Plant Molecular Biology (2005) 57:445–460
DOI: 10.1007/s11103-005-0271-1

Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes

Hong Yao[1,4], Ling Guo[1,6,†], Yan Fu[1,4,†], Lisa A. Borsuk[1,6], Tsui-Jung Wen[2], David S. Skibbe[1,5], Xiangqin Cui[1,4,9], Brian E. Scheffler[8], Jun Cao[1,4], Scott J. Emrich[6], Daniel A. Ashlock[3,6] and Patrick S. Schnable[1,2,4–7,*]

[1]*Department of Genetics, Development, and Cell Biology (*author for correspondence; e-mail schnable@iastate.edu);* [2]*Department of Agronomy;* [3]*Department of Mathematics;* [4]*Inderdepartmental Graduate Programs in Genetics;* [5]*Department of Molecular, Cellular and Developmental Biology;* [6]*Department of Electrical and Computer Engineering and Department of Bioinformatics and Computational Biology;* [7]*Center for Plant Genomics, Iowa State University, Ames, Iowa 50011-3650;* [8]*USDA-ARS, Mid South Area Genomics Facility, Stoneville, MS 38776-0038, USA;* [9]*Present address: Department of Biostatistics, Birmingham, AL 35294, USA;* [†]*these authors contributed equally to this work*

Received 16 August 2004, accepted in revised form 6 January 2005

Key words: ab initio gene prediction, *a1-sh2* interval, maize, maize assembled genomic islands, maize genomic survey sequences

Abstract

Five *ab initio* programs (FGENESH, GeneMark.hmm, GENSCAN, GlimmerR and Grail) were evaluated for their accuracy in predicting maize genes. Two of these programs, GeneMark.hmm and GENSCAN had been trained for maize; FGENESH had been trained for monocots (including maize), and the others had been trained for rice or *Arabidopsis*. Initial evaluations were conducted using eight maize genes (*gl8a*, *pdc2*, *pdc3*, *rf2c*, *rf2d*, *rf2e1*, *rth1*, and *rth3*) of which the sequences were not released to the public prior to conducting this evaluation. The significant advantage of this data set for this evaluation is that these genes could not have been included in the training sets of the prediction programs. FGENESH yielded the most accurate and GeneMark.hmm the second most accurate predictions. The five programs were used in conjunction with RT-PCR to identify and establish the structures of two new genes in the *a1-sh2* interval of the maize genome. FGENESH, GeneMark.hmm and GENSCAN were tested on a larger data set consisting of maize assembled genomic islands (MAGIs) that had been aligned to ESTs. FGENESH, GeneMark.hmm and GENSCAN correctly predicted gene models in 773, 625, and 371 MAGIs, respectively, out of the 1353 MAGIs that comprise data set 2.

Abbreviations: AE, actual exon; CC, correlation coefficient; FN, false negative; FP, false positive; GSSs, genome survey sequences; HC, high C_ot; MAGIs, maize assembled genomic islands; ME, missing exon; MF, methylation filtration; OE, overlapped exon; PE, partial exon; RACE, Rapid Amplification of cDNA Ends; SN, sensitivity; SP, specificity; TE, true exon; TP, true positive; WE, wrong exon.

Introduction

Locating the positions of all the genes and determining their structures is a first step toward deciphering the functions of a sequenced genome. Two approaches are available (reviewed by Stormo, 2000; Mathé *et al.*, 2002; Pertea and Salzberg, 2002). The first is based on sequence similarity. A

significant degree of sequence identity or similarity between a genomic query sequence and cDNA, EST, protein or genomic sequences of a gene from the same or another species can provide evidence that a query sequence contains a gene. This method is, however, highly dependent upon the quantity and quality of pre-existing sequence data. Typically only 50–70% of the genes in a sequenced genome can be found via comparisons to other genomes, although this fraction will increase as the number of sequenced genomes increases (reviewed by Mathé *et al.*, 2002; Pertea and Salzberg, 2002). In addition, sequence similarity searches can provide misleading information due to artifacts in databases. The second approach for identifying genes in a sequenced genome is to use *ab initio* gene prediction programs. *Ab initio* gene prediction uses statistical and computational methods to detect coding regions, splice sites, and start and stop codons in genomic sequences. This approach does not depend on sequence similarity and is therefore not limited by the availability of sequence data. But as compared to predictions based on sequence similarity, *ab initio* predictions are currently typically less accurate because available programs are not yet able to make highly reliable predictions of gene structures. One reason for this is that the quality of predictions is limited by the quality of the training sets. These training sets usually consist of gene sequences that have been characterized in a given species.

To date only two plant genomes, *Arabidopsis* (The *Arabidopsis* Genome Initiative, 2000) and rice (Goff *et al.*, 2002; Yu *et al.*, 2002), have been completely sequenced. Efforts to sequence other crop genomes, including maize, are underway (http://www.nsf.gov/bio/pubs/awards/genome02.htm) (Palmer *et al.*, 2003; Whitelaw *et al.*, 2003). The maize genome consists of about 2400 Mb, i.e., approximately 6-fold larger than that of rice (reviewed by Moore, 2000). It is estimated that the maize genome contains approximately 50,000 genes that account for only 10–15% of the genome (Bennetzen *et al.*, 2001). Much of the genome is repetitive elements, many of which are retrotransposons (SanMiguel *et al.*, 1996). Due to the large size and highly repetitive nature of the maize genome, sequencing efforts are being focused on the gene-rich, low-copy fraction of the genome, i.e., the 'gene space'. Two methods are being used to isolate the 'gene space', methylation-filtration (MF) (Rabinowicz *et al.*, 1999) and high C_ot (HC) selection (Peterson *et al.*, 2002; Yuan *et al.*, 2003).

The identification of genes from sequences generated from the maize genome sequencing project will establish whether the current sequencing approaches are successfully enriching for genes, and will, in addition, define genomic resources necessary to study the function of maize. Given the limitations associated with gene prediction based on sequence similarity, *ab initio* gene prediction programs will necessarily play an important role in maize gene discovery. In an effort to develop an *ab initio* gene discovery strategy for maize, existing versions of five programs (Table 1) including FGENESH (Salamov and Solovyev, 2000), GeneMark.hmm (Lukashin and Borodovsky, 1998), GENSCAN (version 1.0) (Burge and Karlin, 1997),

Table 1. Evaluated gene prediction programs.

Programs	Websites	Trained organisms	Type of prediction			Algorithm models
			Splice site	Exon	Gene model	
FGENESH	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	Monocots	Yes	Yes	Yes	GHMM[a]
GeneMark.hmm	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi?org=H.sapiens	Maize	Yes	Yes	Yes	GHMM
GENSCAN	http://genes.mit.edu/GENSCAN.html	Maize	Yes	Yes	Yes	GHMM
GlimmerR	http://www.tigr.org/tdb/glimmerm/glmr_form.html	Rice	Yes	Yes	Yes	IMM[b]
Grail	http://compbio.ornl.gov/Grail-1.3/	*Arabidopsis*	Yes	Yes	No	neural networks

[a]GHMM, Generalized Hidden Markov Model.
[b]IMM, Interpolated Markov Model.

GlimmerR (Salzberg *et al.*, 1999; Yuan *et al.*, 2001) and Grail (version 1.3) (Xu and Uberbacher, 1997) were evaluated for their accuracy in predicting maize genes. The purpose of this study was to evaluate currently available tools for suitability in the *ab initio* discovery of genes from partial maize genomic sequences rather than to compare the algorithms that underlie these tools. Maize-trained versions of three of these programs, FGENESH, GeneMark.hmm and GENSCAN, are available. For the remaining programs versions that had been trained using rice (GlimmerR) or *Arabidopsis* (Grail) were used. Each program was evaluated using a data set consisting of genomic sequences of eight genes cloned by us (Table 2). Because these gene sequences could not have been included in the data sets used to train the *ab initio* programs, they represent a valuable tool for evaluating these programs. These five programs were also used to help identify and determine the structures of two genes in the 140-kb maize *a1-sh2* interval (Civardi *et al.*, 1994; Yao *et al.*, 2002). FGENESH, followed by GeneMark.hmm and GENSCAN made more accurate gene predictions in these tests. Their ability to predict maize genes was further tested using a larger data set (1353 genic sequences) consisting of maize assembled genomic islands (MAGIs) assembled from genome survey sequences (GSSs) whose exons were identified via alignments to ESTs. In this larger data set, FGENESH was still the most accurate program.

Materials and methods

Data sets

Two data sets were used to evaluate the accuracy of gene prediction programs.

Data set 1: The genomic sequences of eight maize genes (*gl8a*, *pdc2*, *pdc3*, *rf2c*, *rf2d*, *rf2e1*, *rth1*, and *rth3*) that were cloned in our lab but that had not been released to the public prior to the completion of this evaluation constitute data set 1. With the exception of *rf2d* that is incomplete at its 5′ end, all of these genic sequences contain the corresponding start and stop codons. The GenBank accession numbers of each gene sequence are listed in Table 2. Their gene structures were determined by spliced alignment of full-length cDNA sequences to the corresponding genomic sequences using the GeneSeqer program (http://bioinformatics.iastate.edu/cgi-bin/gs.cgi) (Usuka *et al.*, 2000; Usuka and Brendel, 2000; Brendel *et al.*, 2004). To make a fair comparison of the predictions among genes in this data set, genomic sequences that contain complete genes were trimmed at their 5′ and 3′ ends. The incomplete *rf2d* genomic sequence was only trimmed at the 3′ end. Consequently, in this data set, the amount of sequence before the start codon of each gene is 520 bp and after the stop codon is 375 bp. The statistical characteristics of each gene in data set 1 are listed in Table 2.

Table 2. Members and characteristics of data set 1.

Genes	GenBank accession numbers	(G + C)% of Gene[a]	Input sequence length (bp)	#D[b]	#A[c]	#Exons	Exon length (bp)			Intron length (bp)		
							Min	Max	Average	Min	Max	Average
gl8a	AF302098	50.0	3288	2	2	3	70	653	327	583	829	706
pdc2	AF370004	51.2	3974	5	5	6	118	651	297	82	691	259
pdc3	AF370006	54.1	3477	5	5	6	118	651	304	77	391	152
rf2c	AF348412	56.1	4527	6	6	7	62	648	216	70	1604	354
rf2d[d]	AF348414	54.9	2940	6	7	7	62	474	200	72	123	96
rf2e1	AY374447	54.3	4673	9	9	10	69	237	134	75	1080	271
rth1	AY265854	39.9	13621	24	24	25	65	174	107	80	1705	419
rth3	AY265855	61.5	2899	0	0	1	2004	2004	2004	NA[e]	NA	NA
Overall		48.8	39399	57	58	65	62	2004	208	70	1705	327

[a]Begins with the start codon and ends with the stop codon.
[b]#D, number of donor sites.
[c]#A, number of acceptor sites.
[d]The *rf2d* gene sequence is partial in the 5′ end. The first intron is partial and was not included for analysis of intron length here although the A site of this intron is included for counting the #A.
[e]NA, Not applicable.

Data set 2: Data set 2 consists of a subset of the 114,173 ISU MAGIs in version 3.1b (http://plantgenomics.iastate.edu/maize). These MAGIs were assembled from 879,523 GSSs (MF and HC sequences) of the maize inbred line B73 using a strategy similar to that described by Emrich *et al.* (2004) (see Supplementary Materials). MAGIs to include in data set 2 were selected based on the qualities of their GeneSeqer alignments to clustered B73 ESTs generated by Schnable Lab. Detailed methods used to generate data set 2 are provided in the Supplementary Materials. In summary, data set 2 consists of 1353 selected MAGI contigs that contain at least one pair of reliable donor and acceptor sites flanking an intact intron (Figure 3). The statistical characteristics of data set 2 are shown in Figure 1.

Statistical comparison of data set 1 to 74 structure-known genes of maize

The GC contents and lengths of internal exons and introns in data set 1 were compared to those in a data set consisting of 74 structure-known maize genes. The GC contents and lengths of the exons and introns of the 74 structure-known maize genes were calculated by parsing the exons and introns from the sequences that were downloaded from NCBI. Only complete internal exons and introns were used in this analysis. The length of each internal exon and intron was determined and then the number of G's and C's were counted and divided by the total length to determine the percent GC content. Similar calculations were conducted for sequences in data set 1. The two-sample Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1939), which tests the null hypothesis that the data values from two samples have the same continuous distribution, was used to compare these parameters.

Programs evaluated

Five *ab initio* programs were evaluated using data set 1. The features of these programs are listed in Table 1. Available versions of FGENESH, GeneMark.hmm and GENSCAN (version 1.0) that had been trained for maize were evaluated. For those two programs for which a maize trained version was not available, the version trained for the closest

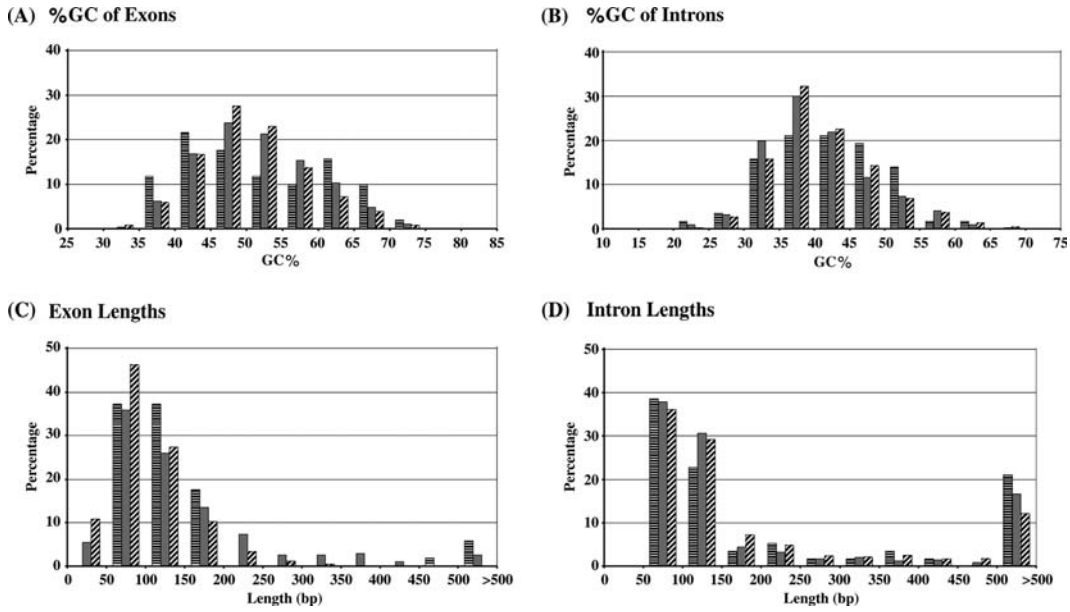

Figure 1. The GC contents and lengths of exons and introns in data sets 1, 2 and structure-known genes. Only internal exons were analyzed in data set 1 and the structure-known genes. The percentages of exons (panel A) and introns (panel B) with the indicated GC contents (bin sizes = 5 percentage points) in each of three data sets are indicated. The percentages of exons (panel C) and introns (panel D) with the indicated lengths (bins sizes = 50 bp) in each of three data sets are also indicated. Data set 1, structure-known genes, and data set 2, are indicated by horizontal stripes, dark gray fill and diagonal stripes respectively.

organism to maize was evaluated. Although all five programs make predictions in both strands of a genomic DNA sequence, only the predictions for the strand containing known genic sequences were analyzed in this study because they could be compared with the known actual gene structures or splice sites in our test data sets 1 and 2. FGENESH, GeneMark.hmm, GENSCAN and GlimmerR predict gene models that can be single or multiple in a genomic sequence and a predicted exon is indicted as initial (starting with the initiation codon and ending with a donor site), internal (starting with an acceptor site and ending with a donor site), terminal (starting with an acceptor site and ending with the stop codon) or single (starting with the initiation codon and ending with the stop codon) exon in the output. Grail predicts a series of non-overlapping exons in both DNA strands but no gene model is produced. All programs were run via their websites by using their organism-specific default parameters to obtain the prediction results for data set 1 (Table 1). To obtain predictions for data set 2, FGENESH, GeneMark.hmm and GEN-SCAN were run locally using the default parameters for monocot sequences (including maize) with usage of the GC donor site (FGENESH) or using the default parameters for maize sequences (GeneMark.hmm and GENSCAN). Additional information is provided in the Supplementary Materials. Programs for prediction of splice site only such as SplicePredictor (Brendel *et al.*, 2004), NetGene2 (Hebsgaard *et al.*, 1996; Tolstrup *et al.*, 1997) and GeneSplicer (Pertea *et al.*, 2001) were not evaluated in this study because they do not predict exons and/or gene models.

Evaluation of gene prediction programs

The performance of each program was evaluated at three levels (splice site, nucleotide and exon) as described by Burset and Guigo (1996) and Pavy *et al.* (1999).

At the splice site level, the accuracy of a program's predictions is measured by SN (sensitivity), SP (specificity) and the average of SN and SP ((SN + SP)/2). If true positive (TP) is defined as the number of correctly predicted splice sites, false positive (FP) as the number of incorrectly predicted splice sites, and false negative (FN) as the number of actual splice sites missed in the prediction, then SN = TP/(TP + FN) and SP = TP/(TP + FP). Since neither SN nor SP alone can represent the accuracy of a program, the value of (SN + SP)/2 is usually used as a measure of accuracy.

SN and SP are also used to evaluate predictions at the nucleotide level. Here TP is the number of nucleotides that are correctly predicted as coding, TN is the number of nucleotides that are correctly predicted as non-coding, FP is the number of nucleotides that are incorrectly predicted as coding, and FN is the number of nucleotides that are incorrectly predicted as non-coding. Under these definitions, SN = TP/(TP + FN), SP = TP/(TP + FP). The value of the correlation coefficient (CC) that reflects both SN and SP is used for evaluation. CC is defined as:

$$CC=((TP\times TN)-(FN\times FP))/((TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN))^{1/2}.$$

At the exon level, if AE is defined as the actual exons, TE as the correctly predicted exons, PE as the predicted exons that are partially correct (i.e., only one boundary correct), OE as the predicted exons that overlap with the actual exons but with both boundaries wrong, ME as the actual exons missed in the prediction, WE as the number of incorrectly predicted exons, then SN = TE/AE, SP = TE/(TE + PE + OE + WE), PE% = 100 × PE/(TE + PE + OE + WE), OE% = 100 × OE/(TE + PE + OE + WE), ME% = 100 × ME/AE, and WE% = 100 × WE/(TE + PE + OE + WE). These values above as well as the average of SN and SP are used to measure the accuracy of a program.

RT-PCR and cDNA library screen to identify the yz1 *gene*

The five *ab initio* programs evaluated in this study predicted a gene, *yz1,* in the maize *a1-sh2* interval. To confirm this gene prediction and to determine the actual structure of this gene, RT-PCR experiments were conducted and maize cDNA libraries were screened. SuperScript™ First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA) was used to obtain first-strand cDNA from total RNA. The sequences of oligonucleotides used as primers in the subsequent PCR are: YZ4b (5′-GAGATGATGTCCCTTGTG-3′) and ZH2587 (5′-GCCTGGTTAGCGAAGTTG-3′). RT-PCR amplification using these two primers revealed a

681-bp fragment in maize RNA isolated from different organs, including husk, tassel, silk, adult leaf, ear and seedling (data not shown). The sequence of this RT-PCR fragment is identical to the predicted *yz1* exons and the predicted introns were missing from the RT-PCR product.

This RT-PCR product was used as a probe to screen maize cDNA libraries. A cross-hybridizing clone with a 2.1-kb insert was identified from a library prepared from seedlings of the inbred CI31A. Sequence analysis of this clone demonstrated that it is chimeric, with only 1.4-kb derived from the *yz1*. Sequence analysis of this cDNA clone, as well as 3′- and 5′- Rapid Amplification of cDNA Ends (RACE) (Invitrogen, Carlsbad, CA) experiments, suggested that this 1.4-kb sequence is full-length or nearly full-length.

Results

Evaluation of gene prediction programs for maize gene discovery

Five *ab initio* gene prediction programs (Table 1) were evaluated for their ability to predict maize genes from genomic sequences. The purpose of this evaluation was to help biologists select a strategy for the *ab initio* discovery of maize genes from partial genomic sequences using currently available tools. Hence, this evaluation did not seek to evaluate the algorithms *per se* upon which the gene prediction tools are based. Of the five evaluated gene prediction programs, Grail predicts splice sites and exons but not gene models; the remaining four programs predict gene models as well as exons and splice sites. Most of the programs had been previously trained using monocots (e.g., maize and/or rice), but Grail was trained using *Arabidopsis*. Evaluations were conducted using a data set (data set 1) consisting of eight maize genomic gene sequences (Table 2) that could not have been included in the data sets used to train any of the five prediction programs because these sequences were released from GenBank only after the evaluation of the gene prediction programs had been completed.

Seven of the gene sequences are full-length; one (*rf2d*) is partial. The GC contents of the genes (from start to stop codons) in data set 1 range from 39.9% to 61.5% with an average of 48.8%. The lengths of these gene sequences range from 2899 to 13,621 bp (Table 2). There are 65 exons in this data set with one to 25 exons per gene. Exon lengths range from 62 to 2004 bp with an average of 208 bp. The average intron length is 327 bp with a minimum of 70 bp and a maximum of 1705 bp. The total numbers of donor and acceptor sites are 57 and 58, respectively. To determine if the genes in data set 1 are representative of maize genes as a whole, we compared several of their features to those of a set of 74 structure-known maize genes downloaded from GenBank (Methods). The Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939) revealed no significant differences (P value >0.2) in the lengths and GC contents of internal exons (i.e., those that begin with an acceptor site and end with a donor site) and introns (Figure 1). Therefore, data set 1 is at least reasonably representative of the maize genes that have been deposited in GenBank to date.

The performance of the five gene prediction programs was evaluated at three levels: splice site, nucleotide and exon (Methods). At the splice site level, the accuracy of a program is measured by the average value of SN (sensitivity) and SP (specificity) since neither SN nor SP alone is sufficient to indicate the ability of a program to predict genes (Methods). FGENESH had the highest values of (SN + SP)/2 (Table 3). These values are 0.91 for donor sites and 0.92 for acceptor sites. GeneMark.hmm was the second most accurate program

Table 3. The accuracy of gene predictions in data set 1 at the splice site level.

Programs	Donor sites			Acceptor sites		
	SN	SP	(SN + SP)/2	SN	SP	(SN + SP)/2
FGENESH	0.91	0.91	0.91	0.91	0.93	0.92
GeneMark.hmm	0.77	0.92	0.84	0.71	0.85	0.78
GENSCAN	0.56	0.91	0.74	0.53	0.86	0.70
GlimmerR	0.61	0.95	0.78	0.59	0.92	0.75
Grail	0.49	0.39	0.44	0.66	0.51	0.58

with values of (SN + SP)/2 = 0.84 for donor sites and 0.78 for acceptor sites. Both the SN and SP of FGENESH's predictions are high. In contrast, GeneMark.hmm exhibits a higher value of SP than SN. Such a differential is also present in the predictions from GENSCAN and GlimmerR which have SPs close to those of FGENESH and GeneMark.hmm but that have much lower SNs. GeneMark.hmm, GENSCAN and GlimmerR predict donor sites better than acceptor sites.

Accuracy at the nucleotide level (Table 4) is measured by the value of the correlation coefficient (CC, Methods). The programs, from the most accurate to the least measured by the CC, are FGENESH (0.93), GeneMark.hmm (0.89), GENSCAN (0.82), GlimmerR (0.71) and Grail (0.43). FGENESH has the highest SN (0.97) and GENSCAN has the highest SP (0.95). The values of SN and SP for FGENESH and GeneMark.hmm are both high (over 0.90). Although the SP values of GENSCAN and GlimmerR are also high (0.95 and 0.91, respectively), their SN values are less favorable (0.81 and 0.70, respectively).

At the exon level, the programs with values of (SN + SP)/2 from the highest to lowest are: FGENESH (0.87), GeneMark.hmm (0.75), GENSCAN (0.68), GlimmerR (0.57) and Grail (0.31). FGENESH has both the highest SN and SP (0.86 and 0.88, respectively). The SPs of GeneMark.hmm and GENSCAN (0.80 and 0.81, respectively) compare favorably with those of FGENESH but their SNs compare less favorably (0.69 and 0.54, respectively). GlimmerR also exhibits better SP than SN. Consistent with its highest SN and SP among the five evaluated programs, FGENESH has the lowest percentage of missing (ME% = 4.6) and wrong (WE% = 3.1) exons. Although the values of WE% in GeneMark.hmm, GENSCAN and GlimmerR predictions are not high (5.4, 7.0 and 7.7, respectively), the values of ME% are 19, 39, and 23, respectively. Grail exhibited the lowest SN and SP and had correspondently high percentages of both MEs (ME% = 17) and WEs (WE% = 31). Predicted exons can have only one correct boundary (PE, partial exon) or can overlap the true exon but lack two correct boundaries (OE, overlapped exon). Of the exons predicted by FGENESH 9.4% and 0% were PEs and OEs, respectively. These are the lowest values of all evaluated programs. Predictions from GeneMark.hmm and GENSCAN also contain no OEs but 14% and 12% PEs, respectively. GlimmerR and Grail predicted both PEs and OEs, but the values of PE% are much higher than that of OE%.

Gene discovery in the a1-sh2 *interval*

To test the ability of the five evaluated programs to discover new maize genes, each was used to predict the structures of genes in the 15,783 bp (GenBank accession no. AF434192) and 6506-bp fragments (GenBank accession no. AF434193) of the 140-kb maize *a1-sh2* interval (Yao *et al*., 2002) (Figure 2). Because the *a1-sh2* sequences were not released to the public until after the completion of this evaluation, these sequences also could not have been included in the training sets of any of the prediction programs.

The Bennetzen lab (Chen and Bennetzen, 1996; Chen *et al*., 1998) sequenced the *a1-sh2* intervals of rice and sorghum (GenBank accession no. U70541 and AF010283, respectively) and predicted a genic sequence between the *a1* and *sh2* loci, which they termed '*Gene X*'. The identification of its maize homologue has been described by Yao *et al*. (2002). Comparison of the genomic and cDNA sequences revealed that the maize *x1* gene contains seven exons (Figure 2). Comparisons of the sequences of both the rice and maize full-length *x1*

Table 4. The accuracy of gene predictions in data set 1 at the nucleotide and exon levels.

Programs	Nucleotide level			Exon level						
	SN	SP	CC	SN	SP	(SN + SP)/2	PE%	OE%	ME%	WE%
FGENESH	0.97	0.94	0.93	0.86	0.88	0.87	9.4	0	4.6	3.1
GeneMark.hmm	0.92	0.93	0.89	0.69	0.80	0.75	14	0	19	5.4
GENSCAN	0.81	0.95	0.82	0.54	0.81	0.68	12	0	39	7.0
GlimmerR	0.70	0.91	0.71	0.51	0.64	0.57	23	5.8	23	7.7
Grail	0.55	0.67	0.43	0.34	0.28	0.31	33	7.7	17	31

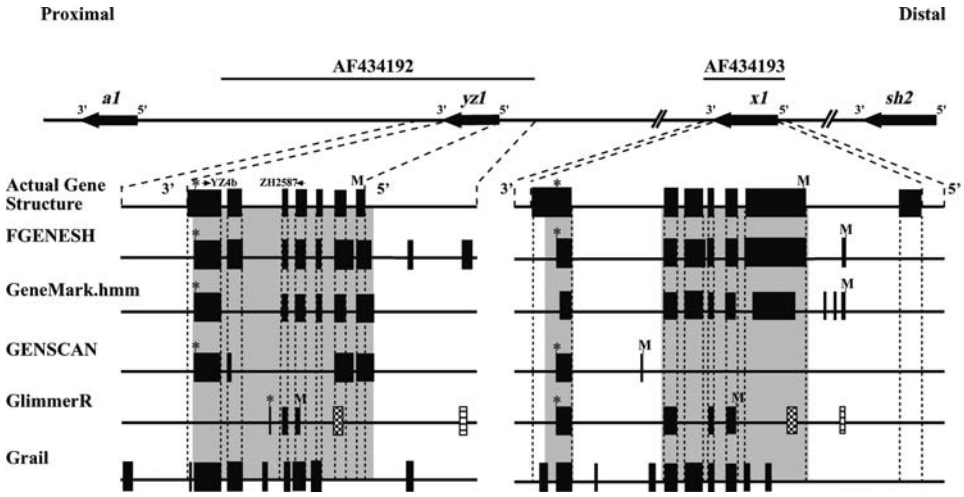

Figure 2. Gene discovery in the *a1-sh2* interval of maize. The gene structures of the *yz1* and *x1* genes predicted by the five indicated programs and their actual structures as verified via RT-PCR and sequencing of cDNA clones are shown. The positions of initiation Mets (M) in the actual genes and predicted gene models are shown. The positions of stop codon are designated by*. Gray regions are conserved among rice, sorghum and maize. In the gene models predicted by GlimmerR, exons filled with different patterns belong to different predicted genes. Primers used in RT-PCR are shown as horizontal arrows.

cDNAs to the predicted rice or sorghum '*Gene X*' showed that only the 5′ portion of the predicted '*Gene X*' corresponds to the actual rice and maize *x1* genes.

Comparisons of *a1-sh2* derived sequences from rice, sorghum and maize revealed a conserved region other than the *a1* and *x1* (Figure 2). This conserved region is located at the distal end of the 15,783-bp portion of the maize *a1-sh2* interval (GenBank accession no. AF434192) and overlaps with the 3′-portion of the predicted '*Gene X*' in rice and sorghum. Since the actual rice and maize *x1* genes do not contain this conserved region and part of this region is single-copy in the maize genome (Yao *et al.*, 2002), we hypothesized that there is another gene in the *a1-sh2* interval. To test this hypothesis, the five gene prediction programs were used to conduct predictions in a 5.4-kb segment from the distal end of the 15,783-bp sequence within the *a1-sh2* interval. All programs predicted a gene in this 5.4-kb sequence (Figure 2), although the predicted gene structures vary. To confirm the validity of these predictions, primers (YZ4b and YZ2587) were designed in the putative exonic regions that were predicted by most programs and that exhibit a high degree of sequence similarity among rice, sorghum and maize (Figure 2). RT-PCR amplification using these primers and comparison of the sequence of the amplified fragment to the 5.4-kb genomic sequence revealed that as predicted by the *ab initio* programs, an additional expressed gene (termed as *yz1*) is present in the *a1-sh2* interval (Methods).

A 1.4-kb *yz1* cDNA was isolated (Methods) that is nearly full-length, probably lacking only the five codons at its 5′ end where a putative initiation Met resides. This putative initiation Met was predicted based on the fact that it and the following four amino acids are conserved among rice, sorghum and maize. Comparison of the 1.4-kb cDNA sequence of *yz1* to the 5.4-kb genomic sequence showed that the genomic sequence of *yz1* is approximately 2.7 kb and consists of seven exons. In a more recently submitted *a1-sh2* sequence from the rice cultivar japonica (GenBank accession no. AF101045), the original '*Gene X*' has been annotated as *x1* and *x2* which are homologs of the maize *x1* and *yz1* genes, respectively.

Comparisons of predicted and actual yz1 *and* x1 *splice sites and gene structures*

Because a complete gene model predicted by FGENESH, GeneMark.hmm, GENSCAN and GlimmerR begins with the start codon and ends with the stop codon, the 5′- and 3′-UTRs of *yz1* and *x1* were not considered in the following comparisons.

FGENESH gave the second best prediction for the *yz1* gene and the best prediction for the *x1* gene at the splice site, nucleotide and exon levels (data not shown). Even so, none of the gene models predicted by FGENESH, GeneMark.hmm, GENSCAN and GlimmerR for *yz1* and *x1* is completely correct (Figure 2). The start and stop codons of *yz1* are located in the first and the last (seventh) exons, respectively. Although FGENESH, GeneMark.hmm and GENSCAN correctly predicted the positions of the stop codons, each of these programs missed the start codon by predicting the first exon as internal rather than initial (i.e., one that starts with the initiation codon and ends with a donor site). The start and stop codons of *x1* are located in the second and last (seventh) exons, respectively. None of the four programs correctly predicted the position of the start codon. Whereas FGENESH, GENSCAN and GlimmerR correctly predicted the location of the stop codon, GeneMark.hmm's prediction is incorrect.

A particular problem with predicting maize genes using the version of GlimmerR that was trained for rice is that it splits maize genes. As shown in Figure 2, GlimmerR predicted multiple genes using the *yz1* and *x1* gene sequences. It predicted three genes in both the *yz1*- and the *x1*-containing sequences: two of these predicted genes each consists of a single exon (which starts with an initiation codon and ends with a stop codon); while in each case the other predicted gene contains multiple exons. GlimmerR was used to predict genes in genomic sequences from the rice *a1-sh2* interval that correspond to the *x1* and *yz1* genes. Although the sensitivity of GlimmerR in predicting *x1* and *yz1* was not improved by using rice sequences, no split genes were predicted (data not shown). Hence, the observed gene splitting could be a consequence of using a version of GlimmerR that had been trained on rice to predict maize genes.

Gene predictions in MAGIs

Gene prediction programs are of particular importance in predicting genes in large genome projects. We therefore extended our evaluations to the maize GSSs being generated as part of the NSF Plant Genome project 0221536 and assembled into MAGIs at Iowa State University (http://plantgenomics.iastate.edu/maize). Data set 2 consists of 1353 MAGI contigs that aligned well with B73 3′ ESTs sequenced by us and that contain at least one pair of reliable donor and acceptor sites flanking an intact intron (Figure 3 and Supplementary Materials). There are 1928 pairs of reliable canonical splice sites and 18 pairs of reliable non-canonical splice sites (16 GC–AG pairs, 2 AT–AC pairs) in this data set that correspond to 592 reliable exons and 1946 reliable introns. Detailed statistical characteristics of data set 2 are provided in Figure 1.

Data set 2 was analyzed with only FGENESH, GeneMark.hmm and GENSCAN because these programs were trained using maize sequences and proved more reliable in the analyses of data set 1 than other two programs. Predictions of data set 2

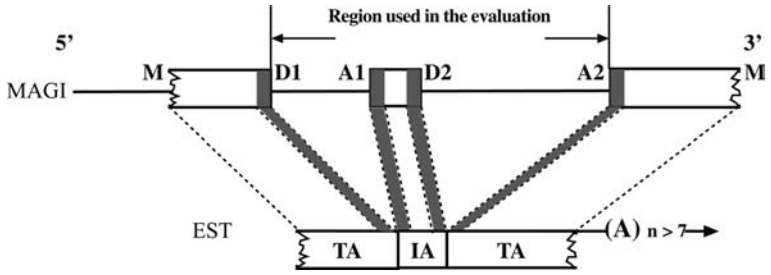

Figure 3. Criteria used to select qualified alignments for data set 2. ESTs that contained polyA tails of at least 8 A's were aligned to MAGI contigs. Alignments between a genomic sequence and an EST can be either terminal or internal. To qualify, terminal alignments (TA) between a MAGI contig and EST must be ≥50 bp with ≥98% nucleotide identity; internal alignments (IA) must be flanked by two qualified terminal alignments and exhibit ≥98% nucleotide identity. In addition, the 10 bp alignments at each splice junctions (shaded regions) must exhibit 100% nucleotide identity. It is possible that the end points of the alignment may not be the real boundaries of exons due to the incompleteness of the MAGI contig (e.g., 3′ end) or the EST (e.g., the 5′ end). These end points were therefore masked and not used to evaluate FGENESH, GeneMark.hmm and GENSCAN. Although not shown in this figure there are MAGI contigs that contain only two qualified terminal alignments and MAGI contigs that contain more than one qualified internal alignments. M's: masked end points of an alignment between a MAGI contig and EST; D's, donor sites; A's, acceptor sites.

from each of the three tested programs were parsed for subsequent analysis. Only regions flanked by two reliable splice sites and without the presence of non-reliable internal splice sites were considered in the evaluation (Figure 3). As was done for data set 1, the accuracy of the predictions by each of the three programs was evaluated at the splice site, nucleotide and exon levels (Methods). As shown in Table 5, the overall accuracy of each the program was somewhat reduced as compared to that obtained using data set 1 (Tables 3 and 4) due to the decreased specificities of the predictions at the nucleotide and exon levels and decreased sensitivities at all three levels. At the splice site level, the specificities of the predictions by the three programs remained high and indeed increased somewhat. Overall, FGENESH performed better than the other two programs because of its much higher sensitivity, even though its specificity is slightly lower than that GeneMark.hmm (Table 5).

At the nucleotide level, FGENESH's values of SN and SP are 0.86 and 0.84, respectively with a CC of 0.83. Consistently, when considering only the MAGIs that have FGENESH predictions at the regions evaluated in our analyses (Figure 3), SN and SP are well correlated at the nucleotide level; this reflects the fact that the majority of these MAGIs have SN and SP values equal to 1 (data not shown). Therefore, if a MAGI is predicted by FGENESH to contain a gene, that prediction is likely to be correct.

When running FGENESH to predict genes in data set 2, the '–GC' parameter was used. This allows FGENESH to predict non-canonical GC donor sites. FGENESH correctly predicted 13/16 (81%) of the non-canonical GC donor sites in data set 2. In contrast, none of these GC donor sites were correctly predicted by GeneMark.hmm and GENSCAN. Neither of the two pairs of non-canonical AT–AC splice sites in the data set 2 was identified by any of the three programs.

FGENESH, GeneMark.hmm and GENSCAN correctly predicted the gene models in 773, 625, and 371 MAGIs, respectively, out of the 1353 MAGIs in data set 2 (Figure 4). FGENESH, GeneMark.hmm and GENSCAN uniquely and correctly predicted 214, 94, and 21 MAGIs, respectively. FGENESH, GeneMark.hmm and GENSCAN failed to predict the evaluated regions as genic in 249, 235, and 540 MAGIs, respectively. FGENESH, GeneMark.hmm and GENSCAN uniquely missed the evaluated genic segments completely in 50, 31, and 275 MAGIs, respectively.

If the predictions from all three programs are considered together, the number of correctly predicted MAGIs increases to 911 and the numbers of MAGIs that were completely missed drops to 112. These results suggest that combining the prediction results from different programs can increase the accuracy of predictions.

Comparisons of predictions of internal vs. initial/terminal exons

In vertebrate and *Drosophila* genomic sequences FGENESH and GENSCAN predict internal exons better than they predict initial and terminal exons (i.e., those that begin with an acceptor site and end with a stop codon) (Burge and Karlin, 1998; Salamov and Solovyev, 2000). This reflects the poorer abilities of these programs to detect the correct start and stop codons than their abilities to correctly identify splice sites. The abilities of the four programs evaluated in this study, FGENESH, GeneMark.hmm, GENSCAN and

Table 5. The accuracy of gene predictions in data set 2.

	Donor sites			Acceptor sites			Nucleotide level			Exon level					
	SN	SP	$\frac{(SN+SP)}{2}$	SN	SP	$\frac{(SN+SP)}{2}$	SN	SP	CC	SN	SP	$\frac{(SN+SP)}{2}$	(PE + OE)%	WE%	ME%
GENSCAN	0.41	0.95	0.68	0.39	0.92	0.66	0.46	0.86	0.60	0.33	0.57	0.45	36	7.6	57
GeneMark.hmm	0.69	0.94	0.82	0.63	0.93	0.78	0.77	0.88	0.80	0.66	0.67	0.66	28	5.9	20
FGENESH	0.73	0.95	0.84	0.71	0.92	0.82	0.86	0.84	0.83	0.74	0.65	0.69	27	8.2	14
FGENESH (score ≥ 0)[a]	0.71	0.95	0.83	0.67	0.94	0.81	0.83	0.86	0.82	0.72	0.68	0.70	27	4.4	17

[a]FGENESH (score ≥0) is a modified evaluation of the FGENESH's prediction, in which predicted exons with negative scores were treated as if there were no predictions and which were therefore not included in the evaluation.

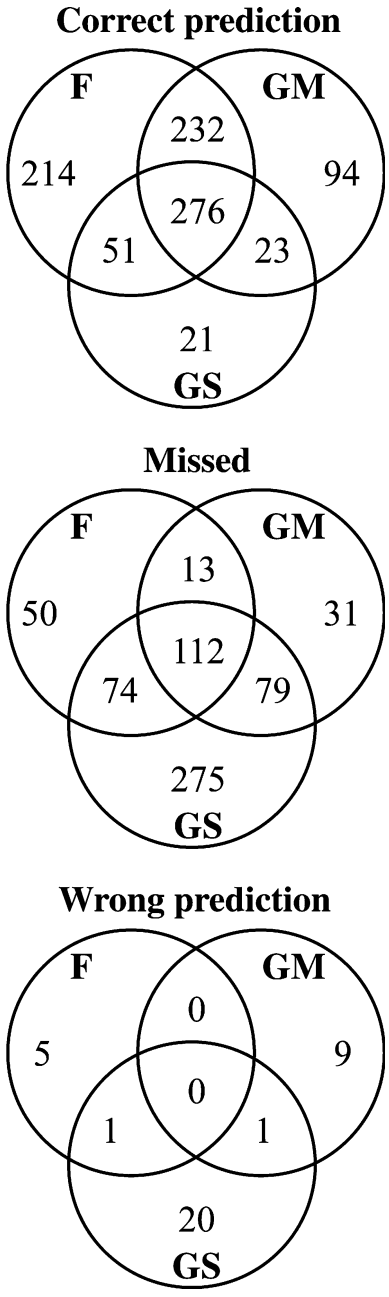

Figure 4. Numbers of MAGIs correctly predicted, missed or predicted completely incorrectly by FGENESH, GeneMark.hmm and GENSCAN. In these comparisons, the entire gene model in the evaluated region of each MAGI (Figure 3) was considered. Predictions that exactly matched the actual gene model were classified as correctly predicted. Predictions that failed to identify the evaluated region as genic were classified as missed. Genic predictions that failed to correctly identify any features of the actual gene model were classified as completely incorrect. F, FGENESH; GM, GeneMark.hmm; GS, GENSCAN.

GlimmerR, to predict initial/terminal exons versus internal exons were compared at the exon level using the eight genes in data set 1 and the *yz1* and *x1* genes from the *a1-sh2* interval. Grail was not included in this analysis because it does not predict exons as initial, internal and terminal.

As expected based on experience from other genomes, FGENESH, GeneMark.hmm and GlimmerR predicted the internal exons better than the initial and terminal exons (Table 6). Surprisingly, GENSCAN predicted initial and terminal exons better than it did internal exons due to its higher SN of the initial and terminal exons in data set 1.

Selecting reliable exon prediction

The accuracy of gene prediction at the exon level, as well as the prediction of gene models is not as high as the accuracy at the nucleotide level. This is because more than 12% of the predicted exons are PEs, OEs and WEs (Tables 4 and 5). In applications such as primer design for RT-PCR experiments or the design of oligos for microarrays, it is highly desirable to be able to exclude WE predictions. Of the two most reliable programs in this evaluation, FGENESH and GeneMark.hmm, only FGENESH reports a confidence score for its exonic predictions. This score is an aggregate of log-odds scores that the base pairs are members of an exon. Unfortunately, we have been unable to locate in FGENESH documentation or its references the method used to estimate the context-sensitive probability that a base is a member of an exon or the method of aggregating these scores into an overall exon score.

To determine if an FGENESH exon score correlates with the quality of prediction, the distributions of exon's scores were compared among the TEs, PEs + OEs and WEs (Figure 5) from predictions using data set 2. About 49% of the WEs have negative scores. In contrast, only 2.2% of the TEs and 6.6% of the PEs and OEs have negative scores. Considering only predicted exons with non-negative scores, 68% are TEs, 27% are PEs + OEs, and only 4.5% are WEs. These results demonstrate that removing exons with negative FGENESH scores eliminates almost half of the WEs, while retaining the majority of the TEs and PEs + OEs. Indeed, in data set 2 all WEs have scores of less than 10 (Figure 5). Hence, if

Table 6. Comparisons of accuracy at the exon level between the prediction of initial/terminal exons and internal exons.

Programs	Initial and terminal exon							Internal exon						
	SN	SP	$\frac{(SN+SP)}{2}$	PE%	OE%	ME%	WE%	SN	SP	$\frac{(SN+SP)}{2}$	PE%	OE%	ME%	WE%
FGENESH	0.77	0.72	0.74	22	0	0	5.6	0.87	0.85	0.86	8.2	0	5.0	6.6
GeneMark.hmm	0.71	0.60	0.65	20	5.0	0	15	0.67	0.80	0.73	14	0	22	6.0
GENSCAN	0.71	0.63	0.67	16	0	12	21	0.42	0.83	0.63	13	0	52	3.3
GlimmerR	0.47	0.44	0.46	17	11	29	28	0.48	0.66	0.57	25	6.8	25	2.3

only those predicted exons with scores greater than 10 were used, WEs could be totally eliminated.

To determine the effect of removing exons with negative scores on the evaluation of FGENESH, the parameters for the evaluation of FGENESH's predictions at the splice site, nucleotide and exon levels were recalculated for data set 2 (Table 5, FGENESH (score ≥ 0)). In this analysis, predicted exons with negative scores were treated as they had not been predicted. This modified analysis resulted in higher SPs at all three levels as compared to the SPs obtained in the original analysis of FGENESH. In contrast, SNs decreased. The values of (SN + SP)/2 for splice site and exon levels, CC, and ME% and WE% were altered only slightly.

Discussion

FGENESH performed better than other evaluated programs for maize gene discovery

The goal of this study was to identify a strategy based on existing *ab initio* gene prediction tools that biologists can use to discover maize genes in genomic sequences. Accordingly the performances

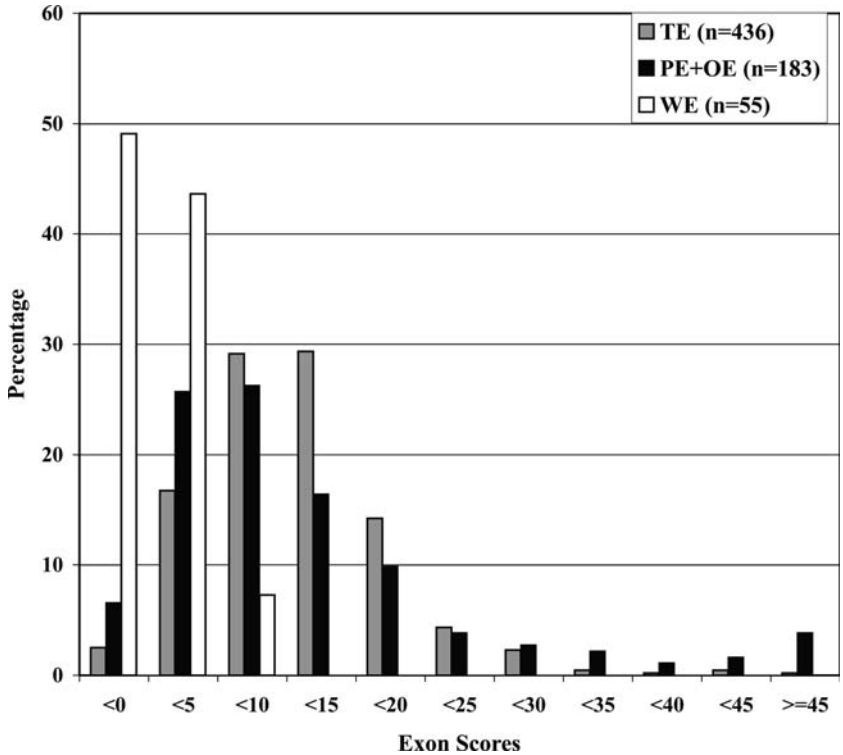

Figure 5. The distributions of exon scores among the TE (true exon), PE + OE (partial and overlapped exon) and WE (wrong exon) predicted by FGENESH using data set 2.

of five *ab initio* gene prediction programs were evaluated using data set 1, which consists of eight maize genes that could not have been used to train these programs. These eight genes are structurally similar to a larger set of 74 structure-known genes downloaded from GenBank (Figure 1). Hence, evaluation of predictions performed on the eight genes in data set 1 are likely to be informative of the ability of these programs to predict other maize genes.

In these evaluations FGENESH performed the best at all three levels of evaluation. FGENESH has also been demonstrated to be more accurate than other tested programs for the discovery of rice and mammalian genes (Solovyev, 2001; Yu *et al*., 2002). GeneMark.hmm is the most accurate program for *Arabidopsis* gene discovery when evaluated using the AraSet that contains contigs of validated genes (Pavy *et al*., 1999). In our evaluation, GeneMark.hmm was the second most accurate program. GENSCAN, which is very good at predicting mammalian genes (Rogic *et al*., 2001), fared less well in our analysis of maize genes. This may be due to the fact that GENSCAN was trained on a smaller data set than FGENESH and GeneMark.hmm (Supplementary Materials). One reason for the poor performance of GlimmerR and Grail in predicting the maize genes may be because they had been trained for other plants (including rice and *Arabidopsis*, Table 1). Since gene features differ among organisms, it is likely that the parameters of these programs have not been optimized for maize gene discovery. Additional evaluations of FGENESH, GeneMark.hmm and GENSCAN using data set 2, which consists of 1353 genic MAGIs, also demonstrated that FGENESH is most accurate at predicting maize genes. It is, however, important to emphasize that this study was not designed to evaluate the algorithms used by these programs. On the other hand, this study does reveal which existing programs will provide maize biologists with the best gene predictions.

Predictions of small exons may be less accurate

The accuracies of FGENESH, GeneMark.hmm and GENSCAN predictions in data set 2 are not as high as those in data set 1. This may be due to the increased fraction of small exons (e.g., ≤100 bp) in data set 2 as compared to data set 1 (Figure 1). This enrichment for small exons in data set 2 is probably a consequence of our stringent EST-guided strategy to select reliable genic regions in MAGIs for analyses (Figure 3, Methods and Supplementary Materials). It has been demonstrated that in rice genes FGENESH is not as successful at predicting small exons (less than 200 bp) as large exons (Yu *et al*., 2002). The finding that in data set 2 the fractions of MEs in exons that are smaller than 50 bp is significantly higher than the corresponding fraction among larger (> 50 bp) exons (χ^2 test, P value = 0.002) is consistent with the hypothesis that this enrichment for small exons is at least partly responsible for the reduced accuracy of predictions in data set 2 as compared to those of data set 1.

Gene model prediction programs need improvement

As shown in Table 7, none of the four gene model prediction programs (FGENESH, GeneMark.hmm, GENSCAN and GlimmerR) precisely predicted the structures of more than half of the eight genes in data set 1 plus the two genes from the *a1-sh2*

Table 7. Comparisons of gene model predictions.

Programs	*gl8a*	*pdc2*	*pdc3*	*rf2c*	*rf2d*	*rf2e1*	*rth1*	*rth3*	*x1*	*yz1*	Number (%) of correct models
FGENESH	Y[a]	Y	N	Y	Y	N	N	Y	N	N	5 (50%)
GeneMark.hmm	Y	N	Y	Y	N	N	N	Y	N	N	4 (40%)
GENSCAN	Y	Y	Y	Y	Y	N	N	N	N	N	5 (50%)
GlimmerR	N[b]	N	Y	N	N	N	N	N	N	N	1 (9%)
Number of programs that predicted correct models	3	2	3	3	2	0	0	2	0	0	

[a]Y, Prediction of the gene model is correct.
[b]N, prediction of gene model is incorrect.

interval. The structures of four genes (*rf2e1*, *x1*, *rth1*, and *yz1*) were not predicted correctly by any of the four programs. These programs each appear to have difficulties predicting gene models that include non-canonical splice sites, start codons that are not in the first exon, or large numbers of small exons and/or large introns. For example, the inability of these programs to correctly predict non-canonical splice sites appears to be the reason they failed to predict correctly the gene model of *rf2e1*. The *rf2e1* gene contains two pairs of non-canonical splice sites, GC/AG and CC/AA, in introns 2 and 3, respectively. Each of the four programs missed both of these two non-canonical donor and acceptor sites.

The start codon of the *x1* gene is in its second exon (Figure 2), which may interfere with the ability of prediction programs to identify it. The reason for the incorrect prediction of the start codon in the *yz1* gene is not clear. The *rth1* gene has 25 exons, each of which is less than the average length of maize exons (i.e., ~200 bp, Table 2). Thirteen of the *rth1* exons are between 50 and 100 bp, eight are between 100 and 150 bp and the remaining four are between 150 and 200 bp in length. In contrast, eight of the *rth1* introns are larger than the average maize intron (i.e., ~300 bp, Table 2) and four are over 900 bp. All of the four assayed programs missed some of *rth1*'s small exons and incorrectly predicted the presence of exons within *rth1*'s large introns. Three of the programs (GENSCAN, GeneMark.hmm and GlimmerR) even split the *rth1* gene, which may indicate a poor ability to predict large genes. As pointed out by Wang *et al*. (2003) because *ab initio* programs predict genes based on statistical analyses of all possible genic features (e.g., splice sites, start and stop codons), longer sequences have an increased probability on containing false genic features that exhibit statistical significance. In addition, stop codons are more likely to be associated with FP predictions in intron, which could split large genes (which usually contain large introns). Our study provides additional evidence that GlimmerR's predictions tend to incorrectly split maize genes. GlimmerR split five genes (*gl8a*, *rf2e1*, *rth1*, *yz1*, and *x1*). Since GlimmerR was trained for rice, the current version may not suitable for the prediction of maize genes. We conclude that *ab initio* gene model prediction remains a field that would benefit from further research.

The ability of FGENESH to predict non-canonical splice sites

Non-canonical splice sites can make the accurate prediction of gene models difficult because until recently no program was trained to recognize non-canonical splice sites due to an insufficient number of non-canonical sites in the training sets. As more genomic sequences have become available, data sets of EST-supported canonical and non-canonical mammalian splice sites have been created and analyzed (Burset *et al*., 2000, 2001). In these mammalian splice site data sets, the canonical GT–AG pairs account for 98.7% of all splice site pairs; non-canonical GC–AG pairs and AT–AC pairs account for 0.56% and 0.05%, respectively, and all other non-canonical pairs account for 0.02%. The collection of GC–AG pairs in this mammalian data set was large enough for training and an updated version of FGENESH (for mammals) incorporates GC donor sites in its predictions.

Analysis of spliced alignments between clustered *Arabidopsis* EST and genomic sequences also showed that the canonical GT–AG pairs account for the majority of the splice sites in *Arabidopsis* (Zhu *et al*., 2003). In that species the frequencies of the non-canonical GC–AG and AT–AC sites have been estimated to be about 1.0% and 0.06%, respectively. These may, however be over-estimates because ambiguous splice sites were included in this analysis (Zhu *et al*., 2003). In our data set 2, 99.1% of all sites were canonical GT–AG pairs and non-canonical GC–AG and AT–AC pairs represent 0.822% and 0.103% of all pairs, respectively. This result indicates that the fractions of non-canonical GC–AG pairs in maize and *Arabidopsis* and AT–AC pairs in maize may be higher than in mammalian genomes.

By using the '–GC' parameter, FGENESH was able to identify 81% (13/16) of the non-canonical GC donor sites in data set 2. Since most donor sites in data set 2 are canonical and the sensitivity of them is 0.73, FGENESH's sensitivity for non-canonical GC sites is at least as good as it's sensitivity for canonical GT sites.

Recommendations for gene prediction

Of the evaluated *ab initio* programs FGENESH provided the highest degree of SP and SN, followed by GeneMark.hmm. Both of these pro-

grams provide high levels of SP with acceptable (but somewhat lower) levels of SN. Consequently, if a sequence is predicted to contain a gene, that prediction is likely to be correct, but some sequences that do contain genes will be missed. Using its '–GC' parameter, FGENESH is able to identify many non-canonical GC donor sites. Removing exons with negative FGENESH scores will eliminate most of the WEs, while retaining the majority of the TEs and PEs + OEs. Therefore, for RT-PCR experiments and microarray design projects it is better to avoid designing primers or oligos in predicted exons with negative scores. If the specificity of exon prediction is the priority, predicted exons with even higher scores (≥10) should be used. Although will result in the loss of correct exons, it will also eliminate essentially all wrong exons.

Combining gene prediction results from multiple *ab initio* programs improves gene model predictions (reviewed by Mathé *et al*., 2002) because even a good program can make incorrect predictions for some genes and even a poor program can make correct predictions for some genes. For example, as shown in Table 7, FGENESH did not correctly predict the gene model of *pdc3*, but the other three programs did. Moreover, analysis of predictions of genes in data set 2 suggests that by considering predictions from FGENESH, GeneMark.hmm and GENSCAN, it is possible to improve the accuracy of *ab initio* gene discovery (Figure 4). Integration of *ab initio* and sequence similarity based approaches is another way to improve the accuracy of gene prediction and is likely to be more widely used as the number of sequenced genomes increases (reviewed by Mathé *et al*., 2002). The Twinscan (Korf *et al*., 2001) and Combiner (Allen *et al*., 2004) programs improve the accuracy of gene predictions via these two approaches. The development of similar programs or the training of existing programs for maize sequences could also contribute to the efficient discovery of maize genes.

No matter which approaches are taken, a good training data set is essential to improve the accuracy of gene predictions in new species. The most straightforward method to improve the quality of a training set is to increase the number of gene sequences. Korf (2004) has recently suggested an alternative computational approach for situations when this is not possible.

Acknowledgments

The *rth3*, *rf2e1*, *pdc2*, and *pdc3* genes were cloned using cDNA clones identified via searches of Pioneer Hi-Bred's proprietary EST database. We thank Tim Fox, Wes Bruce and Carl Simmons (Pioneer Hi-Bred, Intl. Inc., Johnston, IA) for their assistance with these searches. We thank Heike Hofmann (Iowa State University) for assistance with the Kolmogorov-Smirnov tests and Volker Brendel (Iowa State University) for helpful discussions. This research was funded in part by competitive grants from the National Science Foundation Plant Genome Program (awards: DBI-9975868, DBI-0121417, and DBI-0321711) and the United States Department of Agriculture National Research Initiative Program (awards: 98101805, 0001478, 0101869, 0201419, and 0300940). Support was also provided by Hatch Act and State of Iowa funds.

References

Allen, J.E., Pertea, M. and Salzberg, S.L. 2004. Computational gene prediction using multiple sources of evidence. Genome Res. 14: 142–148.

Bennetzen, J.L., Chandler, V.L. and Schnable, P.S. 2001. National Science Foundation-sponsored workshop report. Maize genome sequencing project. Plant Physiol. 127: 1572–1578.

Brendel, V., Xing, L. and Zhu, W. 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20: 1157–1169.

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

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

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

Burset, M., Seledtsov, I.A. and Solovyev, V.V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28: 4364–4375.

Burset, M., Seledtsov, I.A. and Solovyev, V.V. 2001. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 29: 255–259.

Chen, M. and Bennetzen, J.L. 1996. Sequence composition and organization in the *Sh2*/*A1*-homologous region of rice. Plant Mol. Biol. 32: 999–1001.

Chen, M., SanMiguel, P. and Bennetzen, J.L. 1998. Sequence organization and conservation in *sh2*/*a1*-homologous regions of sorghum and rice. Genetics 148: 435–443.

Civardi, L., Xia, Y., Edwards, K.J., Schnable, P.S. and Nikolau, B.J. 1994. The relationship between genetic and physical distances in the cloned *a1*-*sh2* interval of the *Zea mays* L. genome. Proc. Natl. Acad. Sci. USA 91: 8268–8272.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.-J., Narayanan, M., Guo, L., Ashlock, D.A. and Schnable, P.S. 2004. Astrategy for assembling the maize (*Zea mays* L.) genome. Bioinformatics 20: 140–147.

Goff, S.A. *et al*. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92–100.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, 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.

Kolmogorov, A.N. 1933. Sulla determinazione empirica di una legge di distribuzione. Giornale dell' Istituto Italiano degli Attuari 4: 83–91.

Korf, I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Korf, I.P., Flicek, D.D. and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17: 140–148.

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

Mathé, C., Sagot, M.F., Schiex, T. and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30: 4103–4117.

Moore, G. 2000. Cereal chromosome structure, evolution, and pairing. Annu. Rev. Plant Physiol. Plant Mol. Biol. 51: 195–222.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M., Martienssen, R.A. and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. Science 302: 2115–2117.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P. and Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics 15: 887–899.

Pertea, M., Lin, X. and Salzberg, S.L. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 29: 1185–1190.

Pertea, M. and Salzberg, S.L. 2002. Computational gene finding in plants. Plant Mol. Biol. 48: 39–48.

Peterson, D.G., Wessler, S.R. and Paterson, A.H. 2002. Efficient capture of unique sequences from eukaryotic genomes. Trends Genet. 18: 547–550.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R. and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nature Genet. 23: 305–308.

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

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

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24–31.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765–768.

Smirnov, N.V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. Bull. Moscow University 2: 3–16.

Solovyev, V. 2001. Statistical approaches in eukaryotic gene prediction. In: D.J. Balding, M. Bishop and C. Cannings (Eds.), Handbook of Statistical Genetics, John Wiley & Sons, Ltd, New York, pp. 83–127.

Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. Genome Res. 10: 394–397.

The *Arabidopsis* Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Tolstrup, N., Rouze, 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.

Usuka, J. and Brendel, V. 2000. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J. Mol. Biol. 297: 1075–1085.

Usuka, J., Zhu, W. and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203–211.

Wang, J., Li, S., Zhang, Y., Zheng, H., Xu, Z., Ye, J., Yu, J. and Wong, G.K. 2003. Vertebrate gene predictions and the problem of large genes. Nat. Rev. Genet. 4: 741–749.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M. and Quackenbush, J. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302: 2118–2120.

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

Yao, H., Zhou, Q., Li, J., Smith, H., Yandeau, M., Nikolau, B.J. and Schnable, P.S. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. Proc. Natl. Acad. Sci. USA 99: 6157–6162.

Yu, J. *et al*. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79–92.

Yuan, Q., Quackenbush, J., Sultana, R., Pertea, M., Salzberg, S.L. and Buell, C.R. 2001. Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species. Plant Physiol. 125: 1166–1174.

Yuan, Y., SanMiguel, P.J. and Bennetzen, J.L. 2003. High-Cot sequence analysis of the maize genome. Plant J. 34: 249–255.

Zhu, W., Schlueter, S.D. and Brendel, V. 2003. Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. Plant Physiol. 132: 469–484.