction

Analysis at the nucleotide level. The sensitivity and specificity of each exon prediction software were plotted on Figure 4, showing the high accuracy of GeneMark.hmm and of GENSCAN. For all the programs, sensitivity and specificity decreased when only the exon fraction in the correct frame is considered.

Effect of sequencing errors. We tested the effect of sequencing errors on gene prediction accuracy. Insertions or deletions are expected to have the strongest effect on prediction. As a whole, sequencing errors affect more the sensitivity of prediction than the specificity (data not shown but available on our Web site). A low rate of sequence error (10^{-4}) was having only a marginal effect on GeneMark.hmm and GENSCAN prediction sensitivity that was decreased by about 1%. With a sequence error rate of 10^{-3}, the decrease of sensitivity is significant for both programs, with a stronger effect on GENSCAN (−10.2%) compared with GeneMark.hmm (−5.8%). The current genomic sequencing projects fall in this 10^{-4}–10^{-3} range of sequencing errors. The quality of sequencing will thus have a direct effect on the quality of gene finding and annotation. At a high error rate (10^{-2}) which may happen in single-pass sequencing, the decrease of sensitivity is drastic and much more pronounced for GENSCAN (−44%) than for GeneMark.hmm (−29.9%).

Analysis according to the exon type. Sensitivity and specificity were calculated by parsing the exons according to their location in the genes (initial/internal/terminal) and by considering the prediction for each type of exons separately. Indeed, MZEF is adapted to predict internal exons and the gene modelers allow distinguishing all types of exons (Figure 4). When internal exons were specifically analyzed and results compared with the ones for all exons, the sensitivity increased with MZEF and FGENESP, but decreased with GeneMark.hmm and GENSCAN, with a marked loss of sensitivity and specificity for the last one. This is due to wrong gene modeling: as many initial exons are predicted as internal, the specificity for the internal exon prediction is lower than for all exons whatever their type. The decrease in specificity observed with MZEF, reflects the fact that this program predicts exons that overlap actual initial or terminal exons and genes composed of a single exon. When it was used at medium stringency (default, $p = 0.03$), 11.6% of the exons predicted by MZEF overlapped exons which were not internal. When specific prediction of initial exons was

Table 3. Results of the predictions at the exon structure level

	Frame-independent validation										Frame-dependent validation					
Program	Predicted exons	ce correct exons	oe overlapping exons	we wrong exons	me missing exons	Sensitivity Sne	Specificity Spe	Ratio WE	Split exons	Fused exons	Sensitivity Snef	Specificity Spef	Ratio Wef	cef correct exons	oef overlapping exons	wef
GENSCAN	**938**	652	204	82	175	**0.63**	**0.70**	0.09	10	16	**0.63**	**0.69**	0.12	649	182	110
GeneMark.hmm	**1104**	845	172	87	26	**0.82**	**0.77**	0.08	10	4	**0.82**	**0.76**	0.10	844	144	110
MZEF prior $p = 0.01$	**641**	401	153	87	480	**0.39**	**0.63**	0.14	11	10	**0.37**	**0.60**	0.21	382	126	134
MZEF prior $p = 0.04$	**846**	459	236	151	358	**0.45**	**0.54**	0.18	32	14	**0.43**	**0.52**	0.27	438	178	231
MZEF prior $p = 0.10$	**998**	490	298	210	283	**0.48**	**0.49**	0.21	50	16	**0.45**	**0.47**	0.32	467	210	322
FGENE	**1061**	569	300	192	213	**0.55**	**0.54**	0.18	56	6	**0.55**	**0.53**	0.28	562	197	299
GRAIL	**1184**	449	506	229	80	**0.44**	**0.38**	0.19	12	16	**0.43**	**0.38**	0.25	444	440	293
FEX	**1745**	562	484	699	155	**0.55**	**0.32**	0.40	180	23	**0.53**	**0.31**	0.57	547	208	993
FGENESP	**737**	433	195	109	403	**0.42**	**0.59**	0.15	7	8	**0.41**	**0.57**	0.21	423	156	156

considered, GeneMark.hmm kept sensitivity and specificity (Sni = 0.85, Spi = 0.78), whereas GENSCAN was loosing accuracy, especially in sensitivity (Sni = 0.54, Spi = 0.70). For FGENESP, sensitivity for prediction of initial exons is very low (Sni = 0.24, Spi = 0.87). Similar effects were observed for terminal exons, with an even lower figure for FGENESP (Snt = 0.10, Spt = 0.92) (Figure 4).

Evaluation of exon structure modeling

At the exon level, GeneMark.hmm and GENSCAN were the most accurate, with sensitivities being respectively 0.82 and 0.63 (Table 3). With these modelers, the ratio of wrong exons (less than 9%) was low compared with software for exon prediction. GENSCAN predicted fewer exons than GeneMark.hmm. Exons were more frequently fused by GENSCAN than by GeneMark.hmm and much more exons were missed by GENSCAN compared with GeneMark.hmm.

Evaluation of the gene models

Completely correct multi-gene models. Gene models are completely correct only if the borders of all the exons are correct and if the exon types are well defined. Whereas the gene modelers are very successful at the exon level, the accuracy at the gene model level is low. Among the 168 actual genes of the data set, 67, 28 and 10 models are completely correct with GeneMark.hmm, GENSCAN and FGENESP, respectively (Table 4). Nine genes correctly modeled by GENSCAN were faulty with GeneMark.hmm (data not shown but available on our Web site) and 48 genes were correctly predicted by GeneMark.hmm and not correctly by GENSCAN. Five genes were correctly modeled by FGENESP only. As could be anticipated, the performance of the gene modelers varied inversely to the number of exons in the genes. Surprisingly, although similar in sequence, the duplicated genes were not necessarily predicted in the same way.

Wrong genes or new candidate genes? The number of over-predicted genes was quite high for GENSCAN and GeneMark.hmm, which detect 13 and 27 completely wrong genes respectively. All the wrong genes predicted by GeneMark.hmm were re-analyzed with the *blastx* program. Only one of them consisting of two predicted exons gave a hit, showing similarity with a recently cloned well-documented gene, *WUSCHEL*, but the similarity results did not enable us to validate unambiguously the structure of this potential gene which may be a pseudo-gene. We nevertheless cannot exclude that among the ‘wrong’ genes some may be actual genes waiting experimental confirmation.

Prediction of gene borders. GeneMark.hmm and GENSCAN predicted correctly 76 and 50% of the actual initiation sites, respectively (significant difference, $\varepsilon = 8$) (Table 5). NetStart predicts more true ATG sites than GENSCAN and GeneMark.hmm, but with a very poor specificity. In our hands, the filtering of the prediction results according to their score did not help improving the accuracy of that program.

Merging and splitting of genes. In AraSet, 63 intergenic sequences separate two neighbor genes on the same strand. Among these sequences, six (9%) were predicted as introns by GeneMark.hmm and 31 (49%) were seen as introns by GENSCAN (Figure 1). Five of the six genes merged by GeneMark.hmm were also merged by GENSCAN. Whereas GENSCAN merged genes more often than GeneMark.hmm, GeneMark.hmm split genes more often. Whereas only one actual gene was split by GENSCAN, GeneMark.hmm split 18 genes. In these cases, GeneMark.hmm failed to find a splice site but detected instead an initiation site in the vicinity. This fits with the high level of over-prediction of translation initiation sites by GeneMark.hmm. In these cases we observed that GeneMark.hmm was often building incoherent models lacking terminal exon.

Analysis at the protein level

Evaluation of the prediction of protein coding regions. When the frame is taken into account, GeneMark.hmm and GENSCAN did not detect 40 and 197 of the 1028 actual exons, respectively (Table 3). GeneMark.hmm was able to predict 82% of the actual exons in the correct frame with the correct borders, rising up to 96% if overlapping exons in the correct frame were included. Including overlapping exons, GRAIL and GENSCAN predicted respectively 86 and 80.8% of the actual exons in the correct frame, but GRAIL had a higher level of over-predicted exons (28.5%) than GENSCAN (10.7%).

The comparison of the predictions obtained with GeneMark using three different matrices built from codon usage classes (Mathé *et al.*, 1999) shows that better prediction results are obtained with the matrix built with CU_1, a class clustering mostly genes with low expression. The use of gene classification to improve exon prediction was detailed elsewhere (Mathé *et al.* in press).

Fraction of correctly predicted protein sequence. We calculated the total length of correctly predicted protein and the ratio [TPf/size of the actual protein coding sequence] which was obtained from GENSCAN and GeneMark.hmm predictions (Figure 5). For the 168 genes in AraSet, this ratio is higher than 30% with GENSCAN (except for three genes including the completely missing gene) and higher than 50% with GeneMark.hmm

Table 4. Evaluation of the gene models

Gene modeler	Actual genes	Predicted genes	Correct gene model	Missing gene	Partial gene models	Wrong genes	Split genes	Fused genes	Sensitivity	Specificity
GENSCAN	168	150	28	1	139	13	1	60	0.17	0.19
GeneMark.hmm	168	208	67	1	100	27	18	12	0.40	0.32
FGENESP	168	92	10	47	111	3	0	60	0.06	0.11

Table 5. Prediction of the gene borders

Program	Translation initiation sites			Terminator sites	
	Predicted	Correct	Not correctly predicted but in phase with the real ATG site	Predicted	Correct
GENSCAN	127	85	15	125	89
GeneMark.hmm	196	129	31	173	136
FGENESP	68	45	5	44	24
Netstart – all predicted sites	4572	137			
Netstart – probability > 0.9	14	7			
Netstart – probability > 0.8	255	36			
Netstart – probability > 0.7	1007	80			
Netstart – probability > 0.6	2435	112			

(except the missing gene). This ratio is equal to 100% for 89 and 44 proteins deduced from GeneMark.hmm and GENSCAN predictions, respectively. GENSCAN is predicting as a whole a lesser fraction of the actual encoded proteins, but it is able to better find some specific genes. Our results show the interest to use both programs, one complementing the other.

Size of the longest stretch of correct protein sequence. We calculated the longest stretch of correct protein sequence to evaluate and compare the capacity of gene finding software to end up with large contiguous sequence of the actual protein. Figure 6 shows the size of the longest correct protein sequence for each gene as predicted by GENSCAN and GeneMark.hmm. Again, as a consequence of better gene modeling, GeneMark.hmm often ended up with longer correct protein stretches than GENSCAN. GeneMark.hmm predictions ended up with useful protein stretches of significantly bigger size, often double, compared with GENSCAN.

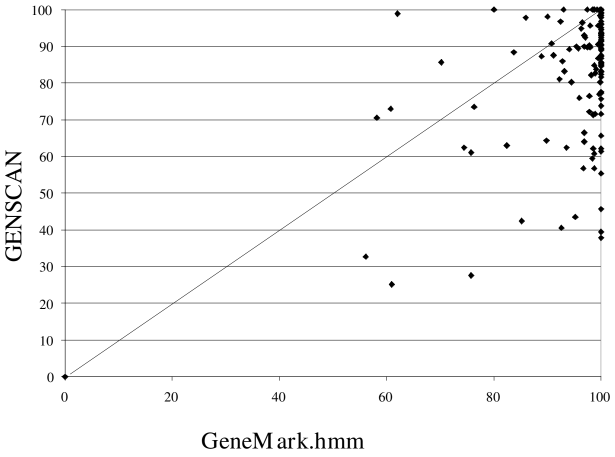

Fig. 5. Portion of the protein sequence, which is correctly predicted by GENSCAN and GeneMark.hmm. Each point represents one of the 168 proteins. The value is the total number of true positive nucleotides, which are predicted, in the correct frame, divided by the length of the actual coding sequence.

Combination of exon prediction software allows validation of exon prediction

We tested the combination of predictions two by two, counting the number of exons with exact same prediction for both programs and parse these common exons into correct, overlap or wrong exons, by comparison with actual exons as done for individual programs (data not shown). The combination of several exon predictions had interesting features. First, wrong predictions were considerably lowered: very few common exons corresponded to over-predictions. For the 11 evaluated combinations involving GeneMark.hmm, GENSCAN, MZEF, GRAIL

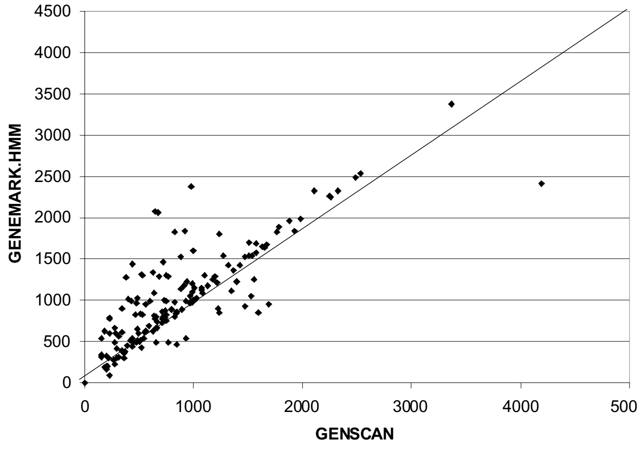

Fig. 6. Comparison of the longest correct coding sequences predicted by GENSCAN and GeneMark.hmm.

and FGENEP, the rate of wrong predictions WE ranged from 1 to 4.2% whereas WE ranged from 7.9 to 20.9% for the individual software. The most useful combinations were the ones involving GeneMark.hmm, GENSCAN and MZEF. Second, combinations of two programs allowed cross-validation of their predictions. On AraSet, 1104 exons are predicted by GeneMark.hmm, among which 845 are correctly predicted. It would be important for the user to be able to identify which exons among the 1104 are most likely the correct ones. The combination of GENSCAN and GeneMark.hmm allowed detection of 626 exons from which 582 are correct, i.e. 57% of the actual exons (Spe = 0.93) (Figure 7). The combination of MZEF and GeneMark.hmm allowed detection of 456 exons from which 431 are correct (Spe = 0.95). The ternary combination led to the consensus prediction of 350 exons, i.e. about one-third of the number of actual exons in AraSet, out of which only nine were not completely correct (Spe = 0.97), three of these being wrong (WE = 0.0086).

Combination of exon prediction programs to improve gene modeling

Some false models from GeneMark.hmm can be easily sorted out from correct models since in these cases either the initial exon or the terminal exon is lacking in the prediction output. These faulty models are easy to detect and can then be improved. Eighteen genes among the 100 faulty models of 168 genes in the data set are concerned. The detailed examination of these abnormal models showed that in 10 cases, one splice site was not predicted and an initiation site was predicted instead; nevertheless the two borders of the actual gene were correctly predicted. In seven cases, one splice site was missing and an initiation site was predicted instead; only one border of the actual gene was correctly found. In one case, a single-exon gene was predicted instead of an internal exon. Among these 18 faulty initial exons, an alternative mode was suggested by using a consensus prediction with GENSCAN, MZEF or GRAIL. Among these 14 exons, 12 appeared to be correct.

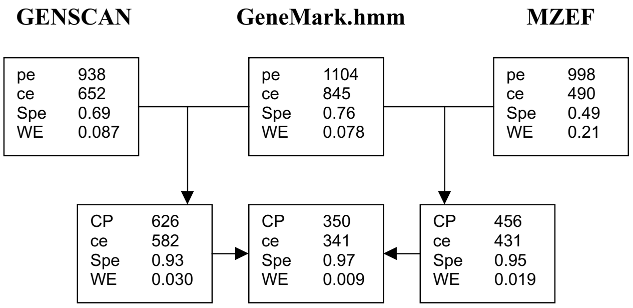

Fig. 7. Combination of the programs GENSCAN, GeneMark.hmm and MZEF (prior probability $p = 0.10$). pe, Predicted exon; ce, number of correct exons; Spe, specificity; WE, ratio of wrong exons; CP, common predictions.

Validation of splice site prediction by combining the results of individual software

With the same reasoning, we determined if a consensus prediction between a splice site prediction program and gene modeler would increase prediction accuracy and help in validating the prediction of the acceptors and donors. We searched for the acceptors (donors) commonly predicted by four chosen combinations, NetGene2 and GeneMark.hmm, GeneMark.hmm and SplicePredictor, GENSCAN and NetGene2, GENSCAN and SplicePredictor and we compared this list with the actual data (Table 6). The increase in specificity is striking, especially for the combination between GeneMark.hmm and NetGene2, which is also the one which kept the highest sensitivity.

Conclusion

Our results in AraSet suggest GeneMark.hmm as the most accurate exon prediction software and gene modeler at all analysis levels for the *Arabidopsis* genome. This encourages its usage for sequence annotation and gene mining of this genome where it is hardly in use yet. Nevertheless, multi-gene modeling remains of limited efficiency since even using GeneMark.hmm, the models built are more often wrong than correct. Clearly, the bottleneck is in finding the proper gene boundaries. Better predictors of these boundaries and better methods to find them are both needed.

Table 6. Results of consensus prediction of the splice sites

	Donors		Acceptors	
	Sn	Sp	Sn	Sp
GeneMark.hmm	0.93	0.81	0.90	0.84
GENSCAN	0.77	0.82	0.73	0.78
GRAIL	0.59	0.46	0.69	0.55
FGENESP	0.58	0.72	0.55	0.70
NetGene2	0.95	0.31	0.85	0.40
SplicePredictor	0.83	0.35	0.68	0.36
GeneMark.hmm and NetGene2	0.90	0.94	0.77	0.94
GeneMark.hmm and SplicePredictor	0.79	0.94	0.63	0.94
GENSCAN and NetGene2	0.76	0.93	0.65	0.92
GENSCAN and SplicePredictor	0.68	0.92	0.53	0.91
GRAIL and NetGene2	0.57	0.84	0.62	0.86
GRAIL and SplicePredictor	0.52	0.88	0.50	0.85
FGENESP and NetGene2	0.57	0.89	0.48	0.89
FGENESP and SplicePredictor	0.52	0.88	0.40	0.85

Having a validation tool did allow us to test and choose the proper combinations of gene predictions. We tested the simplest way to combine the programs (consensus prediction between two programs). Other combination methods were reported (Murakami and Tagaki, 1998). Besides our preliminary results, these methods should be further explored to develop improved strategies for genome annotation using combined prediction results.

The validation method we developed here allows a more realistic evaluation of gene prediction, taking into account the requirements of expert and occasional users for continuous tools adaptation and upgrading for effective genome annotation and gene mining. Our method takes into account the real genomic situation of several genes on both strand on one hand, and stresses the view that the deduced protein sequence is, for the time being, the most informative object of the genome. A special effort has thus been paid to build a real genomic data set of high quality. This approach has the drawback of leaving some uncertainty in the data set, especially on the non-coding regions of genomes which, being defined negatively, could turn out to bear unpredicted function later on. Nevertheless, waiting for more documented data sets to validate *in silico* predictions would be waiting for the time where these predictions will no longer be useful.

Acknowledgements

The authors thank A.Konopka, T.Schiex and anonymous reviewers for their comments on the manuscript. This work was partially supported by a grant from the 'Programme Prioritaire Génomes et Fonction' from the Institut National de la Recherche Agronomique (INRA), France. D.R. was supported by Avesthagen Graine Technologies, India. C.M. is indebted to the Commission of European Communities (BIO4CT950147) for a predoctoral fellowship. N.P., P.L. and P.R. are INRA staff scientists.

References

Bailey,L.C., Fisher,S., Schug,J., Crabtree,J., Gibson,M. and Overton,G.C. (1998) GAIA: framework annotation of genomic sequences. *Genome Res.*, **8**, 234–250.

Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences: where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.

Borodovsky,M. and McIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.

Brendel,V. and Kleffe,J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.*, **26**, 4749–4757.

Brenner,S. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

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

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

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

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

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

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain arrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.*, **1**, 0007.

Guigó,R. (1997) Computational gene identification: an open problem. *Comput. Chem.*, **21**, 215–222.

Harris,N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.*, **7**, 754–762.

Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouzé,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.

Jiang,J. and Jacob,H.J. (1998) EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Res.*, **8**, 268–275.

Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1998) GeneGenerator – a flexible algorithm for gene prediction and its application to maize sequences. *Bioinformatics*, **14**, 232–243.

Korning,P.G., Hebsgaard,S.M., Rouzé,P. and Brunak,S. (1996) Cleaning the GenBank *Arabidopsis thaliana* data set. *Nucleic Acids Res.*, **24**, 316–320.

Lopez,R., Larsen,F. and Prydz,H. (1994) Evaluation of the exon predictions of the GRAIL software. *Genomics*, **24**, 133–136.

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

Mathé,C., Peresetsky,A., Déhais,P., Van Montagu,M. and Rouzé,P. (1999) Classification of *Arabidopsis thaliana* gene sequences:

clustering of coding sequences into two groups according to codon usage improves gene prediction. *J. Mol. Biol.*, **285**, 1977–1991.

Mathé,C., Déhais,P., Pavy,N., Rombauts,S., Van Montagu,M. and Rouzé,P. (1999) Gene prediction and gene classes in *Arabidopsis thaliana*. *J. Biotech.*, in press.

Médigue,C., Rechenmann,F., Danchin,A. and Viari,A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15**, 2–15.

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

Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. 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, pp. 226–233.

Rouzé,P., Pavy,N. and Rombauts,S. (1999) Genome annotation: which tools do we have for it? *Curr. Opin. Plant Biol.*, **2**, 90–95.

Sato,S., Kaneko,T., Kotani,H., Hayashi,R., Liu,Y.G., Shibata,D. and Tabata,S. (1999) A sequence-ready contig map of the top arm of *Arabidopsis thaliana* chromosome 3. *DNA Res.*, in press.

Smith,T.F. (1998) Functional genomics – bioinformatics is ready for the challenge. *Trends Genet.*, **14**, 291–293.

Snyder,E.E. and Stormo,G.D. (1995) Identifying genes in genomic DNA sequences. In Bishop,M.J. and Rawlings,C.J. (eds), *DNA and Protein Sequence Analysis: A Practical Approach* 2nd edn, IRL Press, Oxford, pp. 209–224.

Solovyev,V.V. and Salamov,A. (1997) The Gene-Finder computer tools foranalysis of human and model organisms genome sequences. In and (eds), *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* AAAI Press, Menlo Park, pp. 294–302.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156–5163.

Terryn,N., Heijnen,L., De Keyser,A., Van Asseldonck,M., De-Clercq,R., Verkabel,H., Gielen,J., Zabeau,M., Villarroel,R., Jesse,T., Neyt,P., Hogers,R., Van Den Daele,H., Ardiles,W., Schueller,C., Mayer,K., Déhais,P., Rombauts,S., Van Montagu,M., Rouzé,P. and Vos,P. (1999) Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analysing a 400 kb contig at the *APETALA2* locus on chromosome 4. *FEBS Lett.*, **445**, 237–245.

Tolstrup,N., Rouzé,P. and Brunak,S. (1997) A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res.*, **25**, 3159–3163.

Xu,Y. and Uberbacher,E.C. (1997) Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.*, **4**, 325–338.

Zhang,M.Q. (1998) Identification of protein coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. *Plant Mol. Biol.*, **37**, 803–806.