BIOINFORMATICS

Vol. 21 Suppl. 3 2005, pages iii20–iii30
doi:10.1093/bioinformatics/bti1205

Genome analysis

Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants

Michael E. Sparks[1] and Volker Brendel[1,2,*]

[1]Department of Genetics, Development and Cell Biology and [2]Department of Statistics, Iowa State University, 2112 Molecular Biology Building, Ames, IA 50011-3260, USA

Received on June 13, 2005; accepted on August 16, 2005

ABSTRACT

Motivation: The vast majority of introns in protein-coding genes of higher eukaryotes have a GT dinucleotide at their 5′-terminus and an AG dinucleotide at their 3′ end. About 1–2% of introns are non-canonical, with the most abundant subtype of non-canonical introns being characterized by GC and AG dinucleotides at their 5′- and 3′-termini, respectively. Most current gene prediction software, whether based on *ab initio* or spliced alignment approaches, does not include explicit models for non-canonical introns or may exclude their prediction altogether. With present amounts of genome and transcript data, it is now possible to apply statistical methodology to non-canonical splice site prediction. We pursued one such approach and describe the training and implementation of GC-donor splice site models for *Arabidopsis* and rice, with the goal of exploring whether specific modeling of non-canonical introns can enhance gene structure prediction accuracy.

Results: Our results indicate that the incorporation of non-canonical splice site models yields dramatic improvements in annotating genes containing GC–AG and AT–AC non-canonical introns. Comparison of models shows differences between monocot and dicot species, but also suggests GC intron-specific biases independent of taxonomic clade. We also present evidence that GC–AG introns occur preferentially in genes with atypically high exon counts.

Availability: Source code for the updated versions of GeneSeqer and SplicePredictor (distributed with the GeneSeqer code) is available at http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. Web servers for *Arabidopsis*, rice and other plant species are accessible at http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/AtGDBgs.cgi, http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/OsGDBgs.cgi and http://www.plantgdb.org/PlantGDB-cgi/GeneSeqer/PlantGDBgs.cgi, respectively. A SplicePredictor web server is available at http://bioinformatics.iastate.edu/cgi-bin/sp.cgi. Software to generate training data and parameterizations for Bayesian splice site models is available at http://gremlin1.gdcb.iastate.edu/~volker/SB05B/BSSM4GSQ/

Contact: vbrendel@iastate.edu

Supporting information: http://gremlin1.gdcb.iastate.edu/~volker/SB05B/

INTRODUCTION

Most genes in higher eukaryotic organisms contain intervening sequences (‘introns’), which must be precisely excised from the pre-mRNA transcript prior to being translated into a functional protein. The 5′-terminus of an intron is commonly known as the donor site, whereas the 3′-terminus is referred to as the acceptor site. These terms correlate with the roles of these sites in the biochemical reactions underlying the process of splicing, as catalyzed by the spliceosome. Most introns belong to the class of canonical introns, characterized by a GT donor dinucleotide (first two bases of the intron) and an AG acceptor dinucleotide (last two bases of the intron), and are processed by the U2-type splicing apparatus (Burge *et al*., 1999; Reddy, 2001). The most common deviations from these intron types are those that have GC donors and AG acceptors. In all eukaryotic species studied so far, these introns make up ∼1% of all introns and are presumably also processed by the U2-type spliceosome (Burset *et al*., 2000). More recently, a second type of spliceosome has been identified which recognizes the so-called U12-type introns. These introns share a highly conserved donor site consensus [GA]T/ATCCTT (where [GA] means G or A and ‘/’ indicates the exon/intron border) and a conserved branch point motif CCTTAAC (reviewed by Patel and Steitz, 2003).

Many recent studies have discussed the occurrence and splicing of U12-type introns as well as their potential functions and phylogenetic origin (Burge *et al*., 1998; Dietrich *et al*., 2001; Lynch and Richardson, 2002; Patel *et al*., 2002; Zhu and Brendel, 2003). U12-type introns occur almost invariably in genes with other U2-type introns, with no significant over-representation in any particular functional gene class, appear less common among short introns and may function in post-transcriptional regulation of gene expression. Evolutionarily, they are thought to have an ancient origin, with loss and conversion to U2-type accounting for their sparse occurrence in modern genomes.

Comparatively little attention has been devoted to introns with GC-donors. Thanaraj and Clark (2001) described statistical features of human GC–AG alternative intron isoforms. Kitamura-Abe *et al*. (2004) identified several hundred GC–AG introns in the human, mouse, fruit fly, *Arabidopsis* and rice genomes and provided a descriptive analysis of mono- and dinucleotide frequencies around the splice sites. Their results suggest that GC-donors may show a stronger consensus to maximize base pair formation with complementary positions in the U1 snRNA.

The ability to computationally predict splice sites in pre-mRNA tests our theoretical understanding of the sequence features recognized by the splicing machinery. Several well-supported computational approaches to splice site prediction in pre-mRNA sequences are available, including NetGene2 (Brunak *et al*., 1991; Hebsgaard *et al*., 1996), SplicePredictor (Brendel and Kleffe, 1998; Brendel

*To whom correspondence should be addressed.

et al., 2004), and GeneSplicer (Pertea *et al*., 2001). Because all these approaches require large amounts of trusted exon/intron borders for training that have heretofore not been accessible, the available programs have largely excluded prediction of non-canonical sites. Of the above programs, only NetGene2 predicts GC-donors. SplicePredictor optionally allows scoring of non-canonical splice sites as if the terminal dinucleotide matched the consensus. Similarly, most gene structure prediction tools, whether based on *ab initio* or spliced alignment approaches, have not incorporated explicit models for non-canonical introns or preclude their prediction altogether.

In the current study, we pursue the characterization and prediction of GC–AG introns in plant pre-mRNAs. The motivation for this study derives from the availability of two complete plant genomes, representing dicotylodonous and monocotylodonous plants (*Arabidopsis* and rice). Given the large numbers of public full-length cDNAs and ESTs for these genomes, genome-wide assessment of the occurrence of non-canonical introns should now approximate the final picture very closely. We present general software for the estimation of Bayesian statistical models for splice site prediction (Brendel *et al*., 2004) from reliable spliced alignments of (full-length) cDNAs. Using the models for exon/intron junction prediction in spliced alignments, we show that GC–AG introns can be confidently predicted from cDNA to genomic DNA matches even in the presence of considerable sequence divergence. These models could also be incorporated into *ab initio* gene structure prediction software and should aid in closing the annotation gap for both model and emerging genomes (Mathé *et al*., 2002; Schlueter *et al*., 2005).

SYSTEMS AND METHODS

Training data

We accumulated intron data for two model plant species, *Arabidopsis thaliana* and *Oryza sativa* as follows. All available full-length cDNA sequences for each species (64 840 entries for *Arabidopsis* and 32 136 entries for rice) were aligned to their cognate genomes [assembly version 5 for *Arabidopsis* (Wortman *et al*., 2003) and pseudochromosome assembly version 3.0 for rice (Yuan *et al*., 2003)] using the GeneSeqer spliced alignment tool (Brendel *et al*., 2004). These data can be accessed at http://gremlin1.gdcb.iastate.edu/∼volker/SB05B/original_alignments/.

For model training, we considered only gene structures such that all transcript sequence(s) delimiting them aligned with a perfect overall sequence similarity score (1.0), allowing a very high degree of confidence that exon/intron borders in our training data were correctly resolved. We extracted all exons and introns and classified them into phase 1, 2 or 0, where an intron is in phase 1 if it falls between two codons, phase 2 if it falls between the first and second positions of a codon and phase 0 if it falls between the second and third positions. Only the coding regions of exons are considered (according to the putative translation product of the gene), so that the first exon of any gene is by default classified as phase 1. For all other exons, their phase corresponds to that of the upstream intron.

For each of the three phase classes we identified introns with GC–AG termini and extracted from the genomic templates, for both donor and acceptor termini, 50 nt upstream through 50 nt downstream. Redundant entries were removed. In a similar way we compiled sequence composition data for false within-exon (for the three different phases) and within-intron GC donor and AG acceptor sites. Random sampling of these data after removing redundant entries produced sample sizes equal to those of the corresponding true donor and acceptor site data; the within-intron false sites were randomly sampled to the size of the largest of the three true phase classes. Sizes of the training data corpora are shown in Table 1. Consistent with previous observations (Ruvinsky *et al*., 2005), most introns fall into the phase 1 class.

Table 1. Classification and counts of cDNA-confirmed introns

	Total[a]	GT–AG	GC–AG	Others
Arabidopsis thaliana	67 767	66 733 (98.47%)	721 (1.06%)	313 (0.46%)
TD phase 1		37 904	476	
TD phase 2		14 126	133	
TD phase 0		14 703	112	
Oryza sativa	68 199	65 391 (95.88%)	1103 (1.62%)	1705 (2.50%)
TD phase 1		36 571	644	
TD phase 2		14 310	270	
TD phase 0		14 510	189	

[a]Total counts of each intron type are given for both *Arabidopsis* and rice. Relative abundances are indicated in parentheses. TD, training data (see Systems and Methods section).

These data were then used to parameterize new Bayesian splice site models for GC–AG introns as described previously (Brendel *et al*., 2004). Our software to generate such training datasets and parameterizations from GeneSeqer alignment data and genomic template files is available free of charge for academic or other non-profit use at http://gremlin1.gdcb.iastate.edu/∼volker/SB05B/BSSM4GSQ/.

Information plots and pictograms

It has previously been noticed that there is a stricter adherence to a consensus donor splice site sequence in GC–AG introns than in GT–AG ones (Burset *et al*., 2000; Kitamura-Abe *et al*., 2004). Pooling all sets of cDNA-confirmed training introns (Table 1), we computed the information content for two regions of interest—15 bases upstream through 20 bases downstream of the donor sites and 20 bases upstream through 15 bases downstream of the acceptor sites—using the following formula:

$$I_i = 2 + \sum_{B \in (\mathrm{A,C,G,T})} f_{iB} \log_2(f_{iB})$$

where i indexes each position in the aligned sequences, and f_{iB} is the frequency of base B in position i (White *et al*., 1992; Rogan and Schneider, 1995).

Splice site probability models

Using the Bayesian splice site model framework first described in Salzberg (1997) and later adapted for use by GeneSeqer in Brendel *et al.* (2004), we trained models for GC donor sites of GC–AG introns in *Arabidopsis* and rice. Briefly, this consisted of tabulating dinucleotide relative frequencies over the 102 positions of interest for the donor sites, for seven classes of training data corresponding to seven alternative hypotheses to be evaluated using Bayes rule. The following parameter smoothing technique was used to avoid problematic zero-probability transition probabilities in our models owing to unobserved data, i.e. to emulate 'pseudocounts' (zero-probability transition probabilities are necessary when transitioning into and out from the donor or acceptor dinucleotides and were not adjusted in this process). Dinucleotide frequencies at each position over the training region can be construed as a 4×4 matrix, where rows correspond to the mononucleotide being departed from and columns correspond to that being transitioned into. If certain dinucleotides were not observed at various positions in the training data, then this would erroneously produce rows of zero-likelihood transition probabilities in the matrices; for such cases, we set all four probabilities to 0.25. Otherwise, if any dinucleotide transition probability in a row was below a threshold of 0.0005, then we set it to $P_{\mathrm{fix}} = 0.05$, and all non-zero values in the row were adjusted to $P_{\mathrm{new}} = P_{\mathrm{old}} * (1 - 4 * P_{\mathrm{fix}}) + P_{\mathrm{fix}}$ to produce a valid probability mass distribution, where P_{old} refers to the unadjusted parameter. We elected to use 0.05 for P_{fix} as empirically it gave the most reasonable results of a variety of values we tested (data not shown).

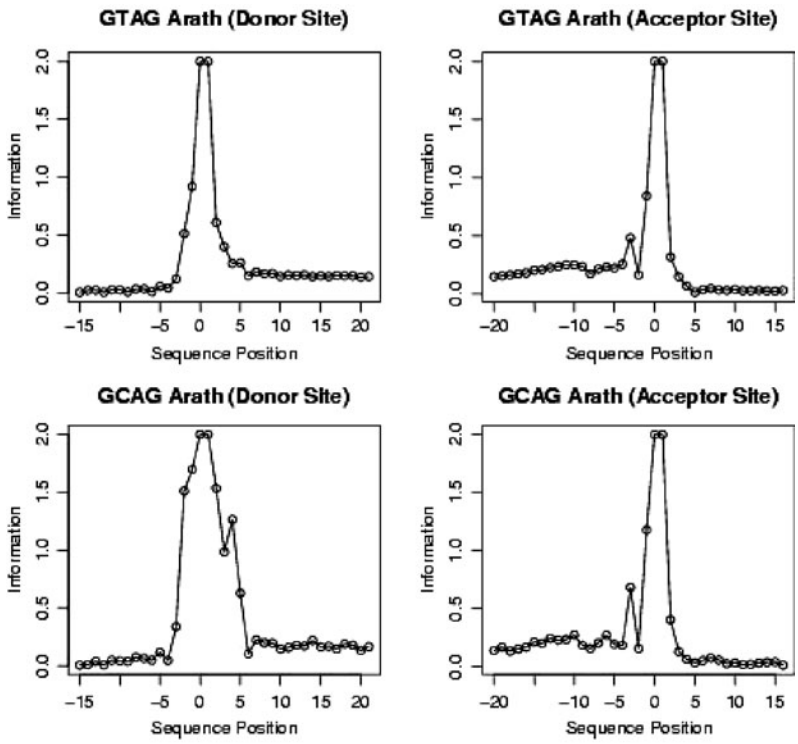

Fig. 1. *Arabidopsis* information content plots. Information content plots were produced for 15 bases upstream through 20 bases downstream of GT and GC donor dinculeotides, and 20 bases upstream through 15 bases downstream of AG acceptor dinucleotides in GC–AG and GT–AG *Arabidopsis* introns identified in the training data set (see Table 1).

Comparison of splice site models

The Bayesian models described above each yield $49 * 16$ first order Markov transition probabilities going into the nucleotide preceding the splice site dinucleotide and the same number of probabilities going out of the nucleotide following it. We considered six models in total: species-specific GT donor models of GT–AG introns trained for *Arabidopsis*, rice, maize and *Medicago truncatula*, and the GC donor models for *Arabidopsis* and rice. To compare the different models, we considered these parameters in order as components of 784-dimensional vectors and calculated distances between the models as the Euclidean distance between specific component ranges of these vectors. Inspection of the information content plots in Figures 1 and 2 indicate that there is significant loss of signal roughly five sites before and after the donor dinucleotides. However, we wanted to determine if there were distinct patterns of dinucleotide usage in these otherwise uninformative regions proximal to the donor splice site. We first considered only the initial 720 elements of the upstream parameter vectors, corresponding to 15 codons prior to the terminal codon of the training exons. Because they are the most abundant category of training exons (Table 1), phase 1 sites were used for this analysis. Phase 2 and 0 data produced similar results (data not shown; necessarily, all distances must be calculated relative to the same codon phase, so pooling the sets is inappropriate). For balance, only the final 720 elements of the downstream parameter vectors were used, representing the final 45 positions of the training region in the downstream intron. We also considered independently the 192 elements of the vectors corresponding to five positions upstream of the donor dinucleotide through five positions downstream, with the modification that the parameters for the GC-donor model donor dinucleotides involving the C-position were shifted to resemble GT-donors, preventing trivially inflated Euclidean distances when compared with GT-donor vectors. The exon, intron and splice site distance matrices were used as input to the neighbor-joining tree building program implemented in the PHYLIP package (Felsenstein, 2004, http://evolution.genetics.washington.edu/phylip.html).

Incorporation of splice site probability models in the scoring of spliced alignments

The GeneSeqer spliced alignments are optimized with respect to several parameters, including weights for identities (default value: 2.0), mismatches (default value: -2.0) and deletions (default value: -5.0 per gap symbol) within exon alignments as well as logarithmically transformed exon/intron state transition probabilities derived from splice site prediction values along the genomic sequence (Usuka *et al*., 2000). Default donor site probabilities are 0.00005 for any GT and 0.00002 for any GC or AT (similarly, 0.00005 for any AG and 0.00002 for any AC as potential acceptor sites), with all other dinucleotides assigned a default donor or acceptor site probability of 0.000001. These default values are replaced by $2 \times (P - 0.5)$ whenever that value is greater, where P is the respective Bayesian a posteriori splice site probability as derived from the training data described above. As a simple rule to recognize U12-type introns, sites matching the U12-type intron consensus sequence ATCCTT downstream of the GT or AT donor site dinucleotide in six or five positions are scored 0.99 and 0.9, respectively (Brendel *et al*., 2004).

Programs used

We used a previous version of GeneSeqer (henceforth referred to as 'GeneSeqerSTD') as a prototype upon which to incorporate the new models. The source code of this older version is available from the authors on request. Source code for the revised GeneSeqer (version of June 13, 2005, referred to as 'GeneSeqerGC' for this paper) is available at http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. For spliced alignment assays, Sim4

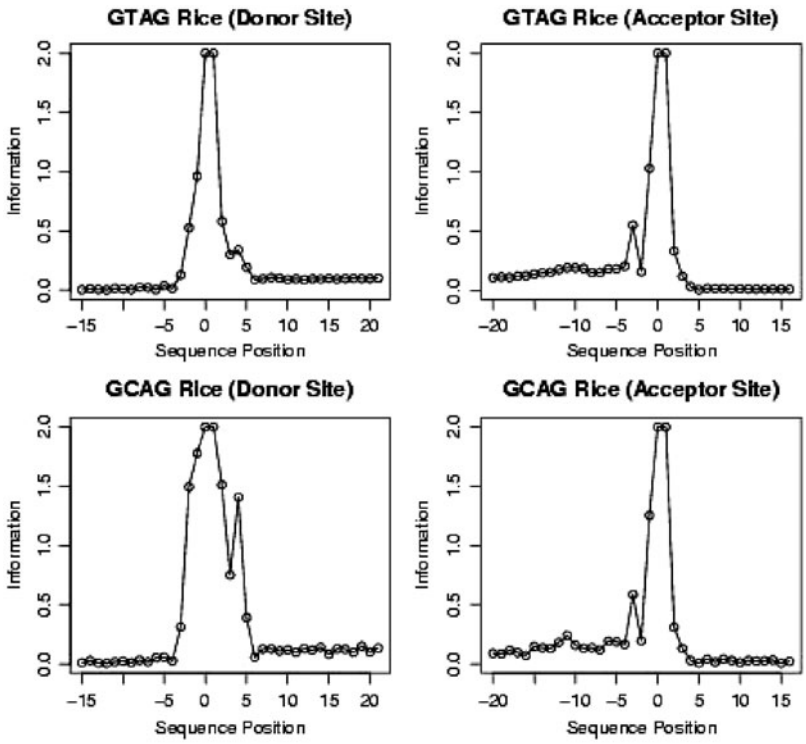

Fig. 2. Rice information content plots. Information content plots were produced for 15 bases upstream through 20 bases downstream of GT and GC donor dinculeotides, and 20 bases upstream through 15 bases downstream of AG acceptor dinucleotides in GC–AG and GT–AG rice introns identified in the training data set (Table 1).

(Florea *et al*., 1998) was downloaded from http://globin.cse.psu.edu/, Spidey (Wheelan *et al*., 2001) was obtained from http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/spideyexec.html, BLAT (Kent, 2002) was downloaded from http://www.soe.ucsc.edu/~kent/src/ and Splign (Kapustin *et al*., 2004, http://recomb04.sdsc.edu/posters/kapustinATncbi.nlm.nih.gov_76.pdf) was obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/splign/. For *ab initio* gene structure prediction, we used GENSCAN (Burge and Karlin, 1997), obtained from http://genes.mit.edu/license.html, and the FGENESH_GC (Solovyev, 2001) and Eukaryotic GeneMark.hmm (M. Borodovsky and A. Lukashin, unpublished data) web servers, available at http://www.softberry.com/berry.phtml?topic=fgeneshgc&group=programs&subgroup=gfind and http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi, respectively. For *ab initio* splice site prediction, we compared the published SplicePredictor version (Brendel *et al*., 2004), referred to here as SplicePredictorSTD, with a new version incorporating the GC-donor site models, referred to here as SplicePredictorGC (obtainable as part of the GeneSeqer code distribution); SplicePredictorSTD was modified to score all GC dinucleotides in the same way as GT dinucleotides. The NetGene2 program (Hebsgaard *et al*., 1996) was used through the NetPlantGene mail server, accessible at http://www.cbs.dtu.dk/services/NetPGene/mailserver.php. GeneSplicer (Pertea *et al*., 2001) gives good results for canonical splice sites, but it does not predict non-canonical splice sites and was therefore irrelevant to this study.

Test data

We accumulated 100 GC donor-containing test loci each for *Arabidopsis* and rice for purposes of comparing spliced alignment programs and assessing *ab initio* gene structure and splice site prediction programs as follows. Using the alignment data mentioned above, we identified Predicted Gene Locations (PGLs) containing four or more introns, exactly one of which had to be of the GC–AG variety, and such that one or more cDNA sequences supporting the gene structure aligned with an overall similarity score of at least 0.975 but not more than 0.995 (and none of the supporting cDNA evidence aligned with a score <0.975). For each such PGL, we extracted the genic locus and 200 nt of upstream and downstream flanking sequence directly from the corresponding pseudochromosome. A total of 100 test loci were randomly sampled from this population. For each test locus, an associated 'pseudotranscript' was also parsed directly from the genome based on the gene structure coordinates given in the PGL.

It was important that the test genes were not a part of our GC donor training dataset, as this would produce artificially elevated accuracy assessments for SplicePredictor with the newly trained models (SplicePredictorGC), which would have been explicitly trained on the test data. Our criteria of having at least one supporting cDNA sequence for a given PGL falling in the 0.975–0.995 similarity score range for a test locus precluded the gene from having been incorporated in the training dataset, nevertheless still permitted resolution of reliable gene structures for our test set using transcript evidence.

We also compiled an independent control test set of 100 genes in *Arabidopsis* and rice just as we compiled the GC–AG intron-containing test sets, with the exception that the control genes had to have at least four introns, but all of the GT–AG type. Similarly, sets of 25 U12 intron-containing genes (with AT donor and AC acceptor dinucleotide termini) for these two taxa were collected. All of these test datasets are available at http://gremlin1.gdcb.iastate.edu/~volker/SB05B/test_data/.

Spliced alignment assays

Performance assessments of spliced alignment software were conducted at the 0, 1, 5, 10 and 25% simulated transcript sequencing error levels by attempting to match mutated transcripts to their cognate genomic loci. To generate

the mutated sequences for a given simulated sequencing error level of X%, X instances of either point substitution or insertion/deletion mutations (of a length randomly selected from 1 to 3 residues) per 100 bases of the pseudotranscript were induced. Software used for this task is available at http://gremlin1.gdcb.iastate.edu/∼volker/SB05B/misc/mutseqer.c.

We compared GeneSeqerGC with GeneSeqerSTD, Splign, BLAT, Sim4 and Spidey. Exact program parameters used for *Arabidopsis* and rice, including the actual spliced alignment results, are provided as supporting data at http://gremlin1.gdcb.iastate.edu/∼volker/SB05B/test_alignments/. At each simulated sequencing error level, 25 replicates were produced, permitting both accuracy and precision assessments for each program on the 100 *Arabidopsis* and rice test loci. These assays tested for competency in both GC-donor site detection alone and complete gene structure resolution. The latter was assessed on the intron level—a gene structure prediction was counted completely correct whenever all predicted intron borders matched those of the true gene structure.

Ab initio assays

We tested the GENSCAN, GeneMark.hmm and FGENESH-GC *ab initio* gene structure prediction and SplicePredictorGC, SplicePredictorSTD and NetGene2 *ab initio* splice site prediction programs on the GC–AG and AT–AC test loci to determine if any of these software were capable of annotating known GC donors or AT–AC introns, respectively. As mentioned above, SplicePredictorSTD was modified to score all GC dinucleotides in the same way as GT dinucleotides. For the gene structure prediction tools, we also assessed each program's ability to delineate complete gene structures.

Identification of GC–AG introns in *Arabidopsis* and rice

After incorporation of the GC–AG models into GeneSeqerGC, we reannotated the *Arabidopsis* and rice genomes using full-length cDNA sequences. We considered only introns derived from gene structures such that all cDNA sequences mapping to a gene yielded an overall score not less than 0.95. For *Arabidopsis*, of a total of 70 803 introns, 69 474 (98.12%) were of GT–AG type, 776 (1.10%) were of GC–AG type and 553 (0.78%) were of other types. For rice, of a total of 71 099 introns, 65 337 (91.90%) were of GT–AG type, 1804 (2.54%) were of GC–AG type and 3958 (5.57%) were of other types. These results show higher proportions of GC–AG and other type introns relative to those identified in the training data set described in Table 1, particularly for rice. This probably results from a combination of incorporation of explicit GC–AG models in GeneSeqerGC and the lower stringency used to cull this dataset (overall score of 0.95 or greater versus 1.0 used to compile training data).

Gene Ontology (GO) annotation

To test whether GC–AG intron containing genes have preferential occurrence in particular functional classes of genes, we evaluated the distribution of GOslim terms (Berardini *et al*., 2004; downloaded from ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/) associated with a set of 622 annotated *Arabidopsis* genes (Wortman *et al*., 2003), each containing at least one GC–AG intron confirmed by our reannotation of the *Arabidopsis* genome, described below. Significance of the distribution was evaluated by random sampling of same-size sets of non-GC–AG intron containing *Arabidopsis* genes. A particular GOslim category was regarded as over- or underrepresented if the frequency of the term in the GC set was in the top five or lower five values in comparison with 99 randomly drawn non-GC sets.

RESULTS AND DISCUSSION

Information plots and pictograms

Information content profiles for donor and acceptor sites of *Arabidopsis* and rice are shown in Figures 1 and 2, respectively. These data support the notion of stricter adherence to a donor consensus site in GC donors, promising good potential for statistical prediction. Information content profiles at acceptor sites are very similar, in either species, between GT–AG and GC–AG intron types. Furthermore, the superimposable acceptor site information content profiles indicate that the elevated information content at GC donors relative to GT donors is not an artifact owing to differing sample sizes between our GT–AG and GC–AG intron populations. This latter observation suggests that, in biological systems, constraints on the degree of adherence to a consensus acceptor sequence, for any particular acceptor site, operate independently of whether a GT or GC donor occurs upstream of it. The consensus sequences of GT–AG and GC–AG donor sites are identical, with the exception of the T ↔ C transition mutation (data not shown). This reduced combinatorial complexity simplifies incorporation of a GC donor model into the framework of the GeneSeqer algorithm: when considering what an appropriate donor site might be for a given intron, the algorithm does not need to modify its procedures for selecting an appropriate acceptor site. This also implies that the parameterizations for acceptor sites in the Bayesian splice site models trained on GT–AG data are sufficient for assessing acceptor sites in both intron types.

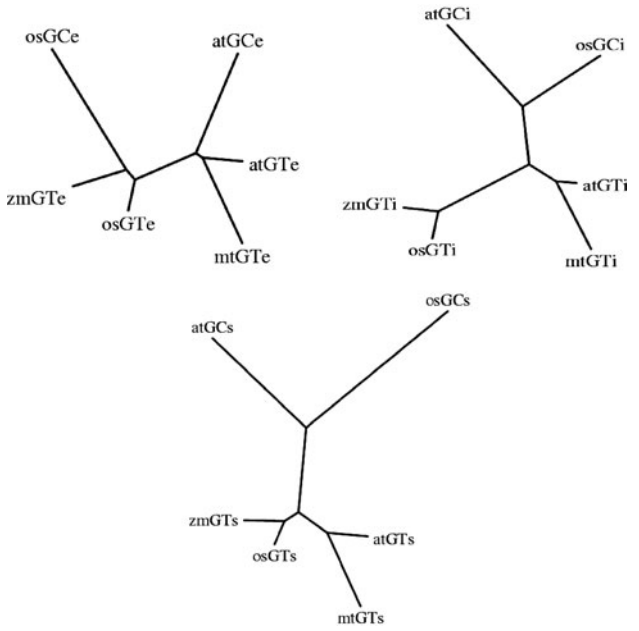

Fig. 3. Neighbor-joining trees derived from donor site model-specific parameter vectors. at, *Arabidopsis thaliana*; os, *Oryza sativa*; mt, *Medicago truncatula*; zm, *Zea mays*. GT, GT-donors; GC, GC-donors; e, exon parameters; i, intron parameters; s, splice site parameters.

Differences between species-specific models

Figure 3 shows neighbor-joining derived topologies based on the Euclidean distances between the parameter sets of the various donor site models (see Systems and Methods section). For GT-donors, exon, intron and splice site parameters cluster according to the monocot/dicot divide. The GC-donor intron and splice site parameters of *Arabidopsis* and rice are nearest neighbors, closer to each other than to the GT-donor intron parameters within the same taxonomic clade. For the exon parameters, the GC-donor parameters group with their taxonomic clade, but with long branch lengths relative to the close pairs of GT-donor parameters within each clade.

Table 2. Accuracy of GC-donor site prediction

Program	0%	1%[a]	5%	10%	25%
Arabidopsis					
GeneSeqerGC	100	99.72 (0.32)	99.76 (0.97)	95.92 (1.43)	80.48 (3.04)
GeneSeqerSTD	100	97.28 (1.32)	83.92 (2.58)	69.84 (3.01)	33.40 (3.49)
Splign	100	98.24 (1.17)	88.88 (2.22)	74.72 (2.61)	0.00 (0.00)
BLAT	100	95.64 (1.93)	74.84 (3.45)	52.08 (4.03)	4.40 (1.74)
Sim4	92	89.72 (1.22)	80.36 (3.28)	67.92 (2.95)	17.60 (2.92)
Spidey	84	79.72 (1.41)	60.04 (2.72)	45.76 (3.39)	7.08 (1.68)
Rice					
GeneSeqerGC	100	98.44 (0.97)	92.32 (1.45)	82.56 (6.83)	31.88 (15.20)
GeneSeqerSTD	100	97.40 (1.37)	80.72 (12.33)	57.60 (17.65)	10.36 (8.88)
Splign	99	97.88 (1.22)	86.40 (12.14)	66.08 (15.32)	0.00 (0.00)
BLAT	93	90.16 (1.08)	63.36 (15.01)	29.96 (15.19)	0.92 (0.83)
Sim4	91	90.16 (1.00)	73.24 (14.67)	57.52 (17.46)	20.60 (12.97)
Spidey	85	78.40 (9.67)	61.64 (17.04)	37.64 (17.28)	6.56 (6.33)

[a]Percent simulated sequencing error induced on pseudotranscripts (see Systems and Methods section for details). The table values give the mean and standard deviations (in parentheses) of successful detection of known GC-donor sites in 100 GC–AG intron-containing test loci for each of *Arabidopsis* and rice over 25 replicates.

Table 3. Accuracy of overall GC–AG intron-containing gene structure prediction

Program	0%	1%[a]	5%	10%	25%
Arabidopsis					
GeneSeqerGC	100	91.52 (2.37)	62.88 (3.94)	41.32 (3.60)	7.40 (1.75)
GeneSeqerSTD	100	89.96 (2.60)	55.84 (4.39)	32.72 (3.88)	3.52 (1.43)
Splign	97	88.12 (2.53)	58.68 (4.07)	22.08 (2.70)	0.00 (0.00)
BLAT	100	31.60 (4.28)	0.36 (0.45)	0.00 (0.00)	0.00 (0.00)
Sim4	88	79.44 (1.63)	51.56 (4.17)	25.80 (3.21)	0.04 (0.14)
Spidey	96	48.20 (4.38)	6.64 (2.31)	0.36 (0.45)	0.00 (0.00)
Rice					
GeneSeqerGC	100	95.68 (1.83)	78.96 (6.83)	66.20 (9.66)	22.36 (15.43)
GeneSeqerSTD	100	94.80 (2.32)	71.00 (12.20)	47.16 (17.38)	6.24 (8.80)
Splign	90	87.60 (1.50)	72.36 (11.83)	50.36 (15.65)	2.56 (5.49)
BLAT	83	71.92 (7.63)	34.88 (17.72)	10.64 (11.32)	3.40 (4.11)
Sim4	79	78.00 (0.97)	60.64 (14.75)	47.08 (16.89)	13.28 (12.56)
Spidey	84	73.36 (10.04)	46.36 (17.51)	23.64 (15.89)	2.56 (5.49)

[a]Percent simulated sequencing error induced on pseudotranscripts (see Systems and Methods section for details). The table values give the mean and standard deviation (in parentheses) of successful detection of entire gene structure in 100 GC–AG intron-containing test loci for each of *Arabidopsis* and rice over 25 replicates.

These results underscore not only compositional differences between monocots and dicots, but in particular they suggest considerable biases associated with GC–AG intron and GC-donor splice site dinucleotide compositions relative to their GT–AG counterparts. It is unclear whether these biases reflect evolutionary history of this class of introns, functional constraints on their splicing or overall compositional biases of gene classes that harbor GC-donor introns (see below).

Spliced alignment

We wanted to compare the performance of GeneSeqerGC to other spliced alignment software on two levels: detection of known GC donor splice sites and correct resolution of full gene structures. Assessments were made using the 100 GC donor-containing spliced alignment test loci for *Arabidopsis* and rice described above. Results of these respective experiments are presented in Tables 2 and 3. It is seen that GeneSeqerGC significantly outperforms spliced alignment programs without explicit GC-donor splice site models, with a >80% GC-donor site detection rate even at the 10% sequence error level. This rate is ~20% higher than the best of the other programs (Splign). More than 40% of the entire gene structures are predicted correctly at the same sequence error level. The increase in performance when compared with GeneSeqerSTD demonstrates that the improvement is because of the specific GC-donor site models, rather than other features of the GeneSeqer algorithm.

To ensure that these improvements do not cause significant detriment to the accurate spliced alignment-based annotation of a typical eukaryotic gene containing exclusively GT–AG introns, we tested the same set of spliced alignment programs on our GT–AG control test data for competency at determining overall gene structures, the results of which are presented in Table 4. Comparison of GeneSeqerGC with GeneSeqerSTD demonstrates that, when presented with a

Table 4. Accuracy of overall gene structure prediction in a non-GC control set

Program	0%	1%[a]	5%	10%	25%
Arabidopsis					
GeneSeqerGC	100	94.60 (1.50)	69.16 (3.91)	44.68 (3.79)	10.08 (1.74)
GeneSeqerSTD	100	94.68 (1.48)	69.80 (3.88)	46.12 (3.62)	11.04 (1.77)
Splign	100	96.16 (1.65)	66.92 (5.99)	25.84 (2.78)	0.00 (0.00)
BLAT	97	34.28 (5.42)	0.12 (0.22)	0.00 (0.00)	0.00 (0.00)
Sim4	97	91.36 (1.78)	64.20 (4.83)	34.80 (2.98)	0.00 (0.00)
Spidey	96	54.44 (4.20)	5.64 (1.71)	0.32 (0.44)	0.00 (0.00)
Rice					
GeneSeqerGC	99	89.12 (1.78)	60.16 (3.12)	36.88 (2.61)	5.12 (1.26)
GeneSeqerSTD	100	90.56 (1.96)	62.76 (2.92)	39.28 (2.82)	6.44 (1.54)
Splign	100	92.00 (2.47)	67.32 (4.15)	25.16 (2.84)	0.00 (0.00)
BLAT	97	21.00 (5.08)	0.24 (0.47)	0.00 (0.00)	0.00 (0.00)
Sim4	99	90.72 (2.53)	64.96 (4.00)	31.72 (2.89)	0.00 (0.00)
Spidey	97	46.48 (3.87)	5.80 (2.04)	0.12 (0.22)	0.00 (0.00)

[a]Percent simulated sequencing error induced on pseudotranscripts (see Systems and Methods section for details). The table values give the mean and standard deviation (in parentheses) of successful detection of entire gene structure in 100 test loci containing exclusively GT–AG introns for each of *Arabidopsis* and rice over 25 replicates.

Table 5. Accuracy of AT–AC intron prediction

Program	0%	1%[a]	5%	10%	25%
Arabidopsis					
GeneSeqerGC	100	95.20 (2.68)	82.72 (4.48)	67.20 (6.31)	30.72 (7.61)
GeneSeqerSTD	100	95.20 (2.68)	83.36 (4.55)	69.76 (5.83)	32.80 (7.18)
Splign	96	93.44 (2.54)	84.80 (3.43)	62.24 (6.81)	0.00 (0.00)
BLAT	92	85.60 (2.80)	67.20 (5.94)	34.88 (6.34)	0.96 (1.22)
Sim4	28	27.52 (2.33)	26.88 (2.86)	20.48 (5.13)	4.64 (0.93)
Spidey	48	41.28 (2.88)	27.20 (4.20)	11.52 (4.87)	0.48 (0.93)
Rice					
GeneSeqerGC	100	97.12 (2.38)	81.76 (6.42)	65.60 (7.23)	29.12 (6.02)
GeneSeqerSTD	100	96.96 (2.59)	81.12 (6.49)	66.40 (7.09)	31.36 (6.55)
Splign	96	97.28 (1.93)	83.52 (6.68)	66.08 (5.78)	0.00 (0.00)
BLAT	88	83.04 (2.83)	60.80 (5.94)	34.24 (5.83)	0.64 (1.05)
Sim4	12	12.96 (2.46)	16.48 (4.38)	18.24 (5.49)	3.84 (2.06)
Spidey	24	21.44 (1.79)	11.20 (3.43)	6.40 (3.33)	0.32 (0.77)

[a]Percent simulated sequencing error induced on pseudotranscripts (see Systems and Methods section for details). The table values give the mean and standard deviation (in parentheses) of successful detection of entire gene structure in 25 AT–AC intron-containing test loci for each of *Arabidopsis* and rice over 25 replicates.

simulated sequencing error challenge, explicit modeling of GC–AG introns does not generally prevent accurate annotation of GT–AG intron-containing gene structures. Better performance relative to the other programs reiterates the value of explicit splice site modeling. The poorer performance statistics for all programs on this rice set compared with the GC–AG intron containing set reported in Table 3 may be a sampling effect.

Spliced alignment detection of U12-type introns

To further probe the potential of gene structure prediction improvements by incorporation of models for non-canonical splice sites, we also assessed the ability of the spliced alignment programs to annotate AT–AC U12-type intron-containing gene structures with a test set of 25 genes (see Systems and Methods section). These spliced alignment assays were conducted similar to their GC–AG counterparts, with results presented in Tables 5 and 6 for AT–AC intron detection (competency at determining both donor and acceptor terminal dinucleotides) and complete gene structure resolution, respectively. Splign and the GeneSeqer variants clearly outperform the other programs. It should be noted that the slight performance discrepancy between GeneSeqerGC and GeneSeqerSTD are probably owing to noise generated by the GC–AG models, as all other aspects of the programs are identical. The GeneSeqer U12-donor prediction accuracy is more sensitive to sequencing errors than its accuracy for GC-donors, which is as expected because the U12-donor site scores are based on (near-)exact matching to the consensus U12 pattern (Brendel *et al*., 2004).

Ab initio gene structure and splice site prediction for GC–AG intron containing genes

We tested the *ab initio* gene structure prediction programs GENSCAN, GeneMark.hmm and FGENESH-GC for their ability to

Table 6. Accuracy of overall AT–AC intron-containing gene structure prediction

Program	0%	1%[a]	5%	10%	25%
Arabidopsis					
GeneSeqerGC	100	86.24 (4.99)	47.68 (5.23)	27.04 (6.58)	2.56 (2.54)
GeneSeqerSTD	100	86.08 (4.92)	48.16 (5.39)	27.52 (6.87)	2.56 (2.65)
Splign	96	84.16 (5.51)	53.44 (6.90)	16.16 (5.14)	0.00 (0.00)
BLAT	92	22.08 (5.31)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Sim4	24	21.12 (2.74)	14.08 (3.80)	6.08 (2.57)	0.00 (0.00)
Spidey	48	18.88 (5.39)	1.44 (1.59)	0.16 (0.56)	0.00 (0.00)
Rice					
GeneSeqerGC	100	85.28 (6.30)	47.84 (9.58)	20.32 (6.81)	1.76 (1.63)
GeneSeqerSTD	100	85.76 (6.96)	47.52 (9.57)	21.60 (6.10)	2.08 (1.64)
Splign	96	86.88 (5.14)	48.80 (9.39)	17.60 (4.98)	0.00 (0.00)
BLAT	88	27.84 (6.97)	0.16 (0.56)	0.00 (0.00)	0.00 (0.00)
Sim4	12	11.52 (3.16)	8.64 (4.48)	5.12 (3.38)	0.00 (0.00)
Spidey	24	10.88 (2.36)	1.92 (1.83)	0.48 (0.93)	0.00 (0.00)

[a] Percent simulated sequencing error induced on pseudotranscripts (see Systems and Methods section for details). The table values give the mean and standard deviation (in parentheses) of successful detection of entire gene structure in 25 AT–AC intron-containing test loci for each of *Arabidopsis* and rice over 25 replicates.

Table 7. Splice site prediction accuracy for GC-donors

		Arabidopsis				Rice			
	c	TP	FP	Sn	Sp	TP	FP	Sn	Sp
SplicePredictorSTD	6.0	96	281	0.96	0.25	81	579	0.81	0.12
	12.0	88	31	0.88	0.74	77	85	0.77	0.48
	15.0	71	4	0.71	0.95	74	41	0.74	0.64
SplicePredictorGC	6.0	95	147	0.95	0.39	82	456	0.82	0.15
	12.0	93	19	0.93	0.83	82	36	0.82	0.69
	15.0	86	9	0.86	0.91	82	10	0.82	0.89
NetGene2		80	7	0.80	0.92	31	4	0.31	0.89

Values are relative to the test sets of 100 genes of *Arabidopsis* and rice, each containing exactly one GC–AG intron (see Systems and Methods section). *c*, critical value for Bayesian splice site prediction (Brendel *et al*., 2004); TP, true positive; FP, false positive; Sn (sensitivity) = TP/100, and Sp (specificity) = TP / (TP + FP) (Burset and Guigó, 1996).

predict GC–AG introns on test sets of 100 genomic regions from *Arabidopsis* and rice. Each region contains a gene with four or more introns, exactly one of which is a cDNA-confirmed GC–AG intron (see Systems and Methods section). GENSCAN and GeneMark.hmm did not predict GC-donor sites, thus failing on all gene structures. FGENESH-GC predicted 54% of the *Arabidopsis* and 64% of the rice GC sites; however only 26% and 3% of the gene structures in the two sets were predicted correctly in their entirety.

Table 7 gives results of splice site prediction algorithms on the same GC–AG test dataset used in the spliced alignment assays. The new GC-donor site models incorporated into SplicePredictorGC significantly reduce the false positive prediction rate (increase specificity) in both *Arabidopsis* and rice compared with SplicePredictorSTD, which uses the strategy of treating each GC in the input sequence in the same way as the GTs. The true positive recovery rate (sensitivity) is about the same for both strategies. This result is consistent with the general conservation of the GT-donor site signal in GC-sites, but with stronger adherence to the consensus sequence as discussed above. Prediction accuracy is lower in rice compared with *Arabidopsis*, with the high false positive prediction rates in part reflecting the longer intron lengths in rice (see below). NetGene2 with default settings shows high specificity, at the cost of diminished sensitivity. The SplicePredictor programs give comparable results at a high threshold for the critical value *c* (Brendel *et al*., 2004). The poor performance of NetGene2 on rice relative to *Arabidopsis* is most probably explained by the monocot to dicot differences discussed earlier; as NetGene2 only offers *Arabidopsis* parameters, those were used also on the rice sequences. Note also for rice the dramatic drop in false positive predictions by SplicePredictorGC at higher *c*-values, without loss in sensitivity. This results from the fact that true GC-donors match well to a consensus signal and thus tend to score very high.

Characterization of GC–AG introns within their gene structure

We used the sets of *Arabidopsis* and rice genes with cDNA-confirmed GC–AG introns to probe for possible significant features of the underlying gene structures. In particular, we addressed the following questions in comparison with non-GC–AG intron containing genes. What is the average length and base composition of the GC–AG

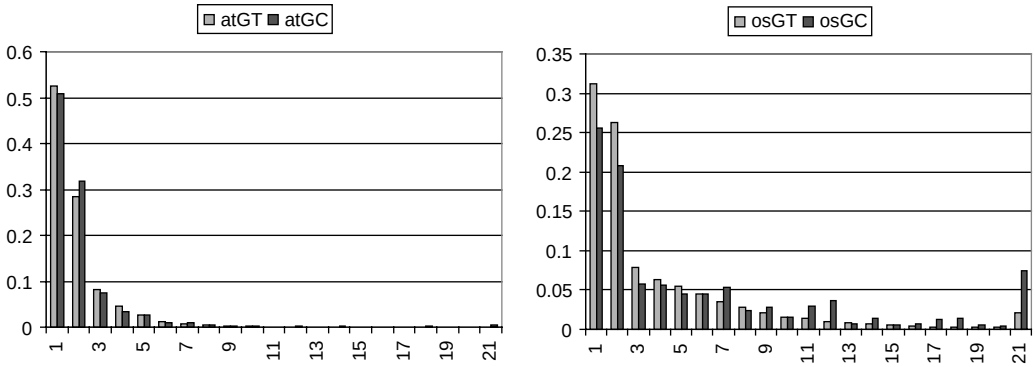

Fig. 4. Length distribution of *Arabidopsis* (at) and rice (os) introns with GT- and GC-donors. Size markers on the x-axis are in hundreds. Relative frequency is given on the y-axis.

introns? What is the average number of exons in these genes? Is there any preference of the GC–AG introns to occur 5′- or 3′-most?

The average length of the *Arabidopsis* GC–AG introns was found to be 168 nt, compared with 159 nt for GT–AG introns; for rice, the means are 691 and 386 nt. Figure 4 shows the length distributions in detail. In general, rice has a higher frequency of long introns compared with *Arabidopsis*. Of note is the particularly high proportion of long GC–AG introns in rice. The base composition of both classes of introns is similar within each species, although the rice GC–AG introns are slightly higher in G + C content. The *Arabidopsis* introns have a strong bias for U (40.3% in GT–AG introns, compared with 35.2% for rice), consistent with previous observations for dicots and monocots (Ko *et al*., 1998).

Of the 622 distinct annotated *Arabidopsis* genes containing GC–AG introns, 37 contained two GC–AG introns apiece, and only one contained three GC–AG introns (At3g10380, putatively encoding the exocyst complex component Sec8), the maximum number encountered in this study for a single gene. We identified 168 putative rice orthologs of the GC–AG intron containing *Arabidopsis* genes. Pairs of rice and *Arabidopsis* genes were considered orthologous if their translation products yielded reciprocally best BLASTP (Altschul *et al*., 1997) hits at a threshold of $E < 10^{-15}$ and no similar next-best hits. Eleven of these rice genes also had a GC–AG intron. LOC_Os01g68330, the rice ortholog for At1g70610, has two GC–AG introns. The genes are thought to encode chloroplast-associated ABC transporters. At1g70610 has only one GC–AG intron. For the better-annotated *Arabidopsis* genome, based on our data the fraction of genes with GC–AG introns is <2.5% of all genes. Thus, conservation of the GC–AG intron in 11/168 ortholog pairs is much more than expected by chance, suggesting these introns may have existed prior to the divergence of dicots and monocots.

The average exon count for the 622 *Arabidopsis* GC–AG intron-bearing genes is 12.18, compared with an average of 5.06 over all annotated *Arabidopsis* genes. Of the identified 168 rice orthologs of these genes, we determined, using the TIGR pseudomolecule version 3.0 annotation, that they contain on average 12.13 exons per gene, in comparison to an average of 5.89 exons per gene over all rice genes. Thus there seems to be a distinct bias for GC–AG introns to occur in genes with high exon count. Our analyses did not indicate that GC–AG genes exhibit any form of polar selectivity within gene structures, as their positions in their host genes were uniformly distributed (data not shown).

Classification of GC–AG intron containing genes

As shown above, mutation of the C in a GC-donor to T will result in a high-scoring GT-donor site that would be predicted to be an efficient splicing site. Thus, it is an intriguing question whether GC-donors merely represent tolerated mutations that are in equilibrium with GT-donors. Alternatively, present day GC-donors may be remnants of an evolutionary lineage, or there may be functional constraints acting on the extant genomes that maintain these sites. If the GC-donors represent tolerated mutations, then one would not expect a particular bias for their association with specific classes of genes. To test for such association, we derived the counts of GC–AG introns per *Arabidopsis* gene class as defined by GOslim terms defining cellular components (Berardini *et al*., 2004). The statistical significance of the observed counts was assessed by comparison with counts derived for randomly sampled genes containing GT–AG introns only.

The results (Fig. 5) show that GC–AG intron-containing *Arabidopsis* genes tend to be overrepresented in the chloroplast, mitochondria, nucleus, plasma membrane, other membranes, other cellular components and other cytoplasmic components and are less frequently encountered than expected in the cellular component unknown category. However, no clear bias emerged, and in view of the premature state of the ontology assignations no firm conclusions can be reached. The rice data, with ontology terms derived from their presumed *Arabidopsis* orthologs, gave a similar distribution (data not shown).

CONCLUSIONS

Current genome annotation reflects the limitations of computational tools available for the task. With increasingly available genome data, some of these limitations can be overcome by more specific training of the software. For *Arabidopsis*, our GeneSeqer spliced alignment tool detected an additional 115 cDNA-confirmed GC–AG introns over the currently annotated 661 cases. For rice, only 500 GC–AG introns were previously annotated, and this study identified an additional 1304 instances. In practice, gene structure is predicted from the consensus of multiple cDNA and EST alignments at a given locus. This allows accurate gene structure prediction from spliced alignment of heterogeneous transcripts, which is of great practical importance for plant genome annotation given the relatively small sets of available cDNAs and ESTs for any given species (Schlueter *et al*., 2003). The demonstrated robustness of the GeneSeqer spliced

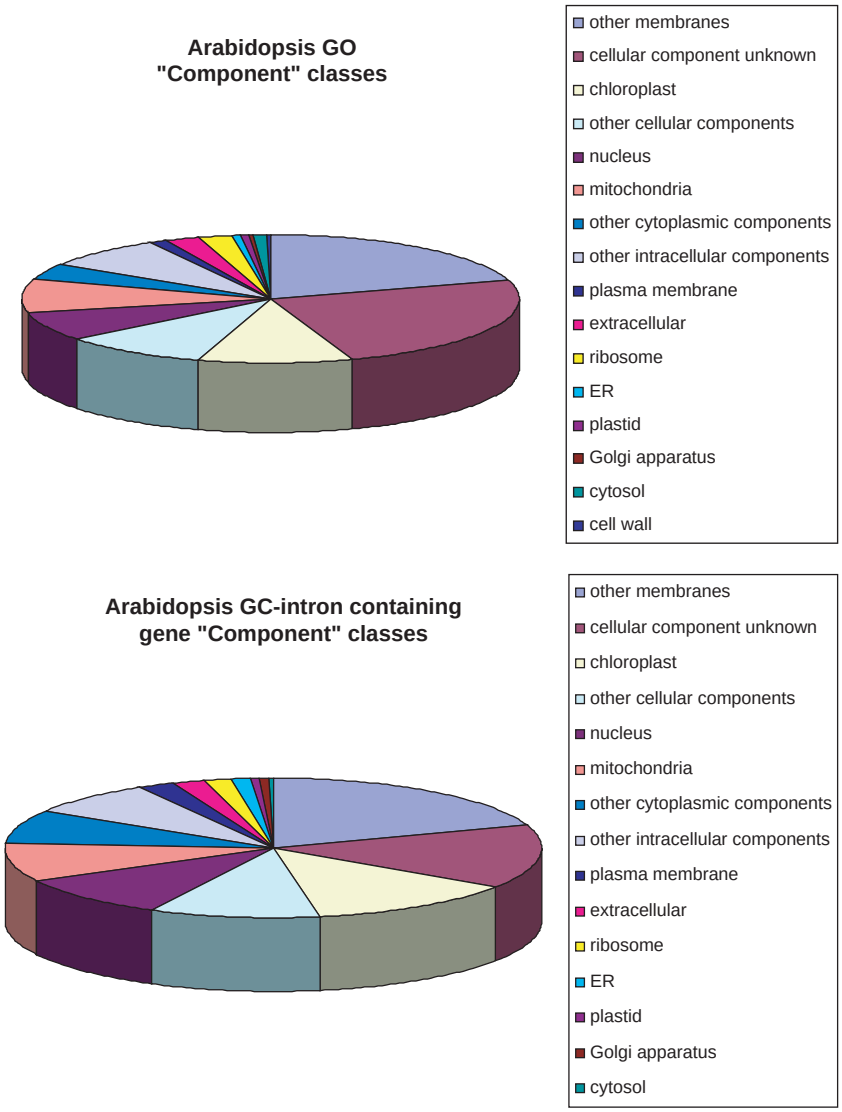

Fig. 5. Distribution of GC-intron containing *Arabidopsis* genes onto gene ontology 'Component' classes compared with the distribution for all genes.

alignments to sequence divergence is derived from the incorporation of species-specific splice site prediction in the scoring of alignments. The BSSM4GSQ software package should facilitate iterative training of updated and novel models for many species with emerging genomic and cDNA sequence data. This in turn will generate more reliable gene structure annotations for the study of the evolutionary origins and functional significance of non-canonical introns.

ACKNOWLEDGEMENTS

The authors are thankful to Valery Sagitov of Softberry, Inc. for allowing unrestricted use of their web services for this study. The authors are grateful to Wendy O. Sparks for a critical reading and editing of this manuscript. This work was supported in part by NSF grant DBI-0321600 and Research Grant No. IS-3454-03 from BARD, the United States–Israel Binational Agricultural Research and Development Fund. M.S. was also supported in part by the USDA with an IFAFS Multidisciplinary Graduate Education Training Grant (2001-52100-11506).

Conflict of Interest: none declared.

REFERENCES

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

Berardini,T.Z. *et al*. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.*, **135,** 745–755.

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**, 4748–4757.

Brendel,V. *et al*. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, **20**, 1157–1169.

Brunak,S. *et al.* (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.

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. *et al*. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.

Burge,C.B., Tuschl,T. and Sharp,P.A. (1999) Splicing of precursors to mRNAs by the spliceosomes. In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World—The Nature of Modern RNA Suggests a Prebiotic RNA*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 525–560.

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

Burset,M. *et al*. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

Dietrich,R.C. *et al*. (2001) Alternative splicing of U12-dependent introns *in vivo* responds to purine-rich enhancers. *RNA*, **7**, 1378–1388.

Felsenstein,J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Florea,L. *et al*. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Hebsgaard,S.M. *et al*. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.

Kapustin,Y., Souvorov,A. and Tatusova,T. (2004) Splign: a hybrid approach to spliced sequence alignments. *Proceedings of RECOMB 2004*, San Diego, CA.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kitamura-Abe,S. *et al*. (2004) Characterization of the splice sites in GT–AG and GC–AG introns in higher eukaryotes using full-length cDNAs. *J. Bioinf. Comp.Biol.*, **2**, 309–331.

Ko,C.H. *et al*. (1998) U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. *Plant Mol. Biol.*, **36**, 573–583.

Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.

Mathé,C. *et al*. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4013–4017.

Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.

Patel,A.A. *et al*. (2002) The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J.*, **21**, 3804–3815.

Pertea,M. *et al*. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.

Reddy,A.S.N. (2001) Nuclear pre-mRNA splicing in plants. *Critical Rev. Plant Sci.*, **20**, 523–571.

Rogan,P.K. and Schneider,T.D. (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.*, **6**, 74–76.

Ruvinsky,A. *et al*. (2005) Can codon usage bias explain intron phase distributions and exon symmetry? *J.Mol. Evol.*, **60**, 99–104.

Salzberg,S.L. (1997) A method for identifying splicing sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.

Schlueter,S.D. *et al*. (2003) GeneSeqer@PlantGDB—gene structure prediction in plant genomes. *Nucleic Acids Res.*, **31**, 3597–3600.

Schlueter,S.D. *et al*. (2005) Community-based gene structure annotation for the *Arabidopsis thaliana* genome. *Trends Plant Sci.*, **10**, 9–14.

Solovyev,V.V. (2001) Statistical approaches in eukaryotic gene prediction. In Balding, D. J. *et al*. (eds), *Handbook of Statistical Genetics*. John Wiley & Sons, Inc., New York, NY, pp. 83–127.

Thanaraj,T.A. and Clark,F. (2001) Human GC–AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.*, **29**, 2581–2593.

Usuka,J. *et al*. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

Wheelan,S.J. *et al*. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.

White,O. *et al*. (1992) Information contents and dinucleotide composition of plant intron sequences vary with evolutionary origin. *Plant Mol. Biol.*, **19**, 1057–1064.

Wortman,J.R. *et al*. (2003) Annotation of the *Arabidopsis* genome. *Plant Physiology*, **132**, 461–468.

Yuan,Q. *et al*. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.*, **31**, 229–233.

Zhu,W. and Brendel,V. (2003) Identification, characterization, and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.*, **31**, 4561–4572.