Plant Biotechnology 26, 9–13 (2009)

Bioinformatics Note

DAGViz: a directed acyclic graph browser that supports analysis of Gene Ontology annotation

Kentaro Yano[1,2], Koh Aoki[1], Hideyuki Suzuki[1], Daisuke Shibata[1,*]

[1] Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Faculty of Agriculture, Meiji University, Kawasaki, Kanagawa 214-8571, Japan
* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received December 30, 2008; accepted January 27, 2009 (Edited by Y. Ogata)

Abstract The number of nucleotide and protein sequences deposited in public databases has rapidly increased in recent times. To provide consistent annotation of orthologous genes across organisms, annotation using Gene Ontology (GO) terms is shared among most of these databases. GO outlines a structured vocabulary used for describing biological properties of gene products. The illustration of GO terms using complete paths to the root term on the basis of a directed acyclic graph (DAG) approach aids in the systematic grasping of the functional information related to a gene product. However, the website provided by the GO Consortium does not present DAGs for two or more GO terms due to difficulties in the depiction of the complex relationships of such GO terms. To overcome these problems, we have constructed a DAG-based browser, termed DAGViz, that shows DAG-based information of multiple GO terms assigned to one or more genes within a single screen (http://www.pgb.kazusa.or.jp/dagviz/). In the current report, we illustrate the advantages of DAGViz in analyzing GO annotation.

Key words: Database, annotation, Gene Ontology, directed acyclic graph.

Due to improvements in high-throughput technologies for comprehensive genomic, transcriptomic and proteomic analyses, the number of nucleotide and protein sequences deposited and stored in public databases has increased rapidly in recent times (Lee et al. 2005, Mueller et al. 2005, Liang et al. 2008, Rice Annotation Project 2008, Sugawara et al. 2008). In order to provide consistent annotation of orthologous genes across organisms, annotations using Gene Ontology (GO) terms and GO identifiers are shared by most gene databases for model organisms. GO provides a dynamically controlled vocabulary for describing functional categories of gene products (The Gene Ontology Consortium 2008). GO terms are grouped into three main categories with different organizing principles and include 1) molecular function (MF), 2) biological process (BP) and 3) cellular components (CC). These terms, representing the main categories, are hereafter referred to as the roots of GO. GO terms are structured in a hierarchy form of a directed acyclic graph (DAG), in which lower level elements are connected to one or more upper level elements in an upward direction, and thus all pathways are acyclic. In a DAG for GO terms, a more specialized lower level gene termed child can be related to a more generalized upper level gene termed parent, representing more general functional categories. These parent-child relationships provide complete paths to the roots of the main functional categories; that is MF, BP and CC (Figure 1A). Accordingly, one or more GO terms may be assigned to a gene product at various levels in these categories.

The generalization of GO terms assigned to a particular gene of interest facilitates the functional analysis of the gene, in which more general functional categories of the gene product can be obtained by tracing paths from the GO terms to the upper level terms of the GO hierarchy. For example, an *Arabidopsis* gene identified as At5g20980 (methionine synthase 3) is associated with the GO term 'methionine biosynthetic process' in the BP category (Figure 1B). Parental GO terms for this gene including 'Sulfur amino acid biosynthetic process' and 'Aspartate family amino acid biosynthetic process', indicate that this gene is involved in the functional categories of sulfur metabolism and aspartate family amino acid metabolism. This example among others, suggests that an overview of DAG-based information regarding GO terms will be helpful in inferring functional categories associated with a gene product of interest. A graphical view of the DAG-based information for each individual GO term is available in public databases such as AmiGO

Abbreviations: BP, biological process; CC, cellular component; DAG, directed acyclic graph; GO, Gene Ontology; MF, molecular function.
This article can be found at http://www.jspcmb.jp/

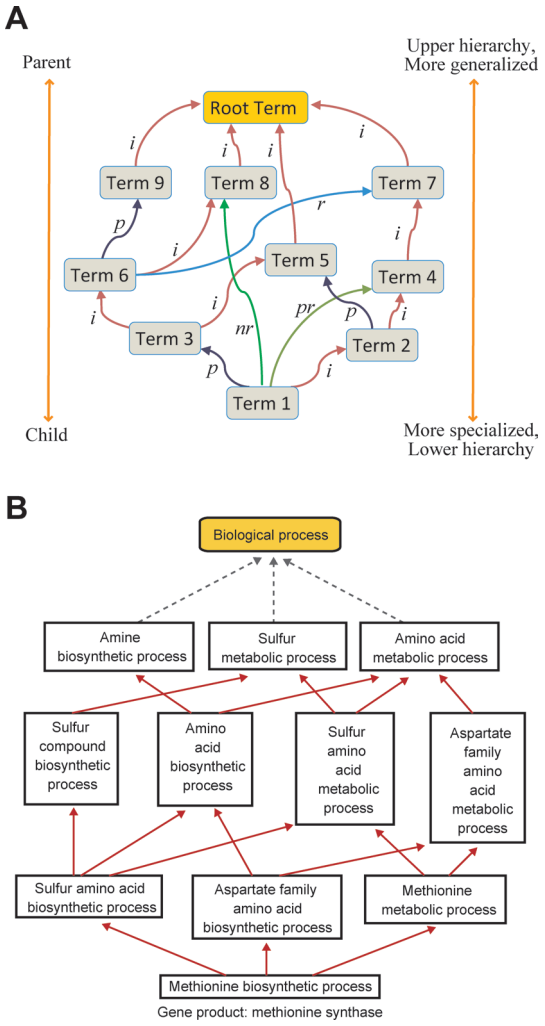

Figure 1. Schematic of DAG and parent-child relationships. (A) A DAG from a given term (Term 1) to a root term. Five types of relationships between terms are used in the current GO database: is a (i), part of (p), regulates (r), positively regulates (pr) and negatively regulates (nr) (also see the GO web site). Each term in DAG may be a "child" of one or more than one "parent" term. (B) Example of a part of DAG for the GO term 'methionine biosysthetic process'.

(http://amigo.geneontology.org/cgi-bin/amigo/go.cgi). However, AmiGO fails to depict DAGs for two or more GO terms on the same screen.

In the current study, we report the construction of a DAG-based browser entitled DAGViz (http://www.pgb.kazusa.or.jp/dagviz/), that facilitates the integrative analysis of functional categories associated with a gene product by displaying all DAG-based information for multiple GO terms using a tabulated color chart screen.

Database and browser construction

DAGViz was constructed using MySQL (http://www.mysql.com/) and Hypertext Preprocessor (PHP, http://www.php.net/) on our Linux server (Red Had Enterprise Linux ES v.4 for x86). GO annotation datasets for DAGViz, which contain GO terms, relationships between

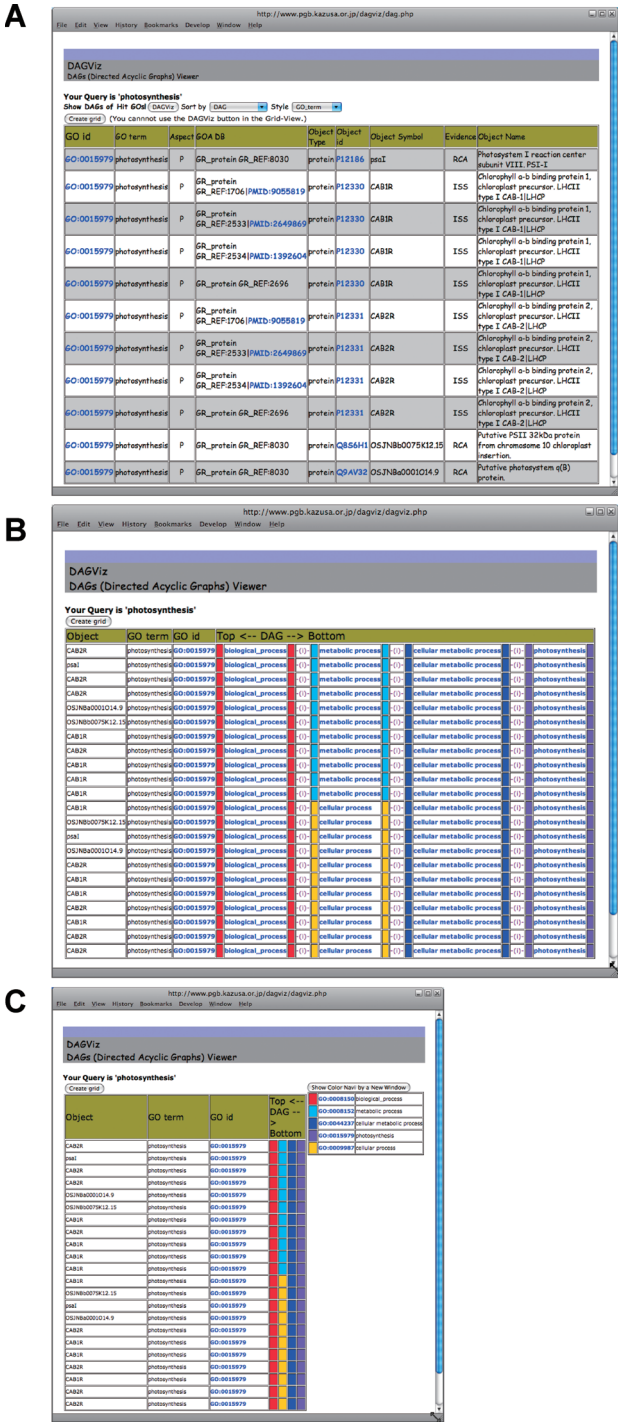

Figure 2. Example of a search result using DAGViz. (A) Result page of a GO term search. GO terms and the keywords "photosynthesis" were searched in a dataset of "*Oryza sativa*" and the gene products are shown in the table. (B) DAG browser. DAGs for all GO terms outlined in Figure 2A are shown on the one web page. (C) DAG browser by color chart. GO terms are presented by color rather than GO terms.

GO terms and evidence codes for individual gene products, were obtained from the web site of the GO project (http://www.geneontology.org/index.shtml). Evidence codes were used to indicate how the annotation to a particular term was supported. Relationships

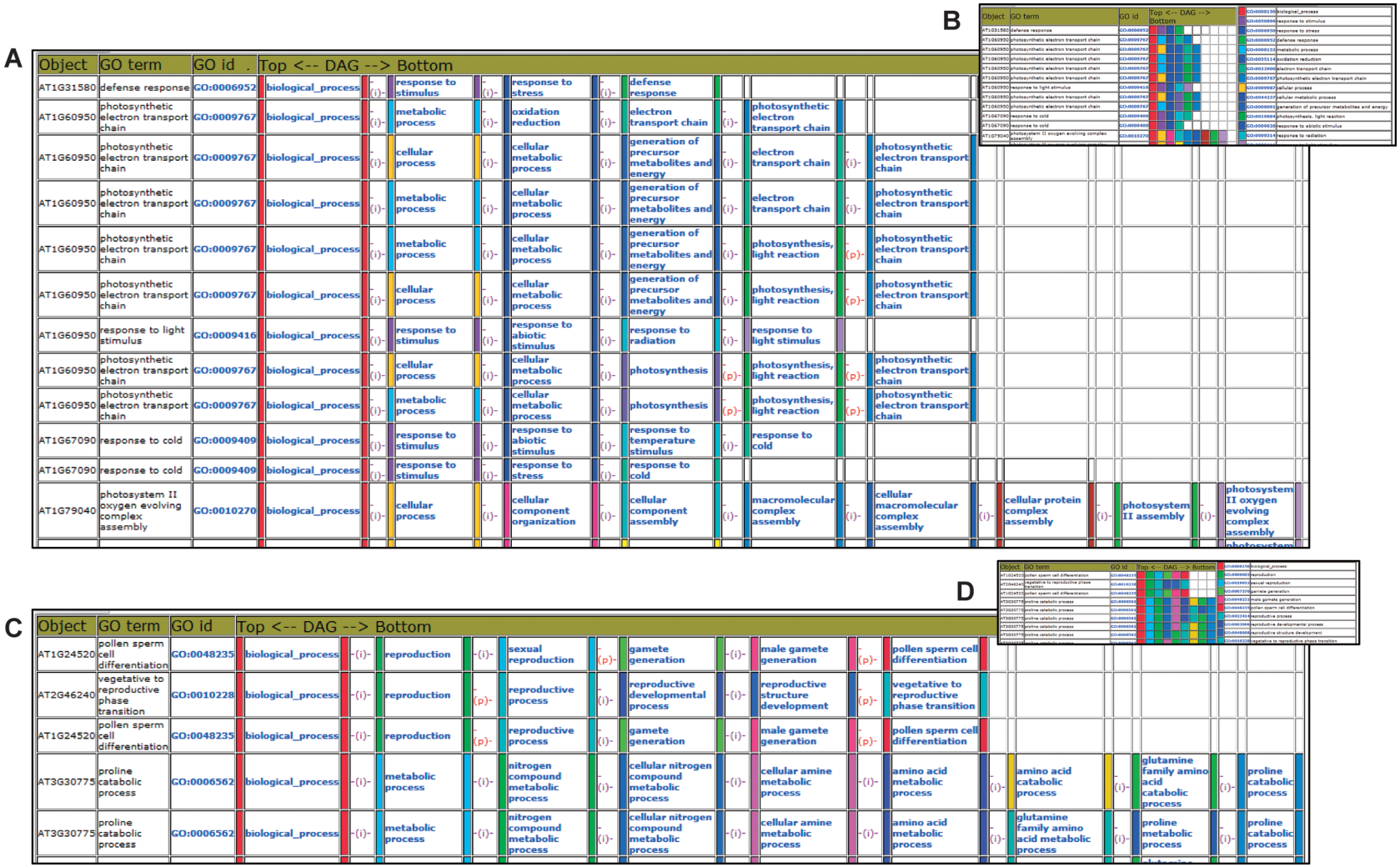

Figure 3. Example of the comparison of the biological processes associated with two gene sets. Each DAG browser outlines a portion of the result page (top of the page). (A) DAGs for GO terms assigned to *Arabidopsis* genes that are highly expressed in the leaf. (B) Color chart for DAGs of the GO terms outlined in Figure 2A. (C) DAGs for GO terms assigned to *Arabidopsis* genes that are highly expressed in the flower. (D) Color chart of DAGs of the GO terms outlined in Figure 2C.

between GO terms fell into five categories: 1) is a, 2) part of, 3) regulates, 4) positively regulates and 5) negatively regulates. We developed a Java program to construct a color-coded table of DAG-based information, in which a unique color was assigned to each GO term obtained from the datasets. The current version of DAGViz contains GO annotation datasets for 45 organisms including fungi, protist, animal, plant and bacterial species.

Search functions of DAGViz

DAG-based information in DAGViz can be retrieved by selecting organisms, searching keywords or by selecting specific categories such as GO terms, GO identifiers or names of gene products. The retrieved data, that consists of GO terms, evidence codes and names of gene products, are shown in the DAGViz web browser. Users can also select one or more categories of evidence codes using the retrieval function. Figure 2A shows an example of the search results using '*Oryza sativa*' as an organism, 'photosynthesis' as a keyword and 'GO terms' as its category. Upon clicking the button 'DAGViz', DAG-based information for all of the retrieved GO terms is displayed (Figure 2B). Using the default setting, DAGViz presents complete paths from the retrieved GO terms to the root terms. When a user selects 'Color Chart' in the 'Style' selector, GO terms are color-coded in colors unique to each individual term, without the description of the terms (Figure 2C). The legend of colors for GO terms is listed on the same page. The legend can also be highlighted on other pages by clicking on the button "Show Color Navi by a New Window'.

An example of a comparative analysis using GO terms between different sets of genes

The functional classification of GO terms using DAGViz facilitates the efficient characterization of biological functions and processes associated with not only a single gene, but also with a set of genes. In the current study, we present the example of a comparative analysis using the GO term BP between different sets of genes. The expression dataset used in the comparison (Bergmann et al. 2004), and whose experimental identifiers were GSM9225 for leaf and GSM9230 for flower, was obtained from Gene Expression Omnibus (Barrett and Edgar 2006). We selected two sets of twenty *Arabidopsis* genes that were highly expressed in both the leaf and flower, in accordance to the following criteria: 1) their detection indices were 'Present' and 2) their expression ratio between leaf and flower were more than 2.0 or less than 0.5. For the selected genes, the GO terms BP were retrieved in the "*Arabidopsis thaliana*' database using the evidence codes of EXP, IDA, IPI, IMP and IGI. The GO terms assigned to the highly-expressed genes in leaf included photosynthesis, light harvesting in photosystem I (GO : 0009768 as a GO identifier), photosystem II oxygen evolving complex assembly (GO:0010270), photosynthesis (GO : 0015979), and response to light stimulus (GO : 0009416) (Figure 3A, 3B). Interestingly, the DAG-based information of GO terms from the genes contained the GO term acid catabolic process (GO : 0046395), which in general concerns biological processes associated with the leaf, although the term is not directly associated with the genes. The set of genes that were highly expressed in flower were associated with the GO term sexual reproduction (GO : 0019953), which is consistent with the expected biological processes undertaken in flower organs. These results demonstrate that DAGViz is helpful in predicting differences in biological function and processes associated with different sets of genes.

Summary

DAGViz allows users to perform functional analyses using GO annotations in various organisms. Using this program, all DAG-based information associated with multiple GO terms of interest may be displayed in a single screen. In contrast, other databases such as AmiGO provide DAG-based information for a single GO term only. For a set of genes of interest, DAGViz is also useful for comparing biological function and processes associated with gene sets.

Acknowledgements

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

References

Barrett T, Edgar R (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO). *Methods Mol Biol* 338: 175–190

Bergmann DC, Lukowitz W, Somerville CR (2004) Stomatal development and pattern controlled by a MAPKK kinase. *Science* 304: 1494–1497

The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucl Acids Res* 36: D440–D444

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucl Acids Res* 33: D71–74

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L (2008)

Gramene: a growing plant comparative genomics resource. *Nucl Acids Res* 36: D947–953

Mueller LA, Tanksley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, Ling H, Xue Y, Seymour G, Bishop G, Bryan G, Sharma R, Khurana J, Tyagi A, Chattopadhyay D, Singh NK, Stiekema W, Lindhout P, Jesse T, Lankhorst RK, Bouzayen M, Shibata D, Tabata S, Granell A, Botella MA, Giuliano G, Frusciante L, Causse M, Zamir D (2005) The Tomato Sequencing Project, the First Cornerstone of the International Solanaceae Project (SOL). *Comp Funct Genom* 6: 153–158

Rice Annotation Project (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucl Acids Res* 36: D1028–1033

Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y (2008) DDBJ with new system and face. *Nucl Acids Res* 36: D22–24