102–103 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

D*At*A: Database of *Arabidopsis thaliana* Annotation

Curtis J. Palm*, Nancy A. Federspiel and Ronald W. Davis

Stanford DNA Sequencing and Technology Center, 855 California Avenue, Palo Alto, CA 94304, USA

Received September 1, 1999; Revised and Accepted October 25, 1999

ABSTRACT

The Database of *Arabidopsis thaliana* Annotation (D*At*A) was created to enable easy access to and analysis of all the *Arabidopsis* genome project annotation. The database was constructed using the completed *A.thaliana* genomic sequence data currently in GenBank. An automated annotation process was used to predict coding sequences for GenBank records that do not include annotation. D*At*A also contains protein motifs and protein similarities derived from searches of the proteins in D*At*A with motif databases and the non-redundant protein database. The database is routinely updated to include new GenBank submissions for *Arabidopsis* genomic sequences and new Blast and protein motif search results. A web interface to D*At*A allows coding sequences to be searched by name, comment, blast similarity or motif field. In addition, browse options present lists of either all the protein names or identified motifs present in the sequenced *A.thaliana* genome. The database can be accessed at http://baggage.stanford.edu/group/arabprotein/

INTRODUCTION

Arabidopsis thaliana is a model system for plant biology research (1,2). Currently, a multinational project is sequencing the 120 Mb genome of this plant. The sequencing of the *Arabidopsis* genome is ~60% completed and is scheduled to be finished by the end of the year 2000. As part of the genome sequencing project, each sequencing group annotates the DNA sequence produced at its site. Access to the annotation data has been cumbersome. Presently, annotation data for only 65% of the sequenced genome is available through GenBank; therefore, viewing the annotation data from the entire sequenced genome requires navigating several web sites, many of which do not have database search capabilities. In addition, older annotation lacks the potential protein similarity data from new entries in the motif and protein databases. This new information would be useful in elucidating function of proteins annotated as hypothetical or unknown function.

The Database of *Arabidopsis thaliana* Annotation (D*At*A) was designed to address these annotation needs. First, D*At*A uses an auto-annotation process to annotate non-annotated GenBank records. This process also demonstrates the potential of auto-annotation to provide high quality gene predictions for genomic sequence data. Second, D*At*A provides a single, convenient site to search annotation for the entire set of extant sequenced *Arabidopsis* genomic DNA sequences. Third, D*At*A provides periodic updates to the protein similarity and protein motif data for all annotated proteins in the sequenced *Arabidopsis* genome.

DATA DESCRIPTION

The primary source of information for D*At*A is the GenBank (3) accessions for the *Arabidopsis* genome sequencing project. Protein coding sequence data (if present) and information about the source DNA are extracted from these entries and entered into the database. If no protein annotation is present in the record, our auto-annotation process is performed on the DNA sequence to generate protein predictions for the DNA clone. Briefly, the auto-annotation process runs the gene prediction program Genscan (4) in a recursive manner to predict coding regions in the DNA sequences. The recursive use of Genscan increases the number of the correctly assigned coding regions by reanalyzing, in smaller segments, regions of the DNA sequence where Genscan is likely to have predicted an incorrect coding sequence consisting of a combination of two or more genes. In the database all genes annotated by the automated process include 'AUTO' in their name (i.e. F1A10_AUTO.1) The auto-annotation was tested by comparing the auto-annotation of the DNA sequence of a set of *Arabidopsis* BAC clones which have annotation in their GenBank records. A summary of the results (Table 1) shows that auto-annotation performs well compared to manual annotation. All proteins are searched for protein motifs using eMotif (5) and for protein similarity using the Blast search program (6). The blast results for each protein are parsed to remove the uninformative matches (self, hypothetical proteins etc.). The remaining top significant (blast score > 80) blast hit is added to the database. In order to keep the protein similarity information current, the blast searches are repeated bimonthly for all protein sequences in D*At*A and the motif searches are repeated when new versions of the eMotif database are released.

The 8/09/99 update to the database contained 17 474 coding sequences from 76 Mb of genomic *Arabidopsis* DNA sequence. This represents ~60% of the *Arabidopsis* genome. A summary of the 8/09/99 update of the D*At*A content is in Table 2 and the current content summary can be found in at the D*At*A website (http://baggage.stanford.edu/group/arabprotein/NAR/summ.html). Information in D*At*A is stored and managed using the MySQL relational database software (http://www.tcx.se/).

*To whom correspondence should be addressed. Tel: +1 650 812 1994; Fax: +1 650 812 1975; Email: curtis@sequence.stanford.edu

Table 1. Auto-annotation comparison[a]

	Total number of annotated genes	Number of genes with functional information		
		Functions with correct size	Wrong size	Number of functions missed
Single pass auto-annotation	353	145	46	35
Recursive auto-annotation	418	191	27	8
Manual	401	226	–	–

[a]The 20 most recent (on 6/25/99) *Arabidopsis* BAC Genomic DNA sequences from GenBank were annotated with our auto-annotation process. Each auto-annotated gene was manually inspected and also compared to the annotation in GenBank for the BAC clones to assess the quality of the auto-annotation.

Table 2. D*At*A content summary[a]

	DNA sequence	No. of genes	Genes annotated as hypothetical or unknown	Genes with Blast hits	Genes with motifs
GenBank annotation	49.6 Mb	10 601	4912	6743	2160
Auto-Annotation	26.4 Mb	6873	3395	3441	789
Total	76.0 Mb	17 474	8307	10 184	2949

[a]As of 8/09/99, the current D*At*A content is available at the D*At*A website (http://baggage.stanford.edu/group/arabprotein/NAR/summ.html).

DATA ACCESS

D*At*A is accessible through the WWW at http://baggage.stanford.edu/group/arabprotein . An interactive figure showing the layout of the D*At*A website with links to example pages is at http://baggage.stanford.edu/group/arabprotein/NAR/map2.html

The D*At*A home page has a brief overview of the database and links to the main search page and to browsable lists of protein names and motif names. These lists allow alphabetical browsing of all the motif names, all the protein names and the number of times each occurs in D*At*A. Clicking on an entry in the list returns a table containing all the genes in the database that contain that entry.

The main search page has two search options, to search by protein function or to search by BAC name. The 'search for Protein by name' form searches for the entered word or phrase in any combination of the protein name, comment, blast results and motif name fields. The protein name and comment field originate from the GenBank records for annotated sequences. The blast results and motif name field are generated from search results for all the proteins in D*At*A. A successful search using this form returns a page listing, by gene name, all coding sequence records that contain the search phrase. The information about the coding sequences included on this page is gene name, protein name, comment field and blast field and motif field if they were selected in the search options. Clicking on the gene name in this table returns the coding sequence information page detailing the information for that gene. The information presented includes a graphic of the gene structure, the amino acid sequence of the protein, exon locations in the BAC, a GenBank link to the accession, the product name, comment field, top blast hit and score, motif name and link to blast search (with the aa sequence preloaded in the form) and a link to the web site annotation for the group that sequenced the clone.

The second search option allows one to search D*At*A by BAC name. The BAC name search also allows one to limit the search to a single sequencing group or chromosome (leaving the BAC name field empty while selecting one of these options will return all BAC clones for that option). The successful search returns a table of BACs listing Clone name, GenBank accession number (linked to the GenBank record), the Sequencing group, chromosome number, size of clone and whether the annotation was generated by our automated process. Clicking on a clone name in this table returns a new table listing the coding sequences annotated for the clone. This new table lists the gene name for each coding sequence in the BAC, along with its location in the BAC and protein name. Selecting a gene name from this table returns a coding sequence information page as described above for the protein name search.

ACKNOWLEDGEMENT

We thank Slava Glukhov for helpful discussions and computer systems support.

REFERENCES

1. Meinke,D.W., Cherry,J.M., Dean,C., Rounsley,S.D. and Koornneef,M. (1998) *Science*, **282**, 662–682.
2. Meyerowitz,E.M. and Somerville,C.R. (1994) *Arabidopsis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
4. Burge,C. and Karlin,S. (1997) *J. Mol. Biol.*, **268**, 78–94.
5. Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.
6. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.