Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W293–W297***
doi:10.1093/nar/gkl031

WEGO: a web tool for plotting GO annotations

Jia Ye[1], Lin Fang[2], Hongkun Zheng[2], Yong Zhang[2,3], Jie Chen[2], Zengjin Zhang[2], Jing Wang[2], Shengting Li[2,4], Ruiqiang Li[2,5], Lars Bolund[2,4] and Jun Wang[1–5,*]

[1]James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou 310008, China, [2]Beijing Genomics Institute, Beijing 101300, China, [3]College of Life Sciences, Peking University, Beijing 100871, China, [4]The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark and [5]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark

Received October 21, 2005; Revised and Accepted November 29, 2005

ABSTRACT

Unified, structured vocabularies and classifications freely provided by the Gene Ontology (GO) Consortium are widely accepted in most of the large scale gene annotation projects. Consequently, many tools have been created for use with the GO ontologies. WEGO (Web Gene Ontology Annotation Plot) is a simple but useful tool for visualizing, comparing and plotting GO annotation results. Different from other commercial software for creating chart, WEGO is designed to deal with the directed acyclic graph structure of GO to facilitate histogram creation of GO annotation results. WEGO has been used widely in many important biological research projects, such as the rice genome project and the silkworm genome project. It has become one of the daily tools for downstream gene annotation analysis, especially when performing comparative genomics tasks. WEGO, along with the two other tools, namely External to GO Query and GO Archive Query, are freely available for all users at http://wego.genomics.org.cn. There are two available mirror sites at http://wego2.genomics.org.cn and http://wego.genomics.com.cn. Any suggestions are welcome at wego@genomics.org.cn.

INTRODUCTION

Unified, structured vocabularies and classifications freely provided by the Gene Ontology (GO) Consortium (http://www.geneontology.org/) are widely accepted in most of the large scale gene annotation projects. Three ontologies (molecular function, biological process and cellular component) were developed to represent common and basic biological information in annotation. Not only the original organizations SGD (*Saccharomyces* Genome Database), FlyBase and MGD (Mouse Genome Database), but also some additional model organism database groups are involved in the project, including TAIR (The Arabidopsis Information Resource), WormBase, RGD (Rat Genome Database), TIGR and so on (1–3).

It is not easy, however, for a biologist with little computer background to analyze and understand genes with the GO information. The difficulties may have two aspects: (i) how to annotate the anonymous sequences with the GO vocabularies, and (ii) how to find the differences or anything new in the dataset. Many tools and software programs have been developed to tackle the first problem through an automatically or manually curated search for the associations between GO terms and genes (4–8). The Web Gene Ontology Annotation Plot (WEGO) is therefore designed as a web application mainly to deal with the second problem. The main purpose of the WEGO is to visualize the annotation of sets of genes, comparing the provided gene datasets and plotting the distribution of GO annotation results into a histogram. General histograms could be drawn by many commercial software programs. However, the GO terms are structured in the form of directed acyclic graph (DAG) to represent a network of complex relationships of 'child' and 'parent' (1). In order to avoid the tedious task of plotting the distribution of GO annotations, WEGO presents the DAG structures of ontologies as hierarchical trees to help users easily choose the levels and GO terms for exhibition.

WEGO is not the only software to address this problem nor is it the most powerful one (9–13), but it is an excellent tool in several aspects. First, it is very user-friendly. For example, biologists could use the output result of InterProScan (http://www.ebi.ac.uk/InterProScan/) as the input data of WEGO without any conversion. Second, WEGO is a web server

*To whom correspondence should be addressed. Tel: +86 10 80491664; Fax: +86 10 80498676; Email: wangj@genomics.org.cn
Correspondence may also be addressed to Lars Bolund. Tel: +45 89421675; Fax: +45 86123173; Email: bolund@humgen.au.dk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

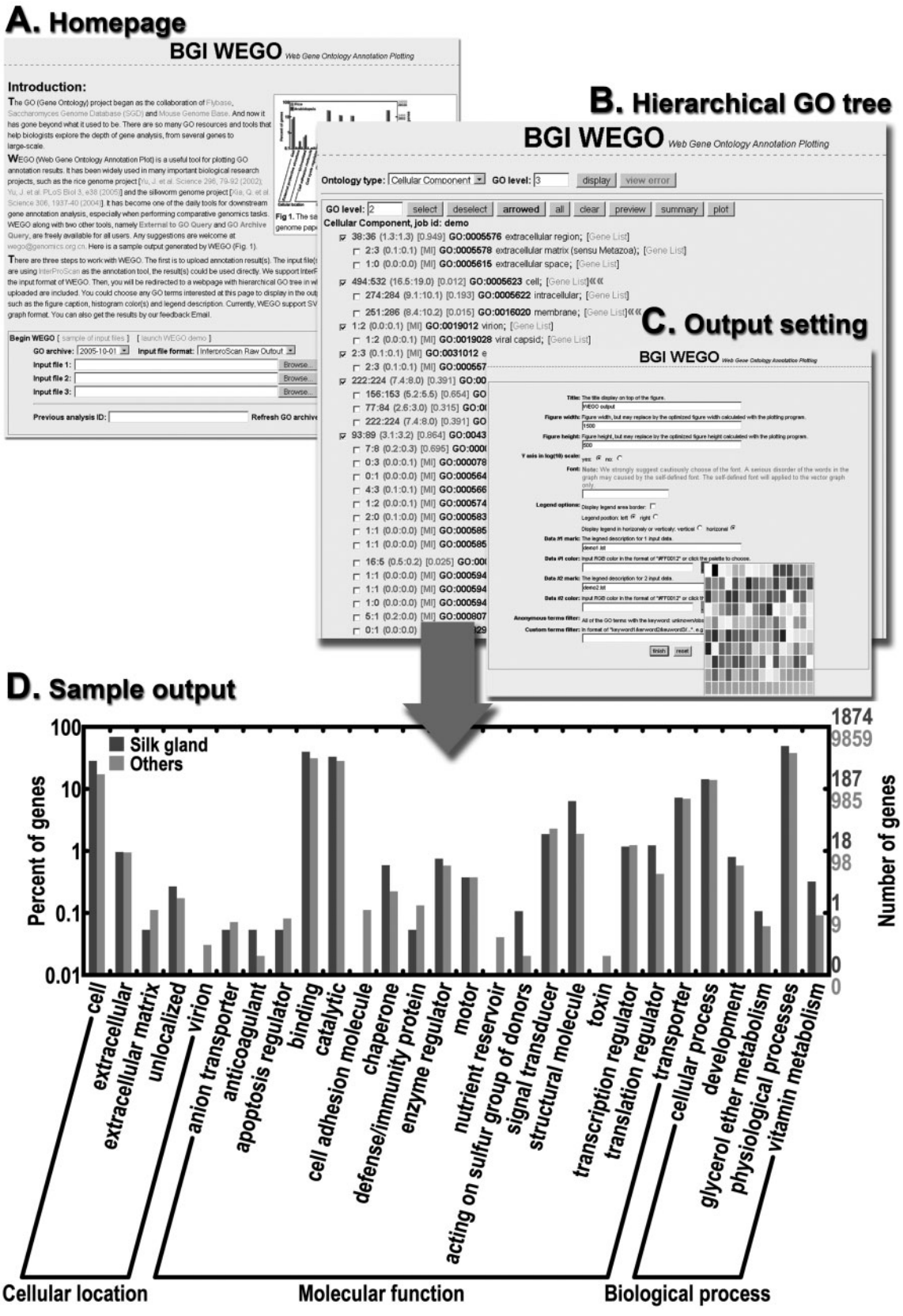

Figure 1. WEGO interfaces. (**A**–**C**) Shows a screenshot montage of the WEGO interface of the three steps of the WEGO procedure: annotation results uploading, hierarchical GO tree editing, output setting. As an example, (**D**) is a sample figure from the analysis of silkworm draft sequences to show how WEGO can help analyze and compare the annotation results. In this histogram, EST-confirmed genes in silk gland are compared with 11 other libraries. Significant differences are obvious in several categories.

that avoids the tedious steps of application installation and testing. It is operating system independent as well. Third, WEGO provides a visualization of the annotation results. It is not only useful for customizing output but is also effective for the understanding of GO annotations. In addition, WEGO does not have the restriction of organism. Finally, WEGO supports the comparison between several gene datasets which is a key characteristic in the post-genomic era.

WEGO has been applied in many important biological research studies, such as the comparative genomics study between the rice genome and the Arabidopsis genome (14,15) and the silkworm genome analysis (16). It has become one of the daily tools for downstream gene annotation analysis, especially when performing comparative genomics tasks. As an example, Figure1.D, which is from the analysis of silkworm draft sequences, illustrates how WEGO can help analyze and compare the annotation results. In this histogram, significant differences in several categories are clearly presented by comparison between expressed sequence tag (EST)-confirmed genes in silk gland and other libraries.

DESCRIPTION OF THE WEB INTERFACE

The web interface of WEGO is based on common gateway interface (CGI) and scalable vector graphics (SVG) technologies. It is implemented by Perl language. There are three freely accessible tools through the web interface: WEGO, External to GO Query and GO Archive Query. The GO data, dated from April 1, 2001, is downloaded from the GO FTP archive and is updated monthly (ftp://ftp.geneontology.org/pub/go/ontology-archive/).

WEGO

Input of WEGO. Currently, WEGO supports four kinds of input format: WEGO native format, InterProScan raw (our default input format), text and XML output formats. The '-goterms' option should be switched on for corresponding GO annotations when performing the InterProScan. WEGO native format is a simple text file with one gene record per line. Each column is tab delimited. The first column is the gene name and the rest are the associated GO IDs.

The InterProScan output formats are acceptable for the convenience of the user, so that the annotation results of InterProScan could be uploaded onto the WEGO without any conversion. We are planning to support more output formats from other GO annotation tools in the near future.

Uses of WEGO. There are two ways to work with WEGO. The first is to upload the annotation files (up to three files at one time). The input files must be in one of the four formats described above. The version of GO archive used for the downstream analysis of the GO annotation results in WEGO should of course be the same as the one used in annotation. Therefore, it is optional in WEGO when uploading the input files. The second way is to simply enter the job ID

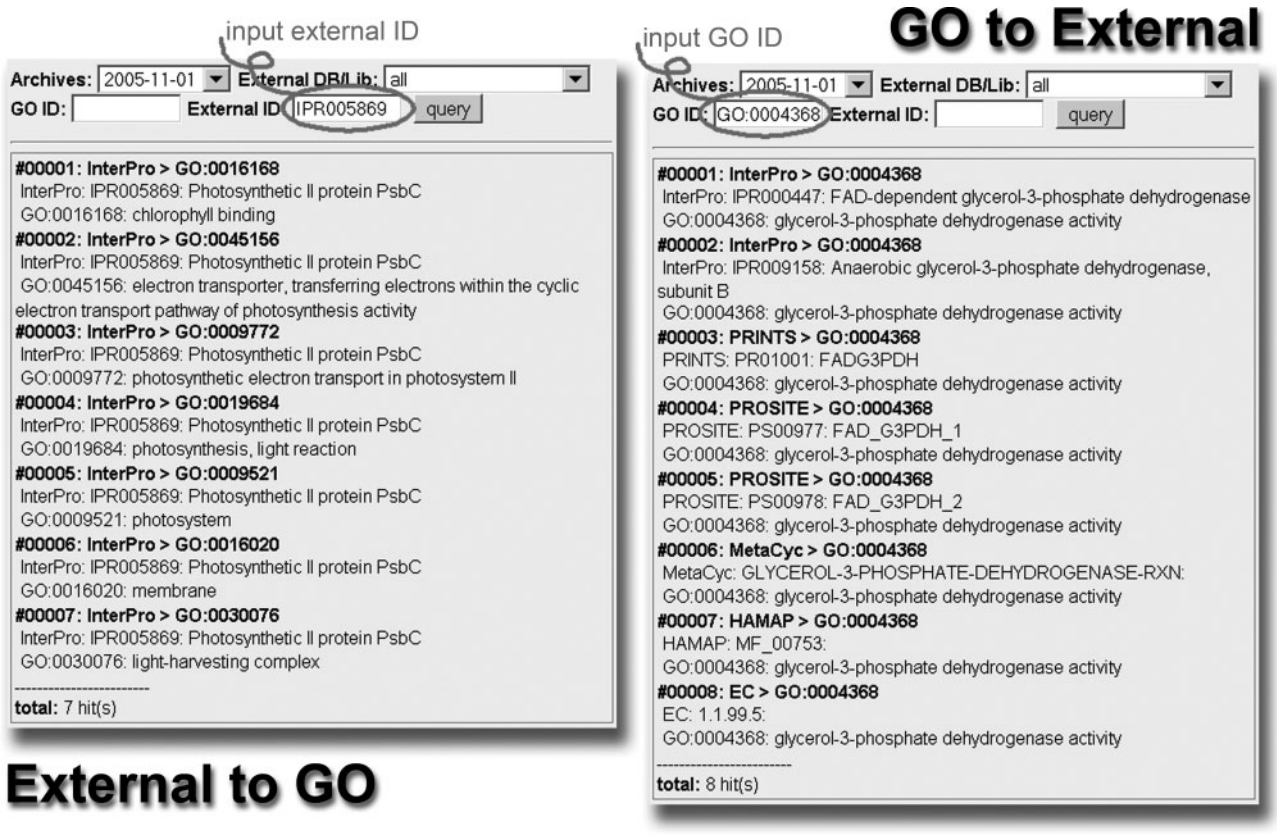

Figure 2. External to GO Query. Screen capture from the External to GO Query, which attempts to make translations between other categories and GO. Users could query both GO ID and entries of external systems by External to GO Query. The complex relationships among the external catalogs are not in the consideration of External to GO Query, so if the entry of external database is queried, only the associated GO terms will be returned.

if the user carried out a WEGO analysis within the previous three days.

A process window shows the job ID after the file is uploaded. Then the user is redirected to a webpage with a hierarchical GO tree which includes all the GO terms contained in the uploaded files. The displayed level of GO tree and the selected GO terms both could be changed by the user. The GO terms that were not contained in the chosen GO archive are listed in the 'view error' page. This error occurs frequently due to the different versions of GO archive used in annotation and WEGO. Another tool, named GO Archive Query, was developed to help users (especially the ones without information of the GO version used in annotation) deal with this problem.

The user could switch between the three ontology trees to choose any GO terms of interest to display in the output histogram. The gene number, percentages and *P*-value of Pearson Chi-square test of each GO term are listed in the same line. The Pearson Chi-Square test is applied to indicate significant relationships between two input datasets. Compared with the Fisher's exact test, the Pearson Chi-Square test is appropriate and efficient for 2×2 matrixes if all the expected counts are greater than 5. Red arrows are used to indicate remarkable relationships with the significant level of 5%. The 'Gene List' function presents all the gene names under special GO term in XML format, so that users can get the gene content of each branch on the GO tree as well as gene number.

Most of the users choose the GO term by the tree level setting, which may result in many GO terms with no exact meaning included. The anonymous terms filter was designed to avoid the useless items. Only two keywords 'unknown' and 'obsolete' have currently been adopted. There is also a custom terms filter, which allows the user to define the filter's keywords. All the GO terms including these keywords will be dropped from the output histogram by the filter. Alternatively, users could use the specially designed function 'arrowed' to select all the independent nodes to present all significant differences between his or her input datasets.

Output of WEGO. SVG is the default output format of WEGO, since it is widely supported by many industrial and open source software programs, such as CorelDRAW®, Illustrilator®, inkscape and ImageMagick. With the help of the SVG plug-in, SVG could be viewed in the browser. Another advantage of SVG is its easy conversion to other graph formats and its suitability for publishing. WEGO also supports other common graph formats, including the bitmap formats PNG, JPEG and GIF, suitable for on-screen display, and the other vector formats PostScript and EPS. The output file will be compressed for downloading and the user could also supply an email address to receive results.

Two associated tools

External to GO Query. The structured vocabularies and classifications of GO are now accepted widely. However, GO is not the only attempt to build structured vocabularies for genome annotation. A series of other catalogs are also in current use, such as EC (Enzyme Commission), Swiss_Prot and Pfam domains. The External to GO Query attempts to make translations between these categories and GO terms. It is an interface based on the database of the GO Consortium's external2go (ftp://ftp.geneontology.org/pub/go/external2go/). Users can query both GO ID and entries of external systems by External to GO Query. Corresponding entries or GO ID will be given as output (Figure 2). Compared with the QuickGO (17,18), which was developed by the GOA (Gene Ontology Annotation project), the External to GO Query is a simpler but handier tool. The External to GO Query is designed to help biologists better understand the annotation results even though these mappings are not currently complete or exact.

Figure 3. GO Archive Query. GO Archive Query provides the interface that allows users to query GO ID in the format of GO:0001955, 0001955 or just 1955. All the versions of GO repositories containing the GO ID will be presented. It is helpful for users choosing the correct version or at least a similar version of GO repository to use.

GO Archive Query. As the GO terms, definitions and ontologies are frequently updated, it is important to choose the correct version of GO archive. The version of GO used in the analysis should be the same as the one used in annotation. As stated above, the choice is difficult for the users without any information of the version of GO archive used in the annotation. Consequently, another tool, GO Archive Query, was developed to help users to solve this problem. Users could query GO ID, especially the GO ID from the 'view error', at which point the user is presented with all the versions of GO archives containing the GO ID and can choose the correct or close version of GO archive (Figure 3).

AVAILABILITY AND PROSPECTS

WEGO, along with the two other tools, namely External to GO Query and GO Archive Query, are freely available for all users

at http://wego.genomics.org.cn. There are two available mirror sites at http://wego2.genomics.org.cn and http://wego.genomics.com.cn. It is operating system independent, and has been tested on Mozilla/Netscape/Firefox, Opera, Galeon and Internet Explorer. An SVG plug-in is necessary for online preview of the figure.

Aiming for the greatest ease of use for biologists, especially for those without computer background, we are trying to develop the WEGO to serve as a GO-application-friendly tool as well as a user-friendly tool. Additional output formats of other GO annotation tools will be adaptable as the WEGO input. And more output choices and better integration with other GO tools will be future features of WEGO.

ACKNOWLEDGEMENTS

We would like to thank Patrick Henry and Su Xu for correcting the English of this manuscript. We would also like to sincerely thank our colleagues at the Beijing Genomics Institute for collaboration and data testing. This work is supported by grants from Ministry of Science and Technology (2002AA104250, CNGI-04-15-7A), National Natural Science Foundation of China (30399120, 90208019, 30200163, 90403130), Zhejiang University, and Chinese Academy of Sciences. Additional funding came from Danish Basic Research Foundation (Danish Platform for Integrative Biology). Funding to pay the Open Access publication charges for this article was provided by National Natural Science Foundation of China (30200163).

Conflict of interest statement. None declared.

REFERENCES

1. 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. *Nature Genet*., **25**, 25–29.
2. The Gene Ontology Consortium. (2001), Creating the gene ontology resource: design and implementation. *Genome Res*., **11**1425–1433.
3. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
4. Khan,S., Situ,G., Decker,K. and Schmidt,C.J. (2003) GoFigure: automated Gene Ontology annotation. *Bioinformatics*, **19**, 2484–2485.
5. Martin,D.M., Berriman,M. and Barton,G.J. (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, **5**, 178.
6. Hennig,S., Groth,D. and Lehrach,H. (2003) Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Res*., **31**, 3712–3715.
7. Zehetner,G. (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res*., **31**, 3799–3803.
8. Groth,D., Lehrach,H. and Hennig,S. (2004) GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res*., **32**, W313–W317.
9. Young,A., Whitehouse,N., Cho,J. and Shaw,C. (2005) OntologyTraverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.
10. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
11. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*., **5**, R101.
12. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontologyhierarchies. *BMC Bioinformatics*, **5**, 16.
13. Lee,J.S., Katari,G. and Sachidanandam,R. (2005) GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, **6**, 189.
14. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
15. Yu,J., Wang,J., Lin,W., Li,S., Li,H., Zhou,J., Ni,P., Dong,W., Hu,S., Zeng,C. *et al*. (2005) The Genomes of Oryza sativa: a history of duplications. *PLoS Biol*., **3**, e38.
16. Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T., Chai,C. *et al*. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
17. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al*. (2003) The Gene Ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*., **13**, 662–672.
18. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.