Plant Biotechnology 22(3), 225–229 (2005)

Bioinformatics Note

KATANA: A web-based guide to public databases for *Arabidopsis* genomic information

Kentaro Yano, Tomoko Dansako, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata*

Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, Chiba 292-0818, Japan
*E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received July 27, 2005; accepted August 3, 2005 (Edited by T. Hashimoto)

Abstract Genomic information of *Arabidopsis thaliana* can be obtained from various public databases. Given that naming conventions often vary between databases and that the same genes can be annotated differently, we developed a web-based tool, KATANA (Kazusa *Arabidopsis thaliana* Annotation Abstract; http://www.kazusa.or.jp/katana/), to guide users searching for *Arabidopsis* genomic information to the relevant public databases from a single site. The tool contains information and annotations of genes and proteins, gene families, gene ontology (GO) terms, metabolic pathways and gene expression data from the Massively Parallel Signature Sequencing (MPSS) experiments. Given that entries in the tool are hyper-linked to those of the databases, detailed information can be accessed at the original sites. Advanced searches for metabolic pathways and GO terms can also be performed.

Key words: *Arabidopsis thaliana*, database, annotation, gene ontology, metabolic pathway.

The genome sequence of the model plant, *Arabidopsis thaliana*, was fully sequenced in the year 2000 (*Arabidopsis* Genome Initiative 2000). Since then, considerable efforts have been devoted to *Arabidopsis* research for identifying gene functions and annotating the genome sequence. To facilitate identification of *Arabidopsis* genes across databases, the *Arabidopsis* Genome Initiative (AGI) has assigned unique identifiers to AGI locus codes (e.g. At1g00010, At1g01030, At5g02430; http://mips.gsf.de/proj/thal/db/about/codes.html). In January 2004, The Institute for Genome Research (TIGR) re-annotated the 26,207 protein-coding genes and 3786 pseudogenes of the *Arabidopsis* genome and released the information as ATH1 version 5.0 (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/). The Gene Ontology (GO) Consortium has been developing shared, structured vocabularies (terms) to improve the consistency with which gene products are annotated among different databases and organisms (The Gene Ontology Consortium 2001). *Arabidopsis* genes have been assigned GO terms by The *Arabidopsis* Information Resource (TAIR) (Berardini et al. 2004).

Arabidopsis genomic information is available from public databases such as TIGR (http://www.tigr.org/tdb/e2k1/ath1/), TAIR (Rhee et al. 2003), Munich Information Center for Protein Sequences (MIPS) (Schoof et al. 2004), *Arabidopsis* Transcription Factor Database (AtTFDB) (Davuluri et al. 2003), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004) and *Arabidopsis thaliana* Biochemical Pathways (AraCyc) (Mueller et al. 2003). The collection and collation of *Arabidopsis* genomic information undertaken by disparate organizations has resulted in data, such as nomenclature of the same genomic characters, differing between databases. This in turn has resulted in the need to develop a tool with which these major databases can be queried from a single web site. The data thus obtained could be used for comparative purposes and the user could then be directed to the original databases. To our knowledge, however, a portal of this kind has not yet been developed to date.

We therefore designed a web-based tool, KATANA (Kazusa *Arabidopsis Annotation* Abstract) that searches user queries to the majr Arabidopsis databases and lists the results on a user's screen. Since the contents of the list are hyper-linked to the original information in the databases, the user can subsequently access the detailed information of interest from the original sites.

The KATANA tool was written using MySQL (http://www.mysql.com/) and Hypertext Preprocessor (PHP) (http://www.php.net/) for the Sun FireV880 web server (Solaris 8). The datasets for searching queries and generating hyper-links were collected from public databases; genomic and complementary DNA (cDNA)

Abbreviations: TAIR, The *Arabidopsis* Information Resource; MIPS, Munich Information Center for Protein Sequences; TIGR, The Institute for Genomic Research; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, gene ontology; RAFL, RIKEN *Arabidopsis* full-length; cDNA, complementary DNA; CDS, coding sequence; MPSS, Massively Parallel Signature Sequencing

Table 1. The datasets collected for creating web-guide tables.

Type of dataset	Web-site	URL
Sequence information	TAIR	http://www.arabidopsis.org/
Genome annotation	TAIR	http://www.arabidopsis.org/
	TIGR	http://www.tigr.org/
	MIPS	http://mips.gsf.de/
	AtTFDB	http://arabidopsis.med.ohio-state.edu/
GO annotation	TAIR	http://www.arabidopsis.org/
GO and definition	GO	http://www.geneontology.org/
Gene family	TAIR	http://www.arabidopsis.org/
	AtTFDB	http://arabidopsis.med.ohio-state.edu/
Metabolic pathway	AraCyc (TAIR)	http://www.arabidopsis.org/
	KEGG	http://www.genome.ad.jp/
Full-length cDNA clone	RIKEN	http://rarge.gsc.riken.go.jp/cdna/cdna_keyword.pl
Gene expression	University of Delaware	http://mpss.udel.edu/at/java.html

sequences, coding sequence (CDS), protein sequences, annotations of genes and proteins from TIGR, TAIR and MIPS, gene family names from TAIR and AtTFDB, gene ontology (GO) (Gene Ontology Consortium 2001) terms from TAIR (Berardini et al. 2004), metabolic pathway names from KEGG and AraCyc, *Arabidopsis* full-length (RAFL) cDNA clone information (Seki et al. 2002) from RIKEN, and gene expression data from the Massively Parallel Signature Sequencing (MPSS) project (Meyers et al. 2004) being undertaken at the University of Delaware (Table 1). Since the annotations and GO terms in the TAIR database and the pathway information in the KEGG database are updated daily to monthly, KATANA uses Perl and bash scripts to update itself. Other datasets are manually updated when the original databases are updated.

The main component of KATANA, the ANNOME (*Arabidopsis* Annotation Comprehensive Search) module, allows users to search datasets using a set of queries. KATANA also has the comparative search tool for metabolic pathways of KEGG and AraCyc (Metabolic Pathway Search) and the advanced search tool for GO terms (Gene Ontology Search). The tool functions are accessible from the main page of KATANA (http://www.kazusa.or.jp/katana/).

ANNOME search

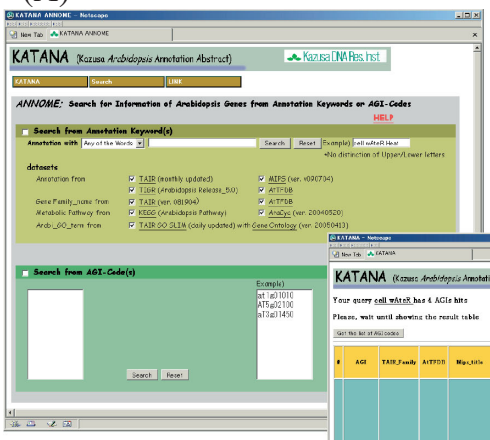
Users can search the datasets using multiple keywords that appear in gene annotations of the public databases, or AGI locus codes (Figure 1A). When multiple keywords are used in a query, one of two options ("All of the Words" or "Any of the Words") can be selected from a pull-down menu to find the term of interest. Users can also select which datasets they want to query, or all of the databases can be selected by default. The results of the search are listed in a comparative table on the user's browser (Figure 1B).

To each AGI code, the table contains the following entries, TAIR family name, AtTFDB entry (if available), MIPS annotation, TAIR gene model(s), TIGR annotation, RAFL clone code(s), names of KEGG and AraCyc pathways that contain the AGI code (if available), TAIR GO terms and gene expression data of the MPSS. These entries are displayed based on to the databases selected. As each entry in the table is hyper-linked to the original site of the corresponding public database, users can access specific information for the character of interest at the site of the host. The annotations for TAIR are shown under "gene models", where the term "gene models" (e.g., At1g05230.1, At1g05230.2) are defined as those genes from the same physical or genetic locus but that have different structural or functional annotations, or that produce different transcribed products (Rhee et al. 2003). The MPSS gene expression data are based on 17-base sequence "signatures" representing transcripts in seventeen DNA libraries using tissues from the callus, inflorescence, leaf, root, silique and seedlings (Meyers et al. 2004).

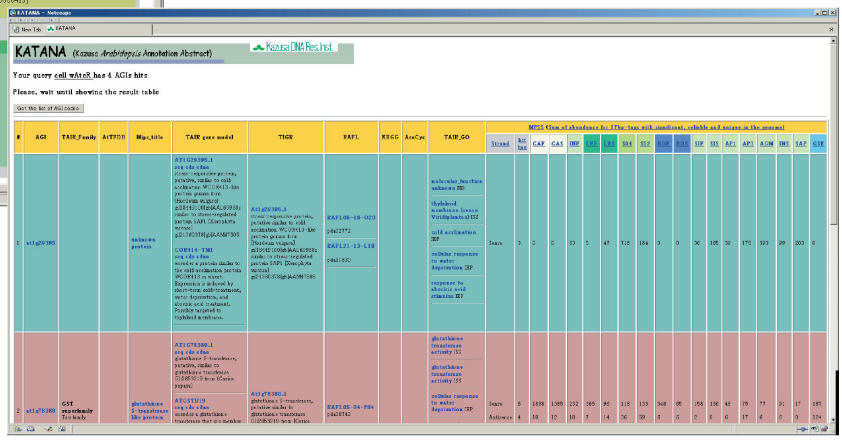
The list of AGI codes can be displayed on a user's browser by clicking "Get the list of AGI codes". The list can be copied and pasted in text format and used for additional bulk searches of the public databases; searches such as FASTA sequences (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp), microarray gene expression data (http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression) and sequence motifs (http://www.arabidopsis.org/tools/bulk/

Figure 1. A web-based guide KATANA. (A) Query input page of ANNOME. (B) Result page of ANNOME. The "strand" column in the MPSS signature table shows the orientation with terms of "Sense" and "Antisense", which are for the coding sequence and the complementary sequence, respectively. The "hit_tag" column shows the number of occurrences of the signatures in the *Arabidopsis* genome. The expression data are displayed for the seventeen libraries (see details at http://mpss.udel.edu/at/Library_Info.php), which were constructed from callus, inflorescence, leaf, root, silique and seedling data (Meyers et al. 2004). (C) Query input page of metabolic pathway search. (D) Result page of metabolic pathway search. (E) Query input page of advanced GO search. (F) Result page of advanced GO search.

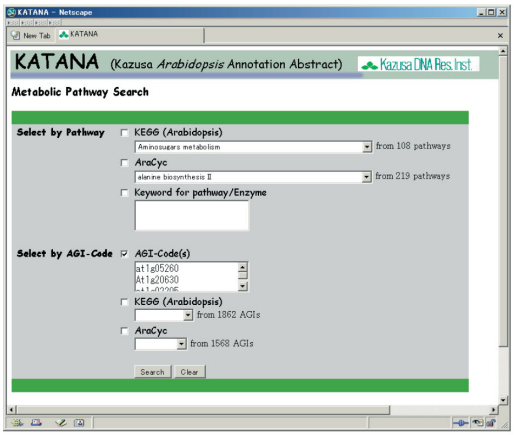
(A)



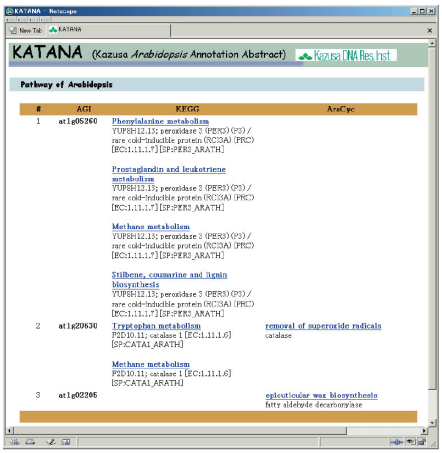
(B)



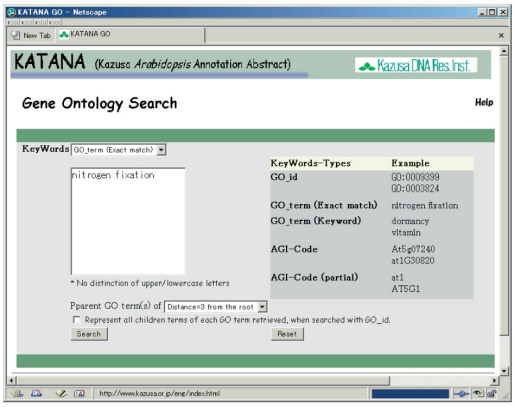
(C)



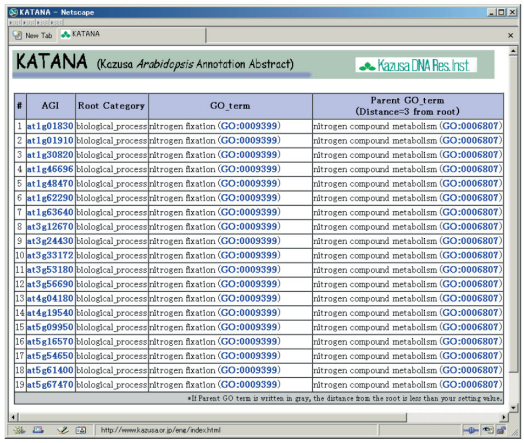
(D)



(E)



(F)



motiffinder/index.jsp).

Search function for metabolic pathway information

A search function for the metabolic pathway data hosted by KEGG/PATHWAY and AraCyc are accessible by clicking "Metabolic Pathway Search" on the main page (Figure 1C). As all pathway names (108 from KEGG/PATHWAY and 219 from AraCyc) and all AGI codes related to the relevant metabolic pathways (1862 from KEGG/PATHWAY and 1568 from AraCyc) are listed in the pull-down menus, users can select one of the pathway names or AGI codes from the lists. Searches can also be conducted using keyword(s) for pathways or enzyme names and AGI code(s) supplied by the user. The result page shows the AGI codes hit by the query and the corresponding KEGG/PATHWAY and AraCyc entries, all of which are hyper-linked to the original web pages (Figure 1D).

A comparative list of the metabolic pathways of interest in the guide table is useful for comparing which database is more complete for that particular query. As of April 2005, 1, 862 and 1568 AGI codes were listed in the KEGG/PATHWAY and AraCyc databases, with the difference in the number of entries probably due to differences in the selection criteria applied for selecting the *Arabidopsis* genes involved in certain metabolic pathways. Interestingly, only 659 codes are overlapped in the databases. Thus, simultaneous representation of the queried metabolic pathways enables users to familiarize themselves with differences in the contents of the databases.

Advanced search function for GO information

The "Gene Ontology Search" function is used to augment a user's query or queries with comprehensive GO information (Figure 1E). It differs from the ANNOME search function in that the latter only provides GO terms for the AGI codes that are selected when using the ANNOME module. To search for GO information, a user first selects one of the keyword categories from the GO identification list (GO_id), GO term (exact match), GO term (partial match), AGI code (exact match) and AGI code (partial match), before querying the selected category using the query box. The tool also provides information on one of the parent GO terms in the structured ontology vocabularies for each GO term based on the user's query, which is useful for knowing the location of the GO term in a broad sense. The level of the parent GO term in the hierarchy can be selected using a pull-down menu, the default setting of which is 3, which is the number of levels from the root category containing the terms, "biological process", "cellular component" or "molecular function". As an option, all child GO terms can be retrieved for each GO term, when searched with GO-id(s). The retrieved data consists of the AGI code, the root category, the GO term and the selected GO parent, all of which are hyper-linked to the AmiGO browser (http://www.godatabase.org/cgi-bin/amigo/go.cgi) (Figure 1F). Since the 33,498 distinct *Arabidopsis* genes are associated with 107,219 GO terms (Jan 21, 2005; http://www.arabidopsis.org/info/ontologies/go/), the GO search function is useful for researchers working on gene function.

Summary

The web-based KATANA tool facilitates user access and queries of available *Arabidopsis* genomic information held by major public databases. Each result generated in response to a user's query is hyper-linked to the original site, which means that users can access detailed information from the original source.

Acknowledgements

This work was supported by New Energy and Industrial Technology Development (NEDO) (as part of the project called "Development of Fundamental Technologies for Controlling the Material Production Process of Plants").

References

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

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G. Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol* 135: 745–755

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4: 25

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucl Acids Res* 32: D277–D280

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S (2004) The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res* 14: 1641–1653

Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol* 132: 453–460

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucl Acids Res* 31: 224–228

Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF

(2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucl Acids Res* 32: D373–D376

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296: 141–145

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res* 11: 1425–1433