94–96 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

Gene discovery using the maize genome database ZmDB

Xiaowu Gai, Shailesh Lal, Liqun Xing, Volker Brendel* and Virginia Walbot[1]

Department of Zoology and Genetics, Iowa State University, Ames, IA 50011-3260, USA and [1]Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA

Received October 5, 1999; Accepted October 7, 1999

ABSTRACT

***Zea mays* DataBase (ZmDB) is a repository and analysis tool for sequence, expression and phenotype data of the major crop plant maize. The data accessible in ZmDB are mostly generated in a large collaborative project of maize gene discovery, sequencing and phenotypic analysis using a transposon tagging strategy and expressed sequence tag (EST) sequencing. ESTs constitute most of the current content. Database search tools, convenient links to external databases, and novel sequence analysis programs for spliced alignment are provided and together serve as an efficient protocol for gene discovery by sequence inspection. ZmDB can be accessed at http://zmdb.iastate.edu . ZmDB also provides web-based ordering of materials generated in the project, including EST and genomic DNA clones, seeds of mutant plants and microarrays of amplified EST and genomic DNA sequences.**

INTRODUCTION

Maize is the primary model plant for addressing fundamental biological questions in monocotyledonous plants. This taxon, which includes all the cereal crops, currently provides >70% of the caloric value of the human diet worldwide (1). Maize is thought to be a segmental allotetraploid reflecting the hybridization of two distinct diploid progenitors ~11–20 million years ago (2); today the genome contains an estimated 50 000–80 000 genes in ~2.3×10^9 base pairs present on 10 chromosomes (3). As in other plants, maize genes are compact; the distance between genes is large as a result of retrotranspson insertions (4). These features make a direct sequencing strategy for gene discovery currently impractical.

EST sequencing has emerged as the most effective way of identifying genes with moderately or highly abundant transcripts (5). Transposon tagging is a complementary technique that requires more effort but provides more information by efficiently combining gene discovery (identification and cloning) and functional genomics (analysis of the phenotypic consequences of altered gene expression) (6). Our project utilizes *Mu* transposons, which insert preferentially into low copy 'gene-like' sequences (7). Transgenic maize carrying a *Mu* element, *RescueMu*, that was engineered to contain a cloning vector, experience new insertion mutations. The *RescueMu* plasmid is cloned directly into *Escherichia coli* creating an immortalized library of insertions. Sequencing from the ends of the transposon yield maize genomic sequences highly enriched for genes.

ZmDB was created to display and analyze data generated in our project of maize gene discovery, and the site also includes details of our techniques and strategies. Most sequence data in ZmDB will consist of ~1.2 kb segments of maize genomic sequence flanking independent *RescueMu* transposon insertion sites. The plan is to sequence 150 000 such insertion locations, yielding 2–3-fold coverage of the haploid gene equivalent. To aid with accurate gene identification on the genomic clones, a collection of 50 000 ESTs is now being developed. One-fifth of these will be sequenced from both the 5′ and 3′ ends. Of the 4000 pairs of forward and reverse sequences available now, approximately half overlap to form continuous EST sequences, ranging from an average of ~750 to 950 bp depending on the source library. The up-to-date EST collection is assembled into tentative unique genes (TUGs) by contig and consensus building. The TUGs are annotated as tentative unique clusters (TUCs) or tentative unique singlets (TUSs). Necessarily, this annotation is temporary, as more sequence data may turn TUSs into TUCs and provide links between previously separate TUCs. The exon/intron structure of a genomic DNA segment can be readily delineated by spliced alignment to a cognate or homologous EST, if such exists, as described in the next section.

Annotation of the TUGs is initially automated. BLAST (8) searches are performed for all TUGs against public databases. The top three highly significant similarities are used as provisional tags for the corresponding TUG. The descriptions and keywords of those entries are carried over to the TUG entry. In that way, a text search for a keyword (for example, glutathione transferase, alternative splicing or nuclear location) will display all TUGs with similarities to entries described in some way by that keyword. Although crude, this automatic annotation procedure proves highly useful in providing the database user with easy starting points for a refined analysis, using ZmDB or their own tools. An ongoing attempt is made to incorporate refined annotation by human experts into the database.

NOVEL TOOLS FOR SPLICED ALIGNMENT

Alignment of an EST sequence with its genomic DNA origin is straightforward in the absence of sequence errors or polymorphisms: the introns, spliced out in the EST, will simply show up as long gaps in the alignment. The alignment task becomes more challenging when the matching is less than perfect because of sequencing errors, sequence variation

*To whom correspondence should be addressed. Tel: +1 515 294 9884; Fax: +1 515 294 6755; Email: vbrendel@iastate.edu

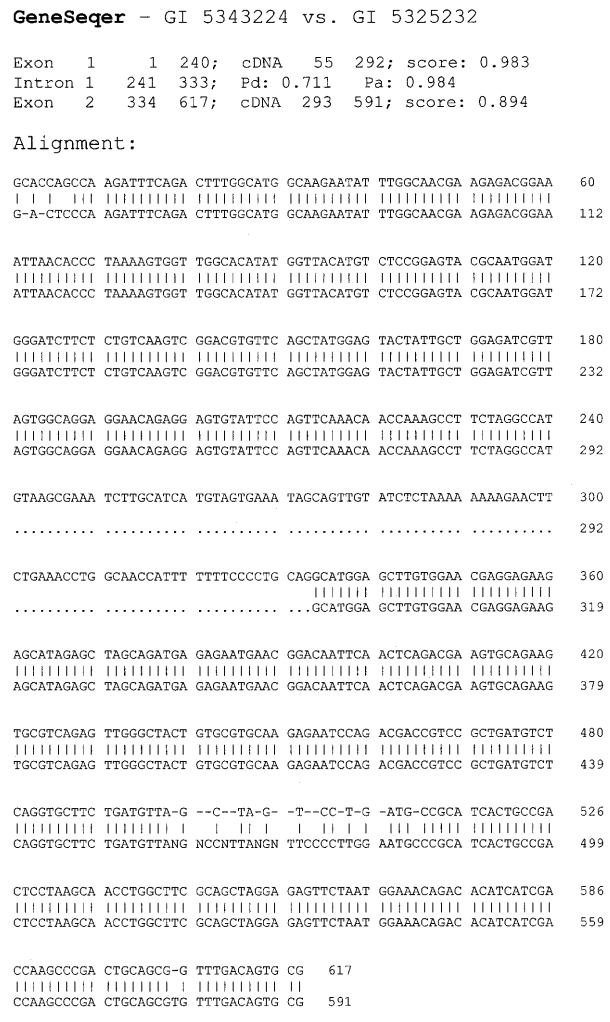

Figure 1. Spliced alignment of maize ESTs GenBank GI 5343224 with GenBank GI 5325232 by the GeneSeqer program (9). The alignment suggests a retained intron in the first sequence. Score, normalized sequence similarity score (1.00 for perfect identity); Pd and Pa, donor and acceptor splice site scores, respectively, as calculated by the SplicePredictor program (10). In the alignment, matching nucleotides are indicated by |, and the intron is indicated by dots.

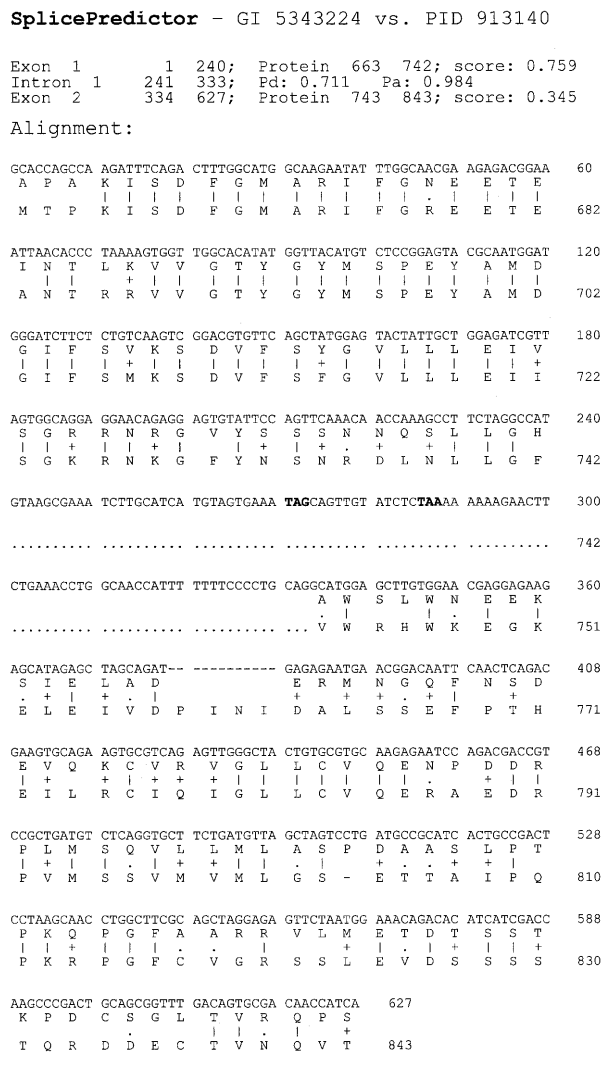

Figure 2. Spliced alignment of maize EST GenBank GI 5343224 with the *Arabidopsis* serine/threonine protein kinase ARK2 (GenBank PID 913140) by a novel function of the SplicePredictor program (10). The alignment confirms the intron assignment of Figure 1. Score, normalized sequence similarity score (1.00 for perfect identity); Pd and Pa, donor and acceptor splice site scores, respectively, as calculated by the SplicePredictor program. In the alignment, matching codons/amino acids are indicated by |, conservative substitutions by +, and neutral substitutions by ·; the intron is indicated by dots. Two in-frame stop codons within the intron are displayed in bold face.

(allelic variants, duplicate genes or gene families) or matching of non-cognate ESTs derived from a homologous locus. Alternative splicing outcomes represented in multiple ESTs can also impede exon identification. ZmDB provides the GeneSeqer program (9), which implements a dynamic programming algorithm to efficiently derive an optimal scoring spliced alignment. GeneSeqer simultaneously assesses the significance of the sequence alignment and the intrinsic quality of the implied splice sites. Figure 1 displays an example. In this case, two EST sequences were aligned and are shown to differ by a putative intron of 93 bases that is retained in one of the sequences.

GENE DISCOVERY AT ZmDB

ZmDB is designed to serve a dual function as both data repository and analysis workbench. Our protocol for potential gene discovery allows many entry points, depending on the data items available as starting points. For example, the relatedness of the ESTs in Figure 1 may have been discovered by either a BLAST (8) search of one of the sequences against ZmDB, or it may have been obtained indirectly from their mutual similarity to a different nucleotide or protein query. The putative retained intron suggests that these ESTs might derive from an alternatively spliced gene. A BLASTX (8) search of GenBank suggests that these ESTs encode a serine/threonine protein kinase. Using a novel function of the SplicePredictor program (10), the putative protein homologs may be used directly for a spliced alignment to the EST (Fig. 2). The strong alignment at the protein level,

the good splice site scores by the SplicePredictor algorithm, and two in-frame stop codons suggest that the intron would normally be spliced out. Retention of the intron would lead to a truncated translation product, that could be functional, or the retained intron could function as a post-transcriptional control point to reduce synthesis of the full-length protein.

FUTURE DIRECTIONS

Data to be added to ZmDB include *RescueMu*-derived maize genomic sequences, phenotypic data from maize plants expressing *RescueMu*-induced mutations, mapping data and microarray data to be generated by this project. The *ResueMu*-derived sequences will be annotated and (in conjunction with the EST sequences) assembled to an approximate unique set of maize genes. In doing so, other data such as possible alternative splicing of some genes will also be identified and added to the database. Any links between the phenotypic data, mapping data, microarray data and genomic and EST sequences will be identified and made accessible in the database. Users will have several new starting points for correlating gene discovery with possible functions, such as particular phenotypes or expression data. In the meantime we will continue developing computational tools to make the ZmDB site a convenient workbench for plant biologists.

AVAILABILITY

ZmDB is accessible at the URL http://zmdb.iastate.edu . Data files and source code for some of the algorithms used at ZmDB can be downloaded by anonymous ftp to ftp.zmdb.iastate.edu The manager of the database can be contacted by Email at zmdb@iastate.edu

ACKNOWLEDGEMENTS

The text of our NSF proposal is available for viewing at the ZmDB web site. Team members include the authors; Vicki Chandler, David Galbraith and Brian Larkins at the University of Arizona, Tucson; Michael Freeling and Sarah Hake at the University of California, Berkeley; Robert Schmidt and Laurie Smith at the University of California, San Diego; Marty Sachs at the University of Illinois, Urbana; and their collaborators at those locations. ZmDB is supported by the USA National Science Foundation grant NSF#9872657 (V.W. principal investigator).

REFERENCES

1. Chrispeels,M.J. and Sadava,D.E. (1977) *Plants, Food, and People*. W.H.Freeman, San Francisco, CA.
2. Gaut,B.S. and Doebley,J.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 6809–6814.
3. Gale,M.D. and Devos,K.M. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 1971–1974.
4. SanMiguel,P., Tikhonov,A., Jin,Y.-K., Motchoulskaia,N., Zakharov,D., Melake-Berhan,A., Springer,P.S., Edwards,K.J., Lee,M., Avramova,Z. and Bennetzen,J.L. (1996) *Science*, **274**, 765–768.
5. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) *Nature*, **377**, 3–174.
6. Walbot,V. (1992) *Annu. Rev. Plant Phys. Plant Mol. Biol.*, **43**, 49–82.
7. Cresse,A.D., Hulbert,S.H., Brown,W.E., Lucas,J.R. and Bennetzen,J.L. (1995) *Genetics*, **140**, 315–324.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
9. Usuka,J., Zhu,W. and Brendel,V. (1999) *Bioinformatics*, in press.
10. Brendel,V. and Kleffe,J. (1998) *Nucleic Acids Res.*, **26**, 4748–4757.