W64–W70 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 30 April 2010*
doi:10.1093/nar/gkq310

agriGO: a GO analysis toolkit for the agricultural community

Zhou Du, Xin Zhou, Yi Ling, Zhenhai Zhang and Zhen Su*

State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China

Received January 16, 2010; Revised April 4, 2010; Accepted April 17, 2010

ABSTRACT

Gene Ontology (GO), the *de facto* standard in gene functionality description, is used widely in functional annotation and enrichment analysis. Here, we introduce agriGO, an integrated web-based GO analysis toolkit for the agricultural community, using the advantages of our previous GO enrichment tool (EasyGO), to meet analysis demands from new technologies and research objectives. EasyGO is valuable for its proficiency, and has proved useful in uncovering biological knowledge in massive data sets from high-throughput experiments. For agriGO, the system architecture and website interface were redesigned to improve performance and accessibility. The supported organisms and gene identifiers were substantially expanded (including 38 agricultural species composed of 274 data types). The requirement on user input is more flexible, in that user-defined reference and annotation are accepted. Moreover, a new analysis approach using Gene Set Enrichment Analysis strategy and customizable features is provided. Four tools, SEA (Singular enrichment analysis), PAGE (Parametric Analysis of Gene set Enrichment), BLAST4ID (Transfer IDs by BLAST) and SEACOMPARE (Cross comparison of SEA), are integrated as a toolkit to meet different demands. We also provide a cross-comparison service so that different data sets can be compared and explored in a visualized way. Lastly, agriGO functions as a GO data repository with search and download functions; agriGO is publicly accessible at http://bioinfo.cau.edu.cn/agriGO/.

INTRODUCTION

The availability of high-throughput techniques allows biologists to monitor changes and regulation at a genome-wide level under certain conditions. Such experiments normally generate huge data sets of genes' expression values under different treatments. There are challenges in the analysis and interpretation of these data sets with one promising strategy to solve these problems being gene-annotation enrichment analysis. The bioinformatics community has developed multiple enrichment tools which were compared and summarized by Huang *et al*. (1). The majority of these tools (2–12) employ Gene Ontology (GO) (3) as their annotation resource, since GO is a controlled vocabulary system with rich content for gene function description at a molecular level and is supported by many consortia focusing on different organisms. Unfortunately, most GO enrichment tools have limited support for agricultural species. Recently, four applications enabling analysis of agricultural species data were evaluated by Berg *et al*. (13). Among four tools, only EasyGO (12) is designed to especially serve the agricultural community. Since its release, this tool has processed >20 000 analysis requests from all around the world and is referenced by 20 publications. After 3 years of continued maintenance, we developed the successor of EasyGO, a web-based toolkit named agriGO with enhanced and novel functionalities.

Retaining the advanced features of EasyGO, agriGO also continues to focus on agricultural species. The enrichment analysis approach used in EasyGO is categorized as SEA (Singular enrichment analysis) in Huang's survey (1). We kept this method because although SEA is the most traditional strategy, it is still very efficient and such continuity will not reduce its accessibility to past users. However, new features were added to meet current complex demands. First, new tools including PAGE (Parametric Analysis of Gene set Enrichment), BLAST4ID (Transfer IDs by BLAST) and SEACOMPARE (Cross comparison of SEA) were developed. The arrival of these tools provides users with possibilities for data mining and systematic result exploration and will allow better data analysis and interpretation. Second, the exploratory capability and result

*To whom correspondence should be addressed. Tel: +86 10 62731380; Fax: +86 10 62731214; Email: zhensu@cau.edu.cn

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

visualization are enhanced. Results are provided in different formats: HTML tables, tabulated text files, hierarchical tree graphs, and flash bar graphs. Third, in agriGO, PAGE and SEACOMPARE can be used to carry out cross-comparisons of results derived from different data sets, which is very important when studying multiple groups of experiments, such as in time-course research. Furthermore, we integrated comprehensive annotations like gene description and protein domain annotation into agriGO, and the information is searchable and downloadable. Technically, working on a more powerful server, agriGO is completely reengineered providing a faster, more robust and flexible tool. Flash technology (http://teethgrinder.co.uk/open-flash-chart-2) is used to generate the result bar graphs. Lastly, this new toolkit is user-friendly with an interactive help system and flexible input requirements.

ENRICHMENT ANALYSIS

Huang *et al.* (1) classified enrichment tools into three categories: SEA, GSEA (Gene Set Enrichment analysis) and MEA (modular enrichment analysis). EasyGO (12) is classed as SEA. In agriGO, the enrichment analysis strategy in EasyGO is kept and improved, and named as 'SEA'.

SEA analysis computes GO term enrichment in one set of genes by comparing it to another set, named the target and reference lists, respectively. As for EasyGO, a default reference list with pre-computed GO term mappings is provided for each data type.

For each supported species, we collected currently popular gene nomenclatures and probe (set) names from different microarray platforms, and computed background GO term mappings. With rice for example, available background includes TIGR (14) and Gramene (15) genes, KOME (16) full-length cDNAs and microarray probe (set) IDs from Affymetrix (www.affymetrix.com), Agilent (www.agilent.com), BGI (www.genomics.org.cn) and other platforms. As a new feature, a custom list with user-defined GO annotation can be uploaded as either the target list or the reference list. agriGO allows arbitrary combination of target and reference lists, to address the data deficiency issue for species that do not yet have a sequenced genome (GO backgrounds from most related species can be used when analyzing gene sets from such species). Such cross-data type combination and interpretation should be conducted with care.

For advanced options, three statistical methods can be selected: hypergeometric, Fisher's exact and χ^2 tests. When the target list comprises a subset of the reference list, the hypergeometric test or Fisher's exact test should be applied. If the target list has few or no intersections with the reference list and its size is large, χ^2 is appropriate.

The multi-testing problem seems inevitable when a large number of GO terms are subjected to statistical calculation. Therefore, SEA performs the Benjamini–Yekutieli method (17) to do the multiple comparison correction by default, while others, such as Benjamini–Hochberg (18), Storey q-value (19) and Holm (20) methods, are also available. The same choices for adjustment methods are provided for the PAGE analysis, as described below.

GSEA is a popular way to do enrichment analysis, since it reduces the arbitrary factors in the gene selection step of SEA and can utilize more information such as gene expression values. Different strategies for GSEA have been introduced already; we chose PAGE which was first proposed by Kim and Volsky (21), because it is relatively straightforward, and accuracy is preserved while computation load is lower. PAGE is based on the Central Limit Theorem (CLT) (22), and according to the CLT, the distribution of the average of randomly sampled n observations tends to follow a normal distribution as n gets larger, whether the parent distribution is normal or not. Here, assuming mean μ and variance σ^2 of the parent, then the sample mean will follow a normal distribution with the same mean μ and the variance σ^2/n. In this context, the parent can be seen as a set of fold change (FC) values between two experimental groups, the random sample is the GO term where n is the number of genes mapped to the term. Thus, for each term having sufficient number of genes mapped to it, a Z-score value, which is used to infer the statistical significance, can be calculated using the following z-test formula:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$\bar{X}_n$ is the mean of sample n, i.e. the average of FC of all genes associated with the GO term. As a z-test is two-tailed, a Z-score can be positive or negative. Using R software (23), a Z-score is converted to P-value, and the P-value will be subjected to multiple test correction. The adjusted P-value generated by the correction is one criterion to estimate whether the term is significant. Apart from the adjusted P-value, an additional criterion is applied in PAGE. Either the term has a positive Z-score and the mean of FC of all genes associated with it is ≥ 1 (upregulated), or the term has negative Z score and the FC mean is ≤ 1 (downregulated).

Generally, PAGE is more objective than SEA. SEA accepts a user-selected target list and uses the adjusted P-value as a single criterion to decide GO term enrichment. In the case of an inappropriately prepared target list, a misleading result might be generated. In contrast, PAGE accepts an arbitrarily large input-list with FC, and identifies significant GO terms associated with groups of genes with significantly deviated change patterns with respect to all the genes. However, PAGE is only applicable for sample comparisons with quantitative measurements (e.g. mRNA abundance and DNA methylation) and factors including precision of measurement and data normalization will influence the PAGE result. In their application, the two approaches serve for different situations and need special attention to the issues mentioned above.

FEATURES AND FUNCTIONALITIES

The ability to present the analysis results in a clear and accessible manner is important in the interpretation step.

In EasyGO, a hierarchical tree graph is used to aid the user in checking the results. We expanded this type of output with more content and functionalities, as described below. A cross-comparison function was developed, to enable users to simultaneously compare multiple data sets.

Enhanced graphical presentation

Elaborate graphical output can facilitate users to explore biological meaning in an intuitive way. The direct acyclic graph or tree structure graph based on the nature of GO can indicate terms are over/under-represented and the inter-relationships between terms. Such graph is available in EasyGO and improved in agriGO. We adopted the testing case from EasyGO (12), which comprised of 168 probe sets from Arabidopsis ATH1 GeneChip with all showing upregulated expression in shoot tissue during cold treatment, data from AtGenExpress project (24). We used SEA in agriGO and EasyGO to do the analysis, and both generated a tree structure graph (see Figure 1 and Supplementary Figure S1, respectively). GO terms are represented as boxes containing detailed description, organized and connected based on their relationship (Figure 1). The detailed pages containing further information, such as gene description and protein domain annotation, are also available. In addition, font and rank direction of the tree are customizable.

As a new feature, we now support another graph format–flash bar chart. All terms in the three categories of GO are free to select for comparison using this functionality. By default, all detectable child-terms (secondary level terms) of three root terms (GO:0008150 biological process, GO:0003674 molecular function and GO:0005575 cellular component) and significantly over/under-regulated child-terms of secondary level terms (if any) are selected to construct a flash bar chart. Parameters for chart setting are customizable (e.g. legend content, font and rotation, and bar style). For example, the bar graph is resizable by simple dragging of the border, and color of bars is controllable by users. Appropriate adjustment, like terms selections or parameters settings, can generate customizable and artistic outputs, which allow users to make graphs and figures suitable for publication. To demonstrate, we selected all significant terms in the analysis results of Figure1 to generate a flash bar chart, and further adjusted size and color of the chart (Figure 2). Though displayed in a new method, with a similar biological conclusion, that cold and stress related terms are overrepresented, can be gained by using a flash bar char. The text tree mode is another unique way available for result inspection in agriGO (Supplementary Figure S2A). Furthermore, we developed a flexible way that users can freely select terms to create custom outputs (Figure 2 and Supplementary Figure S2B and C). These methods will provide users a comprehensive way to explore the analysis results and multiple choices for generating images suitable for publication.

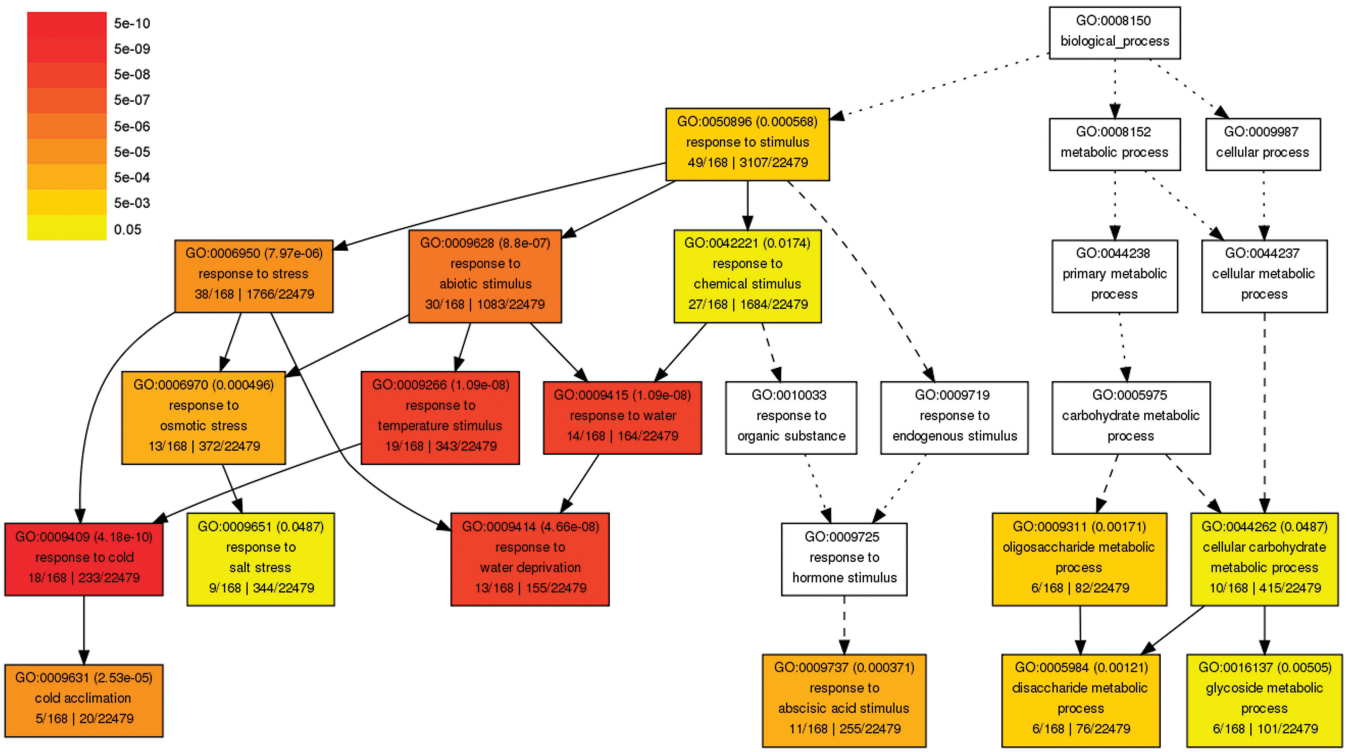


Figure 1. Hierarchical tree graph of overrepresented GO terms in biological process category generated by SEA. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant term (adjusted $P \leq 0.05$) are marked with color, while non-significant terms are shown as white boxes. The diagram, the degree of color saturation of a box is positively correlated to the enrichment level of the term. Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively. The rank direction of the graph is set to from top to bottom.

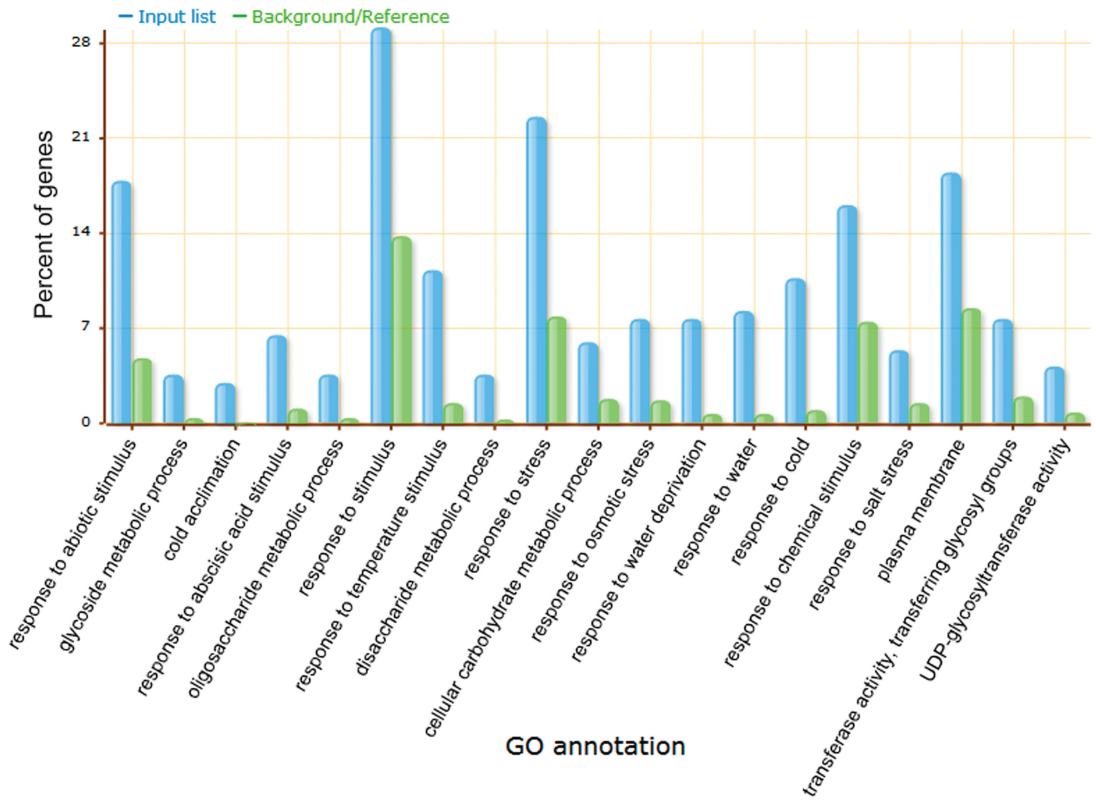


Figure 2. Flash bar chart of overrepresented terms in all three categories. The *Y*-axis is the percentage of genes mapped by the term, and represents the abundance of the GO term. The percentage for the input list is calculated by the number of genes mapped to the GO term divided by the number of all genes in the input list. The same calculation was applied to the reference list to generate its percentage. These two lists are represented using different custom colors. The *X*-axis is the definition of GO terms.

Cross-comparison of analysis results

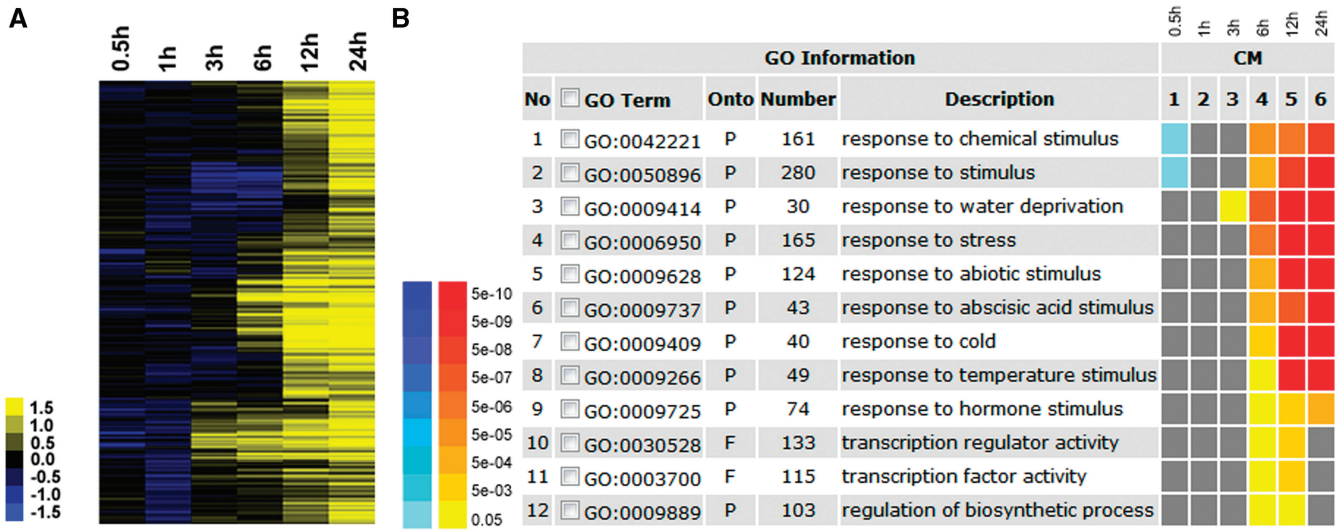
Cross-comparison is essential for interpreting results obtained from experiments involving multiple samples, such as time-series experiments, and this novel functionality is enabled both for SEA and PAGE approaches. Through the SEACOMPARE tool, user can submit multiple SEA job identifiers, and analysis results will be combined for cross-comparison purpose. When using PAGE, user can submit a list of genes with multiple numeric values that were each obtained from separate experiments.

As a test case, we selected a group of 1921 Arabidopsis ATH1 probe sets through hierarchical clustering analysis of the cold-treatment microarray data, from AtGenExpress project (24); (Figure 3A is a heat map representing the clustering result of 1921 probe sets). The $\log_2$ cold/control ratio of these probe sets at six time-points was used as input for PAGE. The results are represented in HTML table mode. For clarification purpose, we selected certain GO terms using the 'suppress GO number' functionality and trimmed out the numerical parts in the image (see the complete snapshot in Supplementary Figure S3). The stress and stimulus-related terms were upregulated and strengthened over time at three later time points (i.e. 6, 12 and 24 h). The transcription factor (i.e. GO:0030528 and GO:0003700) appeared at a relatively early stage (6 h), and were most overrepresented at 12 h, but there were no such activities at the last time point (24 h). Interestingly, two GO terms concerning 'response to stimulus' (GO:0042221 and GO:0050896) were even downregulated at a very early stage (0.5 h). We conclude that comparison can offer users the possibility to quickly and efficiently gather important biological knowledge.

The tree graph and flash bar chart can also be used to do the comparisons. Unfortunately, agriGO can only support mutual comparisons using the tree graph (see Supplementary Figure S4); since when more than two data sets are compared, a much more complex color system will be used to display the terms' changes among different experiments, and this is inconvenient for investigation.

Organisms, identifiers and GO annotation resources

In agriGO, the number of supported organisms and identifiers is substantially increased compared with EasyGO (12). We collected 38 agricultural species including 274 types of corresponding identifiers. The efficiency to map users' input IDs to GO annotation is benefited by the extensive support to different identifier types. Recently released genome sequence data, of which GO annotations are not available in public databases (e.g. tomato and cucumber) are collected and annotated locally, since



	GO Information				CM					
No	GO Term	Onto	Number	Description	1	2	3	4	5	6
1	GO:0042221	P	161	response to chemical stimulus						
2	GO:0050896	P	280	response to stimulus						
3	GO:0009414	P	30	response to water deprivation						
4	GO:0006950	P	165	response to stress						
5	GO:0009628	P	124	response to abiotic stimulus						
6	GO:0009737	P	43	response to abscisic acid stimulus						
7	GO:0009409	P	40	response to cold						
8	GO:0009266	P	49	response to temperature stimulus						
9	GO:0009725	P	74	response to hormone stimulus						
10	GO:0030528	F	133	transcription regulator activity						
11	GO:0003700	F	115	transcription factor activity						
12	GO:0009889	P	103	regulation of biosynthetic process						

Figure 3. Hierarchical clustering of test-sample cluster and cross-comparison of its analysis results by PAGE in HTML table mode. (**A**) Experiments were performed with different cold treatment time (0.5, 1, 3, 6, 12 and 24 h) by AtGenExpress project (24). Probe set signal intensity was computed using RMA (33), and hierarchical clustering based on $\log_2$ Cold/CK ratio of the probe set at each time point was done by Cluster 3.0 (Cluster 3.0, command line version <http://bonsai.ims.u-tokyo.ac.jp/mdehoon/software/cluster>). For the test sample, 1921 probe sets showing coordinated upregulation at later time points of cold treatment were selected. (**B**) The 1921 probe sets in the test sample were analyzed by PAGE, and the comparison is displayed in HTML table mode. The colored blocks represent the level of up/downregulation of each term at a certain time-point. The yellow-to-red, cyan-to-blue and grayscale represent the term is upregulated, downregulated and non-significant, respectively. The adjusted *P*-value of the term determines the degree of color saturation of the corresponding box. Detailed information is provided for each term.

genome-wide data sets can provide completely global perspectives of GO distribution.

The GO annotations in agriGO are either obtained from public databases or produced by computational prediction. We run BLAST (25), Pfam (26) or InterProScan (27) to generate GO annotation for those publicly unsupported identifiers with sequences. Annotations for model organisms are downloaded from publicly databases like TAIR (28), Gramene (15), TIGR (14) or from GO repository server including GOA (29), B2G-FAR (30) and AgBase (31) (see Supplementary Table S1 for detail).

Download, search and BLAST service

Though GO annotations are widely available on the Internet, this is not true for most agricultural species. A GO annotation repository concerning agriculture, AgBase (31), has been established, however, non-model and newly sequenced organisms only have limited support. Therefore, we provide free download and search functions for our annotation data sets as a GO annotation resource for the agricultural research community. Furthermore, we developed a tool called BLAST4ID (Transfer IDs by BLAST) providing a BLAST service, which can be used to do ID mapping for unknown/unannotated identifiers. It can also work as a connection between unidentified IDs and analysis tools in agriGO, for example, users can apply BLAST4ID to generate GO-annotated gene list, and upload the list to do the analysis. However such automatic matching is likely to generate false positives, and thus caution is required when using BLAST4ID.

Interface and usability

The web interface and usability of agriGO has been totally re-engineered. The interactive help system makes agriGO more user-friendly. For example, once the user selects one species, the supported identifier types will be displayed to help users to judge whether their identifiers can be submitted directly or need to be transformed using BLAST4ID. In addition, the identifiers can be automatically recognized without further efforts so that different types of identifiers from one species can be used for one analysis.

Implementation and update

We constructed and configured agriGO upon a typical LAMP (Linux + Apache + MySQL + PHP) platform. Data set was stored in MySQL 5.0 (www.mysql.com), and the web interface was built by PHP scripts (www.php.net) on Red Hat Linux, powered by an Apache server (www.apache.org). Server-side scripts were developed using Python (www.python.org). The hierarchical tree images were generated using Graphviz software (www.graphviz.org) and the flash bar charts were achieved by Open Flash Chart software (http://teethgrinder.co.uk/open-flash-chart-2). The tool is web-based, and no software or plug-in installation effort is required to use it.

We perform regular updates and maintenance to agriGO. As most of the annotation and sequence data is obtained from publicly available databases, manual effects and Python scripts are used to semi-automatically oversee and download source files, to ensure agriGO provides the

most up-to-date data. New agricultural species and functions can be added upon request.

DISCUSSION

One goal of developing agriGO is to provide EasyGO users better service, and the consistency of analysis conclusions is important. We tested agriGO and EasyGO with the same data set (Figure 1 and Supplementary Figure S1), and the conclusions were similar with slight differences caused by updated annotation in agriGO. However, because of different software architectures, EasyGO and agriGO are not compatible in results exploration, i.e. users can not inspect the results generated by EasyGO in agriGO, and vice versa.

One issue is that a lot of GO terms may be detected in the analysis results, which will cause inconvenience in exploration and explanation of the graphical results. To avoid this issue, 'GO slim limitation' in the advanced options can be selected. Alternatively, users can produce custom graph results using custom settings of GO terms (Supplementary Figure S2).

GO annotation coverage is another critical question for GO functional analysis tools including agriGO, as discussed by Berg *et al*. (13). Except for well-studied model organisms like Arabidopsis, GO annotations are mainly generated by computational prediction. Such prediction may lead to two issues: reduced quality of annotation and low annotation coverage. Poor-quality annotation can directly affect the GO distribution and, if not prepared cautiously, can generate biased or misleading analysis results. One issue is that by using a single BLAST search, even with high BLAST scores, it is not guaranteed that sequences will be annotated correctly, thus users should be alert when using BLAST4ID. Effective annotation will be hampered by some sequences that have neither high similarity to already known sequences (for BLAST search) nor sequence signatures (for tools based on pattern recognition). Empirically, automatic annotation methods, e.g. the combination of BLAST (E-value $\leq$ 1e-30 and Coverage ≥ 0.7) and InterproScan (E-value $\leq$ 1e-3), can only annotate $\sim$60% of all protein sequences predicted from one newly sequenced genome by GO. One promising way to overcome these problems is to use a similar annotation strategy to Meng *et al*. (32) by performing comprehensive manual curation. However, such a great workload seems unrealistic for most GO enrichment tools as they may maintain dozens of species. A good approach or resource to generate high-quality GO annotation data for non-model organisms is greatly needed.

Compared to EasyGO, agriGO offers vast improvements; the functionalities have been carefully tested and it has completed >4800 analysis requests since its release. We believe that agriGO will facilitate researchers in the agricultural community to extract biological meanings from data of high-throughput experiments in an easy and systematic way. This new application is freely accessible now at http://bioinfo.cau.edu.cn/agriGO/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Ms Wenying Xu for discussions and critical suggestions. The authors thank Yan Zhang for discussion on logo design. The authors thank anonymous reviewers for their valuable contributions and comments on an earlier version of this article.

FUNDING

Funding for open access: Ministry of Science and Technology of China (90817006 and 2006CB100105).

REFERENCES

1. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
2. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
3. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.*, **25**, 25–29.
4. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
5. Conesa,A., Gotz,S., Garcia-Gomez,J.M., Terol,J., Talon,M. and Robles,M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
6. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
7. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Res.*, **31**, 3775–3781.
8. Kim,S.B., Yang,S., Kim,S.K., Kim,S.C., Woo,H.G., Volsky,D.J., Kim,S.Y. and Chu,I.S. (2007) GAzer: gene set analyzer. *Bioinformatics*, **23**, 1697–1699.
9. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
10. Nogales-Cadenas,R., Carmona-Saez,P., Vazquez,M., Vicente,C., Yang,X., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, **37**, W317–W322.
11. Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.
12. Zhou,X. and Su,Z. (2007) EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, **8**, 246.
13. van den Berg,B.H., Thanthiriwatte,C., Manda,P. and Bridges,S.M. (2009) Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data. *BMC Bioinformatics*, **10(Suppl 11)**, S9.
14. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al*. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.

15. Liang,C., Jaiswal,P., Hebbard,C., Avraham,S., Buckler,E.S., Casstevens,T., Hurwitz,B., McCouch,S., Ni,J., Pujar,A. *et al.* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.*, **36**, D947–D953.
16. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
17. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annal. Stat.*, **29**, 1165–1188.
18. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57**, 289–300.
19. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
20. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
21. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
22. Rice,J.A. (1995) *Mathematical Statistics and Data Analysis*, 2nd edn. Duxbury Press, Pacific Grove, CA.
23. Team,R.D.C. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
24. Kilian,J., Whitehead,D., Horak,J., Wanke,D., Weinl,S., Batistic,O., D'Angelo,C., Bornberg-Bauer,E., Kudla,J. and Harter,K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J.*, **50**, 347–363.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
27. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
28. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
29. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
30. Gotz,S., Garcia-Gomez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talon,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.
31. McCarthy,F.M., Bridges,S.M., Wang,N., Magee,G.B., Williams,W.P., Luthe,D.S. and Burgess,S.C. (2007) AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res.*, **35**, D599–D603.
32. Meng,S., Brown,D.E., Ebbole,D.J., Torto-Alalibo,T., Oh,Y.Y., Deng,J., Mitchell,T.K. and Dean,R.A. (2009) Gene Ontology annotation of the rice blast fungus, Magnaporthe oryzae. *BMC Microbiol.*, **9(Suppl 1)**, S8.
33. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.