Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03

Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

II

In Silico Prediction of Plant Gene Function

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03
Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03

Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

6

Computer Software to Find Genes in Plant Genomic DNA

Ramana V. Davuluri and Michael Q. Zhang

Summary

Gene finding is the most important phase of genome annotation. Eukaryotic genomes contain thousands of protein coding genes, and computational gene prediction would rapidly increase the pace of experimental confirmation of expressed genes at the bench. The purpose of this chapter is to discuss the use of different computer programs that identify protein-coding genes in large genomic sequences. We describe most commonly used gene prediction programs that are available on the World Wide Web and demonstrate the use of some of these programs by an example. We provide a list of these programs along with their Web uniform resource locators (URLs) and suggest guidelines for successful gene finding.

Key Words

gene prediction, protein coding region, gene structure, splice sites, exons, computational gene finding

1. Introduction

The human *(1)* and *Arabidopsis* *(2)* genome projects and the advancement of sequencing technologies within the last decade are driving many other genome projects. The complete genome sequences of more than 800 organisms (many microbes, fungi, plants, and animals) are either complete or being sequenced (http://www.ncbi.nlm.nih.gov). One of the primary goals of any genome project is to provide a single continuous sequence for each of the chromosomes and demarcate the positions of all genes (**Fig. 1A**), along with the annotation of each component of a gene (**Fig. 1B**). Furthermore, recent advances in high-throughput technologies, such as genome-wide micro-array expression analysis, are starting to provide greater insights into the transcrip-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: Erich Grotewold © Humana Press, Inc., Totowa, NJ

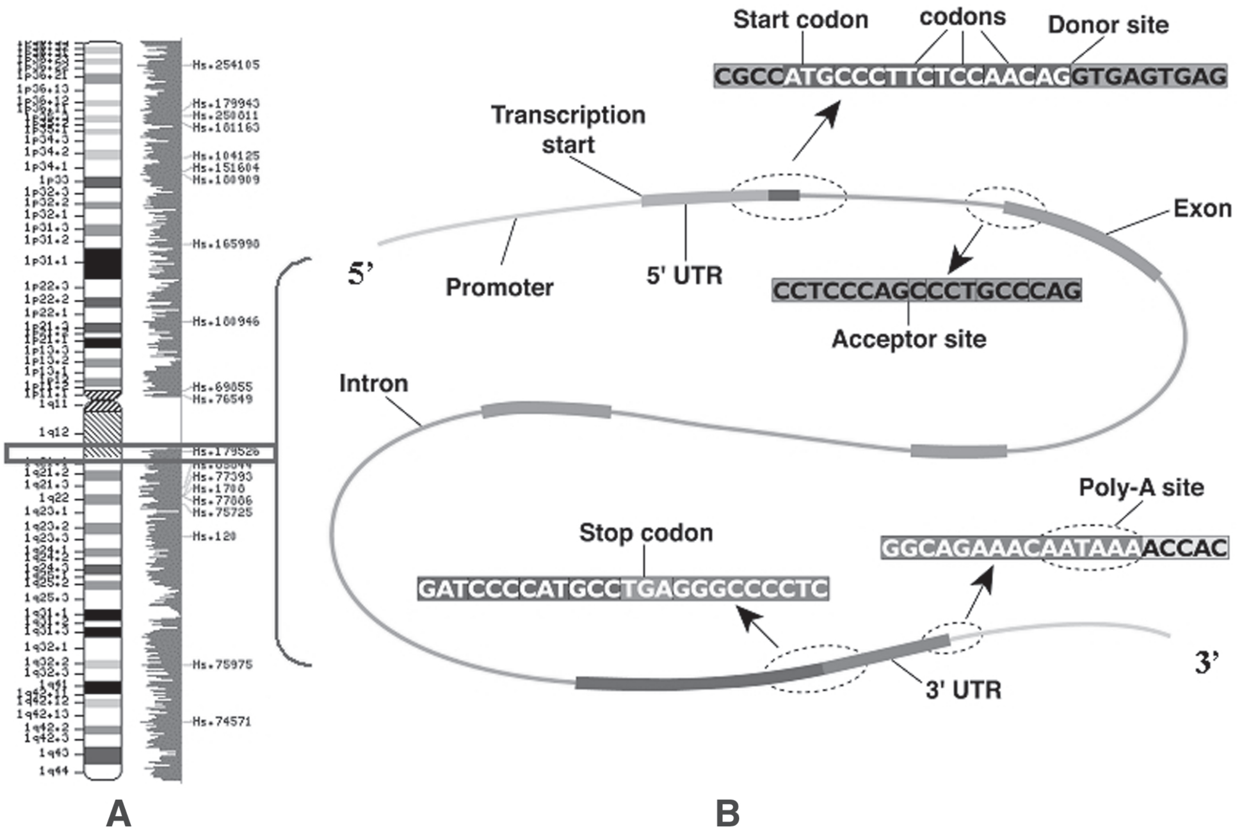

Fig. 1. Genome annotation. **(A)** Annotation of genes at chromosome level. **(B)** Annotation of individual components of a gene (such as exons, start codon, transcription start site, etc.).

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03
Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

tional regulation of eukaryotic cells ***(3–5)***. Integrating the genome sequence information (e.g., gene promoters) and micro-array expression data would provide an initial link to functional genomics. The identification and annotation of genes at genome level will contribute to the understanding of genome-wide gene expression studies. The major focus of this chapter is to introduce different bioinformatics tools that identify genes in genomic sequences.

Gene, defined as a transcribed unit, is usually split into pieces (called exons) that are separated by intervening sequences (called introns) in the eukaryotic genomes (**Fig. 1B**). The identification of genes by computational approaches is relatively straightforward for organisms with compact genomes (such as bacteria and yeast), because exons tend to be large, and the introns are either nonexistent or short. The challenge is much greater for larger genomes (such as those of rice or maize), because the exonic "signal" is buried under nongenic "noise." In the past few years, the accuracy and reliability of computational gene finding programs have improved to a reasonable extent, such that gene predictions within a genomic region can give valuable guidance to more detailed experimental studies. Computational sequence analysis methods, which detect genes in genomic DNA, can be broadly classified into two main categories: homology-based methods, and *ab initio* methods, which we discuss in **Subheading 3**.

2. Materials

User must have access to a computer with Internet access, e.g., a personal computer (PC) running Microsoft® Windows™ or Linux, an Apple® Macintosh®, or a UNIX® workstation. The user should be familiar with the use of Netscape Navigator or Microsoft Internet Explorer. The list of commonly used gene finding and sequence alignment programs and their Web uniform resource locators (URLs) are provided in **Table 1**.

3. Methods

3.1. Gene Prediction by Homology-Based Methods

Sequence homology is a very powerful type of evidence used to detect functional elements in genomic sequences. The homology-based methods to detect genes use either intraspecies or interspecies sequence comparison in at least four different ways, as summarized below.

3.1.1. Comparison with Expressed Sequence Tags/cDNA Database

A direct comparison of a genomic sequence (query) with expressed sequence tags (ESTs) or cDNA (**Fig. 2**) can identify regions of the query sequence that correspond to processed mRNA. BLASTN ***(6)*** is a common program that iden-

Table 1
Web URLs of Gene-Prediction and Sequence Alignment Programs

Program name	Model	Organism	Web URL
AAT	MZEF+homology		http://genome.cs.mtu.edu/aat.html
BCM Search Launcher	Many gene finding programs		http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
BLAST	Sequence alignment programs		http://www.ncbi.nih.gov/BLAST
CDS (search coding region)			http://bioweb.pasteur.fr/seqanal/interfaces/cds-simple.html
Fgenesh: (**Fgenes**; **Hexon**; **TSSW**; **TSSG**; **SPL**; **Polyah**)	HMM	dicots, monocots	http://genomic.sanger.ac.uk/gf/gf.shtml http://searchlauncher.bcm.tmc.edu:9331/seq-search/gene-search.html http://www.softberry.com/nucleo.html
GeneMachine	Integrated gene finder	*Arabidopsis*	http://genome.nhgri.nih.gov/genemachine/
GeneMark.hmm	HMM	*Arabidopsis*	http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
GeneParser	DP-ANN		http://beagle.colorado.edu/~eesnyder/GeneParser.html
GeneSplicer	Marko model and MDD	*Arabidopsis*, rice	http://www.tigr.org/tdb/GeneSplicer/gene_spl.html
GeneWise2	DNA protein alignment		http://www.cbil.upenn.edu/tess/
GenLang			http://www.cbil.upenn.edu/genlang/genlang_home.html
Genomescan	HMM+protein similarity	*Arabidopsis*, maize	http://genes.mit.edu/genomescan/
Genscan	HMM	*Arabidopsis*, maize	http://genes.mit.edu/GENSCAN.html
GRAIL	ANN	*Arabidopsis*	http://compbio.ornl.gov/tools/index.shtml

Program name	Model	Organism	Web URL
MORGAN	Decision tree, HMM		http://www.tigr.org/~salzberg/
VEIL			
GLIMMER			
MZEF	QDA	*Arabidopsis*	http://www.cshl.edu/mzhanglab/
MZEF SPC	MZEF+SpliceProximalCheck		http://industry.ebi.ac.uk/~thanaraj/MZEF-SPC.html
NetGene2	ANN	*Arabidopsis*	http://www.cbs.dtu.dk/services/NetGene2/
NNSplice	ANN	Drosophila, Human, or other	http://www.fruitfly.org/seq_tools/splice.html
OrfFinder			http://www.ncbi.nlm.nih.gov/gorf/gorf.html
PredictGenes			http://cbrg.inf.ethz.ch/subsection3_1_8.html
Procrustes	Spliced alignment		http://www-hto.usc.edu/software/procrustes/index.html
PROCRUSTES	Spliced alignment program		http://www-hto.usc.edu/software/procrustes
RepeatMasker	Identifies and masks interspersed repeats		http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
RiceHMM	HMM and EST similarity	Rice	http://rgp.dna.affrc.go.jp/RiceHMM
SGP-1	Similarity based gene prediction		http://soft.ice.mpg.de/sgp-1
SIM4	Spliced alignment program		http://pbil.univ-lyon1.fr/sim4.html
SplicePredictor	Logitlinear model	*Arabidopsis*, maize	http://bioinformatics.iastate.edu/cgi-bin/sp.cgi
WebGene		*Arabidopsis*	http://www.itba.mi.cnr.it/webgene/
Xpound			ftp://igs-server.cnrs-mrs.fr/pub/Banbury/xpound/
YeastGene			http://tubic.tju.edu.cn/cgi-bin/Yeastgene.cgi

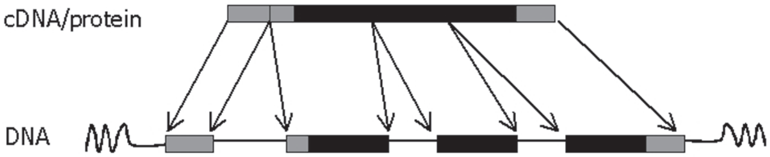

Fig. 2. Sequence alignment. Alignment of a cDNA or protein with a genomic sequence. In the cartoon showing the DNA, the rectangular boxes represent the exons, and the straight lines represent the introns.

tifies similar nucleotide sequences that exist in the databases (nr/EST) to the query sequence (*see* **Note 2**). BLASTN algorithm finds similar sequences by generating an indexed table or dictionary of short subsequences called words for both the query and the database (*see* Basic Local Alignment Search Tool [BLAST] help at [http://www.ncbi.nlm.nih.gov/BLAST] for further details). For identification of gene regions in the query sequence, choose low complexity repeat filter and select expected value as 0.1. If the query sequence is very long MegaBLAST is a better choice, as it is specifically designed to efficiently find long alignments between very similar sequences. MegaBLAST is also optimized for aligning sequences that differ slightly as a result of sequencing errors. The user can select different options. We suggest the use of expected value (e-value) of 0.1 and choose filter for low complexity repeats. When larger word size is used (default value is 28), it increases the search speed and limits the number of database hits. For BLASTN, the word size can be reduced from the default value of 11 to a minimum of 7 to increase sensitivity.

AU: pls. callout Note 1 before Note 2 or renumber

BLASTN is mainly used to pull out similar sequences from the database, and most of the times it is hard to interpret the exon boundaries. After finding a cDNA or EST match to the query sequence, one can use spliced alignment programs such as SIM4 ***(7)***, which efficiently aligns an EST or cDNA with the genomic sequence. RiceHMM ***(8)*** is another program that predicts gene domains in rice genome sequence, based on a hidden Markov model using a database of rice ESTs, composed of nearly 15,000 cDNAs.

3.1.2. Comparison with Protein Sequence Databases

Comparison of genomic sequence with protein sequence database by programs, such as BLASTX, can identify probable protein coding regions. Subsequently, spliced alignment programs such as Genewise ***(9)***, GeneSeqer ***(10)***, or PROCRUSTES ***(11)*** can be used to find the gene structure by comparing the genomic DNA sequence to the target protein sequences. These programs derive an optimal alignment based on sequence similarity score of the predicted gene product to the protein sequence and intrinsic splice site strength of the predicted introns.

3.1.3. Comparison of a Translated Genomic Sequence with Translated Nucleotide Database

A comparison of a translated genomic sequence with nucleotide database, which has been translated in all six reading frames, using TBLASTX can identify similarities among protein coding regions. TBLASTX can be run by selecting "Nucleotide query—Translated db [tblastx]" option from the BLAST Web page. TBLASTX takes a nucleotide query sequence, translates it in all six frames, and compares the translations to a nucleotide database (e.g., nr, est, est_human, est_others, etc.) sequences that are dynamically translated in all six frames.

3.1.4. Comparison of Genomic Sequence with Homologous Genomic Sequences from Related Species

Protein coding DNA from closely related plant species, such as sorghum and maize, show considerable sequence similarity ***(12)***. With the availability of genomes of many different organisms, comparative genomic approaches are gaining importance. VISTA/AVID ***(13)*** and PipMaker ***(14)*** can be used to compare large genomic sequences to find orthologous genomic sequences from closely related species. For example, sequence analysis of orthologous genes from rice, maize, and sorghum showed that the exons are more conserved than introns ***(12)***. The degree of sequence conservation, in terms of sequence identity, across species has been shown to be consistent with the divergence times of the respective species. The rice genes are considerably more diverged than their counterparts in maize and sorghum. For gene prediction programs, it would be best to compare two genomes that are very closely related, but distant enough that their intergenic repeat elements differ significantly. As a rule of thumb, consider two species as closely related, if those two are diverged within the last 25 million yr. For example, maize and sorghum are closely related species as they were diverged 15–20 million yr ago. If homologous genomic sequences from two species are known, then a recently developed gene prediction tool called SGP-1 ***(15)*** can be used to find protein-coding genes.

3.2. Gene Prediction by Ab Initio *Methods*

Homology-based methods provide useful information about gene locations as well as clues about gene function. Similarity-based methods, such as BLAST, combined with more sophisticated spliced alignment methods, such as SIM4, can give most reliable gene structure, provided there exists a full-length cDNA sequence in the database. However, most of the cDNA or EST sequences are partial, and these databases are increasing rather slowly. To help overcome these limitations, several *ab initio* gene finding programs have been developed over the years (**Table 1**). These programs recognize signals or compositional features in an input genomic sequence by pattern matching or statistical methods. The performance of a gene finding program is typically measured in terms of the sensitivity*,* defined as the proportion of true signals (e.g., donor signals, exons) that are correctly predicted, and specificity*,* defined as the proportion of predicted signals that are correct. A program is considered accurate if its sensitivity and specificity are simultaneously high. We describe some of the most commonly used gene prediction programs trained for plant genomes. A comprehensive review of these programs can be found at Weintian Li's Bibliography on Computational Gene Recognition Web site (http://linkage.rockefeller.edu/wli/gene/). A recent review by Lincoln Stein ***(16)*** surveys the various ways the genome annotation is being carried out.

3.2.1. Splice Site Prediction Programs

Since most vertebrate, invertebrate, and plant genes have several exons; precise gene structure prediction in these organisms very much depends on the ability of splice site prediction. Many first generation gene prediction programs used simple position weight matrix methods to model the compositional biases present in the 5' and 3' splice sites. Most recent programs have investigated the correlations between different positions by using Markov models, maximal dependence decomposition models, decision tree models, and artificial neural networks. GeneSplicer, Netplantgene, Netgene2, and SplicePredictor are some of the splice site prediction programs that use splice site models. The specificity of these programs is just around 35% at a 50% sensitivity threshold in large genomic sequences ***(17)***. This is because the selection of splice sites not only depends on the strength of the splice sites but also on other factors, such as exonic and intronic enhancer signals located some distance from splice junctions ***(18)***. To get an initial assessment of potential splice sites we recommend the use of GeneSplicer ***(19)***, SplicePredictor ***(20)***, or NetGene2 ***(21)***.

3.2.2. Exon Prediction Programs

Most of the gene prediction programs have been trained to predict protein coding exons; exons corresponding to the region from translation initiation codon (ATG) to stop codon (TAA/TAG/TGA). The protein coding exons typically are of four types: *(i)* initial exons (ATG to first donor site); *(ii)* internal exons (acceptor site to donor site); *(iii)* terminal exons (acceptor site to stop codon); and *(iv)* single exons (ATG to stop codon without introns). The accuracy of splice site prediction, and hence exon prediction, by second generation programs (e.g., Genscan ***[22]***, GeneMark.hmm ***[23]***, MZEF ***[24]***, or SPL ***[25]***) is significantly higher than simple splice site prediction programs, because these programs integrate splice site models with additional types of information, such as compositional features of exons and introns. MZEF, based on quadratic discriminant analysis, was specifically trained to predict internal exons. It was shown ***(25)*** to perform better than FGENESP, GRAIL, Genscan, and GeneMark.hmm in predicting internal exons for *Arabidopsis* genome. For predicting initial and terminal exons, Genscan and GeneMark.hmm are the best options, even though the accuracy of predicting these exons is significantly lower than that of internal exon prediction.

3.2.3. Gene Modeling Programs

The accuracy of individual exon prediction further increases by combining the reading frame compatibility of adjacent exons to make a full coding transcript. Probabilistic models, such as Hidden Markov models, have been used to incorporate this information in Genscan and GeneMark.hmm, which model different states (exon, intron, intergenic region, etc.) of a gene. In gene modeling and predicting multiple genes in large genomic contigs, Genscan and GeneMark.hmm were shown to give comparable results and by far the best available programs for plant genomes ***(25)***.

3.3. Gene Prediction by Integrated Methods

Gene prediction by homology-based methods is perhaps the most efficient way of finding genes in genomic sequences, since the evidence of support (mRNA, EST, protein) was already derived experimentally. On the other hand, *ab initio* gene-prediction programs miss some known genes (false negatives) and predict some that are not real (false positives). Traditionally, *ab initio* gene prediction programs and homology-based approaches were used independently and combined later manually by an annotator. This process has been automated in recent programs, such as Genomescan ***(27)*** and RiceGAAS ***(8)*** that combine gene predictions with similarity comparisons to produce more reliable predictions of protein-coding regions. GenomeScan incorporates protein homology information (BLASTX hits) with the exon–intron predictions of Genscan. The input to this program consists of a genomic sequence, a selection of appropriate organism (from vertebrate, *Arabidopsis*, and maize), and a set of protein sequences (in fasta format), which may be similar to the genomic sequence. GenomeScan first masks the interspersed repetitive elements in the genomic sequence with RepeatMasker and then combines the Genscan predicted peptides with BLASTX hits. The program determines the most likely "parse" (gene structure), conditional on the given similarity information under a probabilistic model of the gene structural and compositional properties of genomic DNA for the given organism.

AU: pls. cite ref 26 in text between 25 & 27

RiceGAAS runs Genscan (with *Arabidopsis*, maize models), RiceHMM, MZEF (with *Arabidopsis*, model), and SplicePredictor (with *Arabidopsis*, maize models) programs and combines these predictions with BLASTN (against MAFFRICE database) and BLASTX (against nr database) homology comparisons. It also masks the repeats of *Arabidopsis thaliana* repeats by using RepeatMasker program. For RiceGAAS, the input is the genome sequence to be analyzed, which can be pasted in a window or uploaded from a file (as fasta format).

3.4. Worked Example

We discussed various gene-finding strategies in the previous sections. Now let us discuss which programs to choose and how to use those programs in a real practical scenario. Given a large genomic sequence, we suggest the following steps in arriving at probable exons that the sequence may contain.

1. Blast the sequence against nr and EST databases by using BLASTN (Megablast in case of very long sequence) program. Note the list of accession numbers of cDNAs or ESTs with "% identity" score ≥99, from the blast output.
2. Use SIM4 program to align each of the cDNA/ESTs with the genomic sequence so as to identify exons with canonical splice sites.
3. Blast the sequence against nr database by using the BLASTX program. From the output, note down the BLASTX matches that may belong to genes.
4. Submit the sequence to at least 4 different gene prediction programs and select the consensus predictions (exons). We consider a prediction as consensus prediction if it is predicted by at least half of the programs either fully (both ends of the predicted exons are same) or partially (there exists an overlapping region among the predicted exons).

To demonstrate the above steps, we use the genomic sequence in rice bacterial artificial chromosome (BAC) in GenBank® with Accession no. AP005190, which has not yet been annotated at the time writing of this chapter. Since the length of the sequence is very large (138,893 bp), we used Megablast to identify the homologous sequences from the GenBank. The program was run twice, each time by choosing nr and EST databases. **Table 2** gives the list of high scoring segment pairs (HSPs) from the Megablast output. As BLAST is mainly a sequence similarity program, it helps us to identify the regions in the input sequence (query sequence) that are similar to known sequences (subject sequences) in the database. As the output suggests, it is hard to interpret the gene structure (exon–intron boundaries) from the output. Hence, we ran SIM4 program to align each of the EST/cDNA sequences (from the output of Megablast) with the genomic sequence AP005190. **Table 3** gives the list of exons inferred by combining various EST/cDNA alignments with AP005190 using SIM4.

Table 2
List of HSPs of AP005190 (Query) against EST Database from Megablast Output

Subject ID	% Identity	Alignment length	Mismatches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
AU173904	100	375	0	0	47222	47596	87	461	0	743.9
AU173904	100	87	0	0	46592	46678	1	87	6.70E-38	173
AU173465	100	363	0	0	24137	24499	433	71	0	720.1
AU173465	100	72	0	0	25919	25990	72	1	6.00E-29	143.2
AU031146	100	313	0	0	14463	14775	138	450	9.00E-173	621
AU093845	99.4	317	1	1	14463	14778	381	697	2.00E-170	613
AU093845	100	116	0	0	13601	13716	266	381	3.30E-55	230.4
AU093845	100	75	0	0	12946	13020	194	268	9.70E-31	149.2
AU093845	100	74	0	0	12788	12861	125	198	3.80E-30	147.2
C97606	99.7	313	0	1	14463	14775	527	838	9.00E-170	611.1
C97606	100	116	0	0	13601	13716	412	527	3.30E-55	230.4
C97606	100	75	0	0	12946	13020	340	414	9.70E-31	149.2
C97606	100	74	0	0	12788	12861	271	344	3.80E-30	147.2
C73253	99.3	286	1	1	42747	43031	425	140	7.00E-152	551.6
C73253	100	142	0	0	43214	43355	142	1	1.00E-70	282
BI798584	100	267	0	0	14463	14729	252	518	3.00E-145	529.8
BI798584	99.1	116	1	0	13601	13716	137	252	8.00E-53	222.5
BF430535	100	259	0	0	105549	105807	35	293	2.00E-140	513.9
BF430535	100	112	0	0	106534	106645	473	584	8.00E-53	222.5
BF430535	100	99	0	0	106759	106857	585	683	4.60E-45	196.7
BF430535	100	65	0	0	106228	106292	354	418	9.00E-25	129.3
BF430535	100	64	0	0	106041	106104	291	354	3.50E-24	127.4
BF430535	100	60	0	0	106373	106432	414	473	8.60E-22	119.4

Table 2
Continued

Subject ID	% Identity	Alignment length	Mismatches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
D40524	99.6	235	1	0	82675	82909	235	1	8.00E-124	458.4
D40946	99.6	230	1	0	82680	82909	230	1	2.00E-121	450.5
AU090572	99.1	231	2	0	53526	53756	78	308	5.00E-119	442.6
AU163696	100	163	0	0	120870	121032	1	163	3.00E-83	323.6
AU163696	100	125	0	0	121227	121351	161	285	1.40E-60	248.3
AU183284	100	133	0	0	12464	12596	315	447	2.40E-65	264.1
AU183284	100	120	0	0	11103	11222	195	314	1.40E-57	238.4
AU183284	100	54	0	0	9806	9859	142	195	3.30E-18	107.5
AU093296	99.2	120	0	1	11103	11222	236	354	1.30E-54	228.5
AU093296	100	70	0	0	9591	9660	117	186	9.30E-28	139.3
AU093296	100	54	0	0	9806	9859	183	236	3.30E-18	107.5
AU173536	100	112	0	0	82229	82340	112	1	8.00E-53	222.5
BQ281772	100	108	0	0	120925	121032	72	179	2.00E-50	214.6
BE599115	100	108	0	0	120925	121032	85	192	2.00E-50	214.6
BE593685	100	108	0	0	120925	121032	76	183	2.00E-50	214.6
AW680979	100	108	0	0	120925	121032	63	170	2.00E-50	214.6
BG560418	99.1	108	1	0	120925	121032	85	192	4.80E-48	206.7
AU166259	100	84	0	0	29212	29295	356	439	4.10E-36	167
AU166259	100	38	0	0	28319	28356	322	359	1.20E-08	75.82
BI813425	100	79	0	0	83259	83337	466	388	4.00E-33	157.1
BM347731	100	77	0	0	120956	121032	736	660	6.20E-32	153.1
BM079469	100	77	0	0	120956	121032	615	539	6.20E-32	153.1
BI813794	100	77	0	0	83261	83337	476	400	6.20E-32	153.1
D39271	100	77	0	0	27502	27578	185	109	6.20E-32	153.1

Table 2
Continued

Subject ID	% Identity	Alignment length	Mis matches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
BI245296	100	69	0	0	120964	121032	481	413	3.70E-27	137.3
BI813113	100	64	0	0	83274	83337	549	486	3.50E-24	127.4
BE643512	100	64	0	0	120969	121032	1	64	3.50E-24	127.4
AU082326	100	63	0	0	133964	134026	69	131	1.40E-23	125.4
BE593268	100	60	0	0	120973	121032	1	60	8.60E-22	119.4
BJ450012	100	59	0	0	48886	48944	9	67	3.40E-21	117.5
BQ667839	100	49	0	0	120984	121032	393	345	3.20E-15	97.63
BF292448	100	44	0	0	120986	121029	1	44	3.00E-12	87.72
BF145477	100	44	0	0	120986	121029	1	44	3.00E-12	87.72
BM368889	100	43	0	0	120987	121029	1	43	1.20E-11	85.73
BE639720	100	43	0	0	120990	121032	1	43	1.20E-11	85.73
BE426858	100	43	0	0	120987	121029	1	43	1.20E-11	85.73
BQ608952	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606868	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606799	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606785	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BJ321890	100	41	0	0	120989	121029	809	769	1.90E-10	81.77
BJ210114	100	41	0	0	120989	121029	62	102	1.90E-10	81.77
BI125789	100	41	0	0	120992	121032	187	227	1.90E-10	81.77
BG313503	100	41	0	0	120989	121029	24	64	1.90E-10	81.77

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03

Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

Table 3
List of Exons Derived from the Alignments of EST/cDNAs with AP005190 by Using SIM4

Gene no.	Exon no.	Strand	Exon begin—exon end	Supported EST/cDNA
1	1	+	*9475–9658	AU093296, AU183284
	2	+	9808–9859	AU093296, AU183284
	3	+	11104–11222	AU093296, AU183284
	4	+	12464–12704	AU093296, AU183284, AU093845, C97606
	5	+	12790–12857	AU093845, C97606, BI798584
	6	+	12947–13019	AU093845, C97606, AU031146, BI798584, AY072931
	7	+	13603–13715	AU093845, C97606, AU031146, BI798584, AY072931
	8	+	14463–14778*	AU093845, C97606, AU031146, BI798584, AY072931
2	2	–	24499–24137	AU173465
	1	–	25990–25921	AU173465
3	3	–	27370–27214	D39271
	2	–	27577–27502	D39271
	1	–	27787–27678	D39271
3	1	+	27998–28354	AU166259
	2	+	29214–29295	AU166259
5	2	–	43029–42747	C73253
	1	–	43355–43215	C73253
6	1	+	46592–46677	AU173904
	2	+	47222–47596*	AU173904
	3	+	*48878–48950	BJ450012
	4	+	49354–49793*	BJ450012
7	1	+	*53449–53756*	AU090572
8	1	+	*81944–82003	AU173536
	2	+	82219–82340*	AU173536
	3	+	*82407–82909	D40524, D40946
	4	+	83253–83711	BI813425, BI813794
9	7	–	90106–90089*	BF430535
	6	–	105804–105550	BF430535
	5	–	106103–106041	BF430535
	4	–	106290–106228	BF430535
	3	–	106432–106376	BF430535
	2	–	106645–106535	BF430535
	1	–	*106857–106759	BF430535
10	1	+	*120870–121031	AU163696, BQ281772, BG560418
	2	+	121229–121436	AU163696, BQ281772, BG560418

Table 3
Continued

Gene no.	Exon no.	Strand	Exon begin—exon end	Supported EST/cDNA
	3	+	121560–121626	BQ281772, BG560418
	4	+	122609–122625*	BQ281772, BG560418
11	1	+	*133895–134146	AU082326
	2	+	134200–134215*	AU082326

*Might be an incomplete exon due to partial EST/cDNA.

Table 4
List of HSPs of AP005190 (Query) against nr Database from BLASTX Output

Subject ID	% Identity	Alignment length	Subject start	Subject end	Query start	Query end	E-value	Bit score
AAC19401	27%	212	225	376	16622	15987	4e-24	189
AAC19401	42%	69	371	439	15925	15719	4e-24	62.8
AAC19401	41%	51	66	116	18173	18021	0.11	44.7
AAC19401	38%	39	155	193	17145	17029	0.11	42.4
AAB17501	30%	213	223	377	16625	15987	2e-25	88.6
AAB17501	38%	70	372	441	15925	15716	2e-25	55.8
AAB17501	42%	50	66	115	18170	18021	1e-06	47.0
AAB17501	37%	37	122	158	17318	17208	7e-05	40.0
AAB17501	30%	36	157	192	17136	17029	7e-05	34.7
AAB17501	41%	31	32	62	18360	18268	1e-06	33.5
AAD27547	97%	1520	1	1520	62266	66825	0	2915
AAM08795	98%	1520	265	1784	62266	66825	0	2942
AAM08795	98%	203	1	203	61125	61733	1e-113	414
AAK92543	97%	1520	194	1713	62266	66825	0	2929
AAK92543	97%	140	1	140	61314	61733	7e-73	281
BAB86564	98%	1100	1	1100	86635	83336	0	2175
AAD19359	32%	1065	832	1876	119222	116118	1e-129	466

Next, we ran "Nucleotide query—Protein db [BLASTX]" program. Select "TRANSLATED query—PROTEIN database [BLASTX]" for Choose a translation options and nr for database options. Since the sequence is very long, we submitted the sequence as three pieces (1–50 K, 50–100 K, and 100 K to rest) to save running time, which was done by entering corresponding values of each subsequence in "from" and "to" windows of Set subsequence options. The rest of the values were left as default. **Table 4** gives the list of HSPs from

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03
Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

Table 5
List of Consensus Exons Predicted by at Least Two Gene-Pprediction Programs in the Genomic Sequence with Accession No. AP005190

Strand	Type	Ex. Begin– Ex. End	Programs predicted
+	Intr	370–459	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Intr	668–712	Genscan (A), GeneMark.hmm (M)
+	Intr	802–872	Genscan (A), GeneMark.hmm (M)
+	Intr	1501–1633	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	1945–2033	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Term	4279–4049	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Init	5382–5320	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Init	8153–8162	Genscan (A), Genscan (M)
+	Intr	9743–9859	Genscan (A), Genscan (M)
+	Intr	12464–12704	Genscan (A), GeneMark.hmm (M)
+	Intr	12790–12857	Genscan (A), GeneMark.hmm (M)
+	Intr	12947–13019	GeneMark.hmm (M), Mzef (A)
+	Intr	13603–13715	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Term	14463–14615	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	15500–15279	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	15912–15632	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	16226–16112	Genscan (A), GeneMark.hmm (M)
–	Intr	16634–16347	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	16829–16779	Genscan (A), GeneMark.hmm (M)
–	Intr	18173–18003	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	20200–19268	Genscan (A), Genscan (M)
–	Term	24499–24380	Genscan (A), GeneMark.hmm (M)
–	Intr	25684–25613	Genscan (A), GeneMark.hmm (M)
–	Intr	25997–25921	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	27571–27141	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	29214–29427	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	30478–30644	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	31529–31653	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	32807–32902	Genscan (A), GeneMark.hmm (M)
+	Intr	32961–33009	GeneMark.hmm (M), Mzef (A)
+	Intr	33144–33198	Genscan (A), GeneMark.hmm (M)
+	Intr	39059–39180	Genscan (A), Genscan (M)
+	Term	41035–41106	Genscan (A), Genscan (M)
+	Init	43393–43699	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	44245–44360	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	44447–44535	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)

Table 5
Continued

Strand	Type	Ex. Begin– Ex. End	Programs predicted
+	Intr	45293–45338	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	46050–46218	Genscan (A), Mzef (A)
+	Intr	46595–46677	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	47222–47602	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	48259–48950	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	49354–49909	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	50151–50468	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Term	50751–50795	Genscan (M), GeneMark.hmm (M)
–	Term	53795–53682	Genscan (A), GeneMark.hmm (M)
–	Intr	53973–53875	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	54140–54068	Genscan (A), GeneMark.hmm (M)
–	Intr	54335–54225	Genscan (A), GeneMark.hmm (M)
–	Intr	54605–54432	Genscan (A), GeneMark.hmm (M)
–	Intr	55400–54715	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	55547–55402	Genscan (A), Genscan (M)
–	Intr	55814–55673	Genscan (A), Genscan (M)
–	Intr	57329–55889	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
–	Init	58233–57914	Genscan (A), Genscan (M)
+	Init	60906–60917	Genscan (A), Genscan (M)
+	Intr	61125–61718	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	62266–66693	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	67890–67955	Genscan (A), GeneMark.hmm (M)
+	Intr	68046–68188	Genscan (A), GeneMark.hmm (M)
+	Intr	69099–69391	Genscan (A), GeneMark.hmm (M)
+	Intr	72191–73594	Genscan (A), GeneMark.hmm (M)
+	Term	73703–73858	Genscan (A), GeneMark.hmm (M)
–	Intr	82264–82166	Genscan (A), Genscan (M)
–	Intr	86635–83343	Genscan (A), Genscan (M)
+	Init	94228–94246	Genscan (A), Mzef (A)
–	Sngl	98915–97443	Genscan (A), Genscan (M)
+	Intr	103554–103766	Genscan (A), Genscan (M)
–	Intr	10103–106041	GeneMark.hmm (M), Mzef (A)
–	Intr	106290–106228	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	106432–106376	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	106645–106535	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Init	107034–106759	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	112457–112600	Genscan (A), GeneMark.hmm (M)

Table 5
Continued

Strand	Type	Ex. Begin– Ex. End	Programs predicted
+	Intr	112696–113452	Genscan (A), GeneMark.hmm (M)
+	Intr	113495–114083	Genscan (A), GeneMark.hmm (M)
+	Intr	114248–114667	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Intr	114743–114802	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Term	115053–115739	Genscan (A) GeneMark.hmm (M)
–	Term	118976–116094	Genscan (A), GeneMark.hmm (M)
–	Init	119460–119294	Genscan (A), GeneMark.hmm (M)
+	Init	120929–121031	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	121229–121436	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Term	121560–121680	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Term	126660–126599	Genscan (A), GeneMark.hmm (M)
–	Intr	126961–126811	Genscan (A), GeneMark.hmm (M)
–	Init	127447–127307	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Init	129895–131341	Genscan (A), GeneMark.hmm (M)
+	Intr	132275–132331	Genscan (A), Mzef (A)
+	Intr	133577–133610	Genscan (A), Genscan (M)

In the column headings: type stands for type of exon; *Init, Intr,* and *Term* stands for *Initial, Internal, and terminal* exons, respectively, and ex. stands for exon.

BLASTX output. The values in columns query start and query end would give the regions in the genomic sequence AP005190 that may belong to probable genes.

Finally, we submitted the genomic sequence AP005190 to four gene-finding programs Genscan with *Arabidopsis* model, Genscan with Maize model, GeneMark.hmm with rice model, and MZEF with *Arabidopsis* model. Default values were selected for other parameters for each of the programs used. As none of the programs is good enough to predict the complete gene structure, we considered only the exon predictions. We compiled the list of all consensus exons that were predicted by at least two programs. We consider an exon as a consensus prediction if there exists an overlapping region among the predictions of at least two different programs. **Table 5** gives the list of all such exons.

AU: Notes 1, 3, 4, & 5 Pls provide callouts in text

AU: Pls provide fll reference & add to ref list

4. Notes

1. Despite great progress, gene prediction by computational approaches alone is still far from perfect. The existing programs have reached a reasonable sophisti-

cation in identifying >90% of the nucleotides in a given genome as coding or noncoding (Stormo, 2000). We suggest using computational tools to identify a nucleotide as either coding or noncoding. But, identifying the exact boundaries of all the exons and assembly of the exons into different genes might be much harder and is not possible by computational approaches alone. However, even the partial predictions are of immense value to design the experiments that can determine the complete gene structure faster than would be possible by experimental methods alone.

2. Similarity-based methods (e.g., BLASTN, BLASTX) are perhaps the best to determine a given region of the genome is transcribed or not. A BLASTN match to a cDNA/EST or BLASTX match to a protein is good evidence that the region belongs to a gene. However, these methods have their own limitations. Most of the cDNAs or ESTs are incomplete and may contain one or more introns, which could lead to misclassification of intron region as exon. Some cDNA sequences may contain repetitive elements that will cause false genomic matches. Protein databases may contain potentially incorrect predicted proteins. BLASTX matches to predicted protein sequences should be avoided. Partial BLASTX alignment to a target protein should not be considered, as the protein may not be a true ortholog of the source gene and only shares some domains. We should note that the similarity data (cDNA/EST data) is never complete. Even the most comprehensive cDNA projects will miss low copy number transcripts and those transcripts whose expression is low, cell- or tissue-specific, or expressed only under unusual conditions.
3. Almost all gene finding programs can predict only protein coding regions and have not been trained to predict untranslated exons and untranslated portion of first and last coding exons.
4. Before running any gene-finding program, we suggest the use of programs such as RepeatMasker, which identifies known classes of interspersed repeats, and LINEs and SINEs, which exist in noncoding regions of the genome.
5. Most of the gene finding programs are based on statistical pattern recognition methods that require a training data. This makes the program organism-specific depending on the training data. So, while running a gene prediction program, select the organism of the genomic sequence. If the program was not trained on the organism of your choice, select the most closely related one. If the genome of your choice does not exist and has low gene density, then there may be more false positive predictions by choosing another genome with high gene density.

References

1. Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
2. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.
3. Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728.

4. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297.
5. Finkelstei, D., Ewing, R., Gollub, J., Sterky, F., Cherry, J. M., and Somerville, S. (2002) Microarray data quality analysis: lessons from the AFGC project. *Arabidopsis* Functional Genomics Consortium. *Plant Mol. Biol.* **48,** 119–131.
6. Altschul, S. F.,Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
7. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8,** 967–974.
8. Sakata, K., Nagamura, Y., Numa, H., et al.. (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30,** 98–102.
9. Birney, E. and Durbin, R. (2000) Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10,** 547–548.
10. Usuka, J., Zhu, W., and Brendel, V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16,** 203–211.
11. Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*, **93,** 9061–9066.
12. Schmidt, R. (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. *Plant Mol. Biol.* **48,** 21–37.
13. Mayor, C., Brudno, M., Schwartz, J. R., et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047.
14. Schwartz, S., Zhang, Z., Frazer, K. A., et al. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10,** 577–586.
15. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigo, R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11,** 1574–1583.
16. Stein, L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2,** 493–503.
17. Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354.
18. Berget, S. M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270,** 2411–2414.
19. Pertea, M., Lin, X., and Salzberg, S. L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29,** 1185–1190.
20. 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.
21. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452.
22. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

23. Lukashin, A. V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115.
24. Zhang, M. Q. (1998) Identification of protein-coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. *Plant Mol. Biol.* **37,** 803–806.
25. Solovyev V. V., Salamov A. A., and Lawrence C. B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22,** 5156–5163.
26. Pavy, N., Rombauts, S., Dehais, P., et al. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*. **15,** 887–899.
27. Yeh, R. F., Lim, L. P., and Burge, C. B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816.

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 6
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03

Compositor: Nettype
Date: 3/15/2003
Revision: First Proof