Tutorial section

Get ready to GO! A biologist's guide to the Gene Ontology

Abstract

The Gene Ontology (GO) project provides a controlled vocabulary to facilitate high-quality functional gene annotation for all species. Genes in biological databases are linked to GO terms, allowing biologists to ask questions about gene function in a manner independent of species. This tutorial provides an introduction for biologists to the GO resources and covers three of the most common methods of querying GO: by individual gene, by gene function and by using a list of genes. [For the sake of brevity, the term 'gene' is used throughout this paper to refer to genes and their products (proteins and RNAs). GO annotations are always based on the characteristics of gene products, even though it may be the gene that is cited in the annotation.]

Keywords: *Gene Ontology, microarray, gene annotation, model organism database, controlled vocabulary*

Functional annotation of genes is a way of capturing, usually in some shorthand or tabulated form, what is known about that gene. These annotations provide an ever-growing wealth of information for biologists, but with different species being annotated by different databases, how do we best harness and use that information? By making annotations to a common, shared set of vocabularies, the Gene Ontology (GO) resource provides a powerful way to capture, query and analyse this information in a way that is independent of species.[1–5] At present over 87,000 species have some GO annotation, comprising over 6.8 million annotations.[6]

There are three GO vocabularies, each providing a specific type of information about a gene or protein: (i) its molecular function, (ii) the broader biological processes it is involved in and (iii) the cellular compartment it acts in. The GO vocabularies are constantly updated, with new terms and relationships being added by curators in consultation with biological experts to reflect current knowledge of biology. For more information about the structure of the GO vocabularies, see Gene Ontology Consortium,[1–3] Harris *et al.*[4] and Clark *et al.*[5]

A GO annotation includes four essential pieces of information: the gene identifier, the GO term or terms it is associated with, the type of evidence supporting the association and a reference for that evidence. For example, if a protein was experimentally determined to have alcohol dehydrogenase activity in an enzyme assay, the protein would be annotated with the GO term 'alcohol dehydrogenase activity' with the evidence 'inferred by direct assay' (abbreviated as IDA), and the annotation would also include the literature reference for the paper describing the assay.

The evidence type provides a level of confidence or quality for annotations, and is crucial in their interpretation. Broadly, the evidence codes 'inferred by electronic annotation' (IEA), 'reviewed computational annotation' (RCA) and 'inferred by sequence similarity' (ISS) are the least precise evidence codes. Although these annotations have been shown to be accurate (91–100 per cent[7]), they tend to use less specific GO terms than manual annotation, and should be used with that

limitation in mind. The evidence types 'traceable author statement' (TAS) and 'non-traceable author statement' (NAS) refer to remarks made by authors in the literature where the data are not shown, and the code 'inferred by curator' (IC) is used where a curator has deduced the annotation from other available data. Several evidence codes, including IDA, refer to experimental data and as such can generally be used with the most confidence, depending on the experimental technique. A full description of the evidence types used in GO can be found at the GO website.[8]

GO annotation does not capture all known information about a gene; some information – such as phenotype, anatomy and pathways – can be obtained only from other sources such as the individual model organism databases, eg Mouse Genome Informatics (MGI),[9] *Saccharomyces* Genome Database (SGD)[10] and FlyBase,[11] or protein databases such as UniProt.[12] Pathway information can also be obtained from multi-species resources such as Reactome[13] and MetaCyc.[14]

This tutorial aims to provide an introduction for biologists to the GO resources. It is organised by some common ways you may want to query GO: by gene list, by individual genes and by GO term.

QUERYING GO WITH A GENE LIST

If you have done an expression experiment such as a microarray, or some other genome-scale experiment, you may want to query GO with a list of genes.

Various tools have been developed by groups external to the GO Consortium to perform this sort of analysis, such as GoMiner (Figure 1),[15] Onto-Express[16] and GO Term Finder[17] (for a full list see the GO website[18]). These tools work in a similar way: you upload your full gene set, and a list of all 'interesting' genes within that set, usually those that have been up- or down-regulated in an expression experiment. The tool then allows you to view which GO categories have been enriched for your genes of interest, and usually provides some sort of

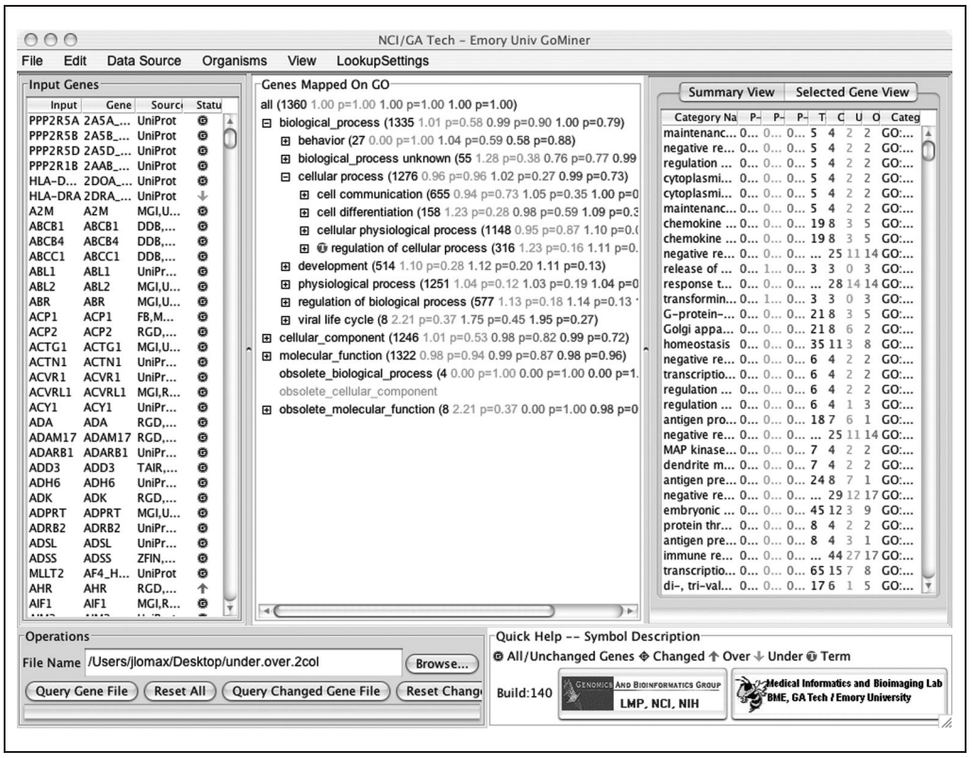

Figure 1: A screen shot of GoMiner displaying a tree view of GO terms with enrichment for a set of human genes

statistical measure to guard against GO categories that appear by chance alone. For information on interconverting gene identifiers from different databases, see Box A.

GO SLIMS

A useful way that you can obtain a high-level GO categories for a given gene list or annotation set is by using GO slims. GO slims are reduced sets of GO terms that provide a 'bird's eye' view of a given gene or annotation set. This can be particularly powerful when you want to display something like the functional annotation for a whole genome (Figure 2[24–26]). Several GO slims are currently available from GO, including a generic slim for all species and GO slims tailored to plant and yeast.[27] It is also possible to create your own GO slim using OBO-Edit, the ontology editor developed by the GO consortium (the GO website[28] provides instructions).

To group your genes into the GO categories listed in your GO slim, the GO Consortium provides a script, map2slim.pl.[29] The input for this script is an annotation file in the Gene Ontology format,[30] and the output will be an annotation file with the genes grouped into the GO slim categories.

QUERYING BY GO TERM

Browsing through GO terms, for example biological processes, and finding the genes and proteins associated with them, can be a useful way for you to learn more about specific areas of biology. If you want to query GO in this way, the best resource is the GO browser AmiGO.[31] Several other GO browsers are also available, including the EBI browser QuickGO.[32] The GO website[33] gives a full list of GO browsers.

AmiGO displays the GO terms as a tree (Figure 3); the branches of the tree can be expanded by clicking the + icon next to the terms. AmiGO also allows you to search for a particular GO term using the search box on the left (Figure 4).

Box A: Interconverting gene identifiers

In functional genomics, the interconversion of identifiers from different systems and databases can pose a problem. GO uses database identifiers from the individual databases that provide the GO annotations, for example UniProt identifiers, FlyBase identifiers and Mouse Genome Informatics (MGI) identifiers, but there are several possible ways to query GO with identifiers from other systems.

UniProt provides conversion tables for *Arabidopsis*, human, mouse, rat and zebra fish UniProt protein identifiers to a range of identifiers in other systems,[19] including the EMBL/Genbank/DDBJ nucleotide sequence databases, HUGO, and Entrez Gene and RefSeq at NCBI.[20] In addition, for the individual model organism databases, translation files are available for the model organism database identifier to a UniProt protein identifier.[21]

There are also tools available that you can use to automatically map identifiers from one system to another, mainly designed for use in expression analysis. *Drosophila*, human, mouse and rat gene identifiers of a range of different types – GenBank accession numbers, NCBI gi numbers and Affymetrix probe identifiers – can be interconverted with Onto-Translate,[16] and MatchMiner[22] provides a similar resource for mapping human and mouse IDs.

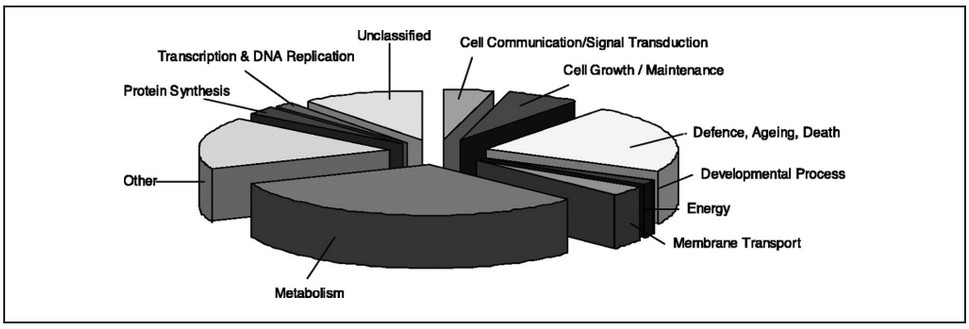

Figure 2: *Oryza sativa* genome distribution among GO slim categories[23]
Reprinted with permission from Goff, S. A. *et al.* (2002), 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)', *Science*, Vol. 296, pp. 92–100. Copyright 2002 AAAS

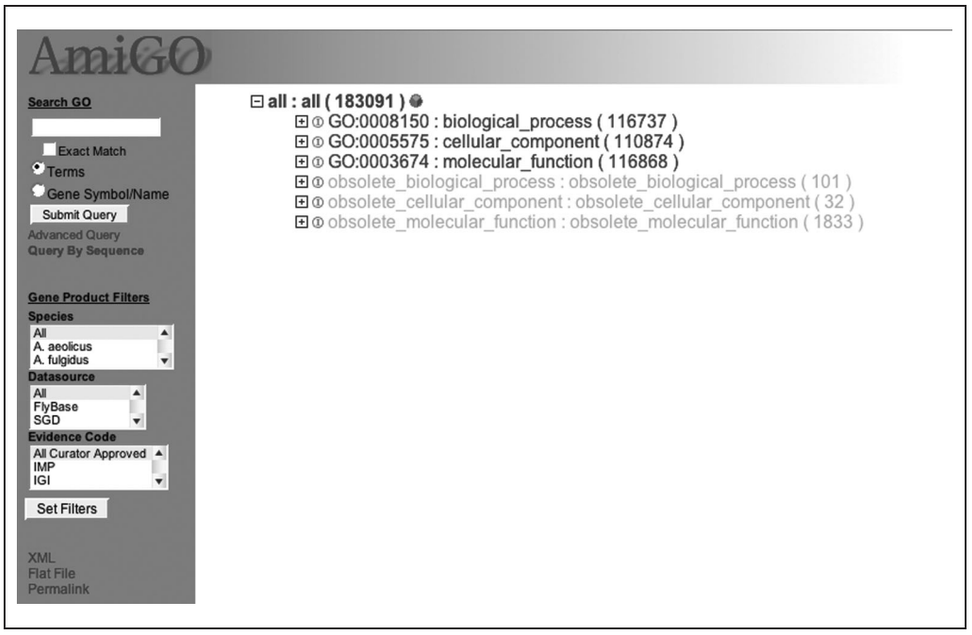

Figure 3: AmiGO treeview

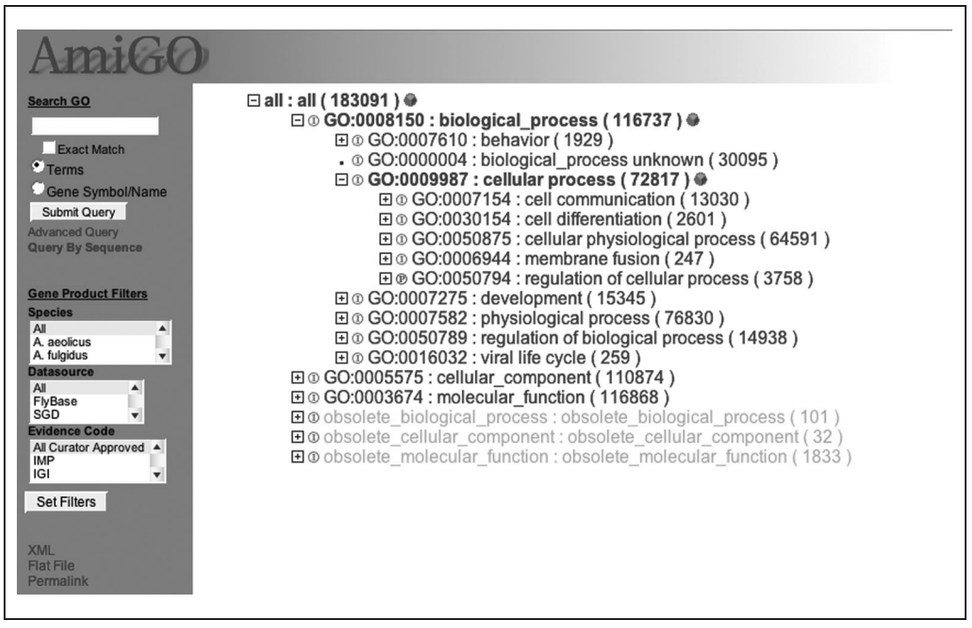

Figure 4: AmiGO treeview expanded

The number at the end of each term is the number of genes annotated to that GO term and its child terms. Note that AmiGO displays only curator-reviewed annotations, and so excludes all annotations using the IEA evidence code. The associated genes are listed in a detailed term view when the term is clicked (Figure 5).

QUERYING GO FOR ONE OR MORE INDIVIDUAL GENES

If you wish to query GO for detailed information on individual or small groups of genes, AmiGO also allows you to search by gene product. AmiGO displays all of the GO terms with which a gene is associated, from all species (Figure 6). To

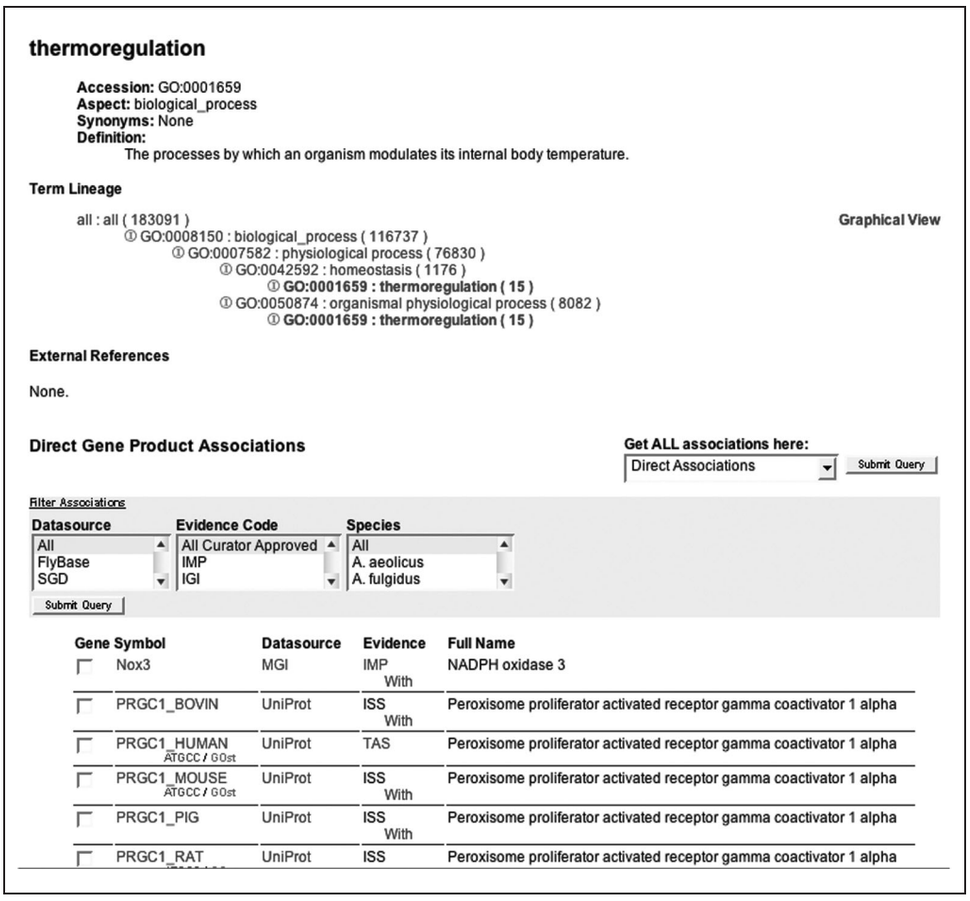

Figure 5: AmiGO detailed view for GO term 'thermoregulation'

see the annotations for just one or a few species, use the species filter settings in the left-hand panel.

It must be remembered, however, that the Gene Ontology project is a work in progress and that if you cannot find any information on your favourite gene, it is probably because it has not been annotated yet. The current status of genes annotated for the major model organism species is shown in Figure 7.

FUTURE DIRECTIONS

One of the continuing aims of the GO project is to encourage contributions from the scientific community, both to increase the number of annotations and the breadth of species they cover, and to improve the quality of the GO vocabularies. To submit annotations, contact the most relevant GO Consortium database (for a list see the GO website[34]). To contribute to ontology development, you can join one of the GO interest groups[35] or submit specific requests for changes to the ontologies to our curator requests tracker[36] (the website[37] provides help with submissions).

The GO Consortium also plans to develop methods to ensure consistency of GO annotation across species and between different annotators, improving the utility of GO annotations. Other plans include aligning and integrating the GO vocabularies with external vocabularies from the Open Biomedical Ontologies (OBO),[38] in areas such as chemicals and cell types; this will extend our ability to reason over the vocabularies and facilitate error checking, automatic addition of new GO terms, and other improvements.

Acknowledgments

Many thanks to Midori Harris, Amelia Ireland, Valerie Wood and Judith Blake for useful

AmiGO

Gene Product: BRCA1_HUMAN

Full Name: Breast cancer type 1 susceptibility protein
Synonyms: IPI00218982
Datasource: UniProt

Associated to Terms:

Evidence	Term	Aspect
TAS	apoptosis	biological_process
NR	damaged DNA binding	molecular_function
TAS	DNA binding	molecular_function
TAS	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	biological_process
TAS	DNA repair	biological_process
TAS	extracellular space	cellular_component
NAS	gamma-tubulin ring complex	cellular_component
NAS	negative regulation of centriole replication	biological_process
TAS	nucleus	cellular_component
NAS	positive regulation of DNA repair	biological_process
TAS	protein binding	molecular_function
TAS	regulation of apoptosis	biological_process
TAS	regulation of cell proliferation	biological_process
TAS	regulation of DNA repair	biological_process
TAS	regulation of transcription from Pol II promoter	biological_process
TAS	regulation of transcription from Pol III promoter	biological_process
TAS	response to DNA damage stimulus	biological_process
TAS	transcription coactivator activity	molecular_function
TAS	transcription factor complex	cellular_component
TAS	transcriptional activator activity	molecular_function
NAS	tubulin binding	molecular_function

Figure 6: AmiGO detailed view of human BRCA1 (breast cancer type 1 susceptibility protein) annotations

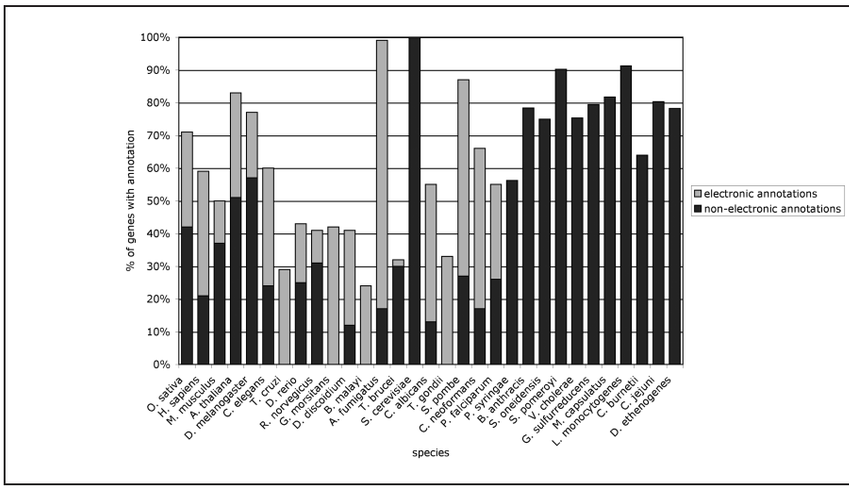

Figure 7: Percentage of genes with some GO annotation in the major model organism species, based on estimated total gene numbers in these species

comments on the manuscript, to Jennifer Clark for help with figures and to David Hill and Mike Cherry for providing data on GO annotations.

Jane Lomax
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Cambridge CB10 1SD, UK
E-mail: jane@ebi.ac.uk
The GO Consortium

References

1. The Gene Ontology Consortium (2004), 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Res.*, Vol. 32, pp. D258–261.
2. The Gene Ontology Consortium (2001), 'Creating the Gene Ontology resource: Design and implementation', *Genome Res.*, Vol. 11, pp. 1425–1433.
3. The Gene Ontology Consortium (2000),

'Gene Ontology: Tool for the unification of biology', *Nat. Genet.*, Vol. 25, pp. 25–29.

4. Harris, M. A., Lomax, J., Ireland, A. and Clark J. I. (2005), 'The Gene Ontology Project', in Subramaniam, S., Ed., 'Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, Part 4. Bioinformatics, 4.7. Structuring and Integrating Data', John Wiley & Sons, Ltd.
5. Clark, J. I., Brooksbank, C. and Lomax, J. (2005), 'It's all GO for plant scientists', *Plant Physiol.*, Vol. 138 (3), pp. 1268–1279.
6. URL: http://www.geneontology.org/GO.current.annotations.shtml
7. Camon, E. B., Barrell, D. G., Dimmer, E. C. *et al.* (2005), 'An evaluation of GO annotation retrieval for BioCreAtIvE and GOA', *BMC Bioinformatics*, Vol. 6 (Suppl 1), p. S17.
8. URL: http://www.geneontology.org/GO.evidence.shtml
9. Blake, J. A., Richardson, J. E., Bult, C. J. *et al.* (2003), 'MGD: The Mouse Genome Database', *Nucleic Acids Res.*, Vol. 31, pp. 193–195.
10. Dwight, S. S., Balakrishnan, R., Christie, K. R. *et al.* (2004), '*Saccharomyces* genome database: Underlying principles and organisation', *Brief. Bioinformatics*, Vol. 5(1), pp. 9–22.
11. The FlyBase Consortium (2003). 'The FlyBase database of the *Drosophila* genome projects and community literature', *Nucleic Acids Res.*, Vol. 31, pp. 172–175.
12. Bairoch, A., Apweiler, R., Wu, C. H. *et al.* (2005), 'The Universal Protein Resource (UniProt)', *Nucleic Acids Res.*, Vol. 33, pp. D154–159.
13. Joshi-Tope, G., Gillespie, M., Vastrik, I. *et al.* (2005), 'Reactome: A knowledgebase of biological pathways', *Nucleic Acids Res.*, Vol. 33 (Database issue), pp. D428–432.
14. Krieger, C. J., Zhang, P., Mueller, L. A. *et al.* (2004), 'MetaCyc: A multiorganism database of metabolic pathways and enzymes', *Nucleic Acids Res.*, Vol. 32(1), pp. D438–442.
15. Zeeberg, B. R., Feng, W., Wang, G. *et al.* (2003), 'GoMiner: A resource for biological interpretation of genomic and proteomic data', *Genome Biol.*, Vol. 4(4), p. R28.
16. Khatri, P., Bhavsar, P., Bawa, G. and Draghici, S. (2004). 'Onto-Tools: An ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments', *Nucleic Acids Res.*, Vol. 32, pp. W449–456.
17. Boyle, E. I., Weng, S., Gollub, J. *et al.* (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*, Vol. 20(18), pp. 3710–3715.
18. URL: http://www.geneontology.org/GO.tools.microarray.shtml
19. URL: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/
20. Camon, E., Magrane, M., Barrell, D. *et al.* (2003), 'The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro', *Genome Res.*, Vol. 13 , pp. 662–672.
21. URL: ftp://ftp.geneontology.org/pub/go/gp2protein/
22. Bussey, K. J., Kane, D., Sunshine, M. *et al.* (2003), 'MatchMiner: A tool for batch navigation among gene and gene product identifiers', Genome Biol., Vol. 4(4), p. R27.
23. Goff, S. A., Ricke, D., Lan, T. H. *et al.* (2002), 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)', *Science*, Vol. 296 , pp. 92–100.
24. Yu, J., Wang, J., Lin, W. *et al.* (2005), 'The genomes of *Oryza sativa*: a history of duplications', *PLoS Biol.*, Vol. 3, pp. 266–281 (e38).
25. Berardini, T. Z., Mundodi, S., Reiser, L. *et al.* (2004), 'Functional annotation of the *Arabidopsis* genome using controlled vocabularies', *Plant Physiol.*, Vol. 135, pp. 745–755.
26. Stein, L. D., Bao, Z., Blasiar, D. *et al.* (2003), 'The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics', *PLoS Biol.*, Vol. 1, pp. 172–192 (e45).
27. URL: http://www.geneontology.org/GO.slims.shtml?all#maint
28. URL: http://www.godatabase.org/dev/java/dagedit/docs/index.html
29. URL: http://www.geneontology.org/GO.slims.shtml?all#script
30. URL: http://www.geneontology.org/GO.annotation.shtml?all#file
31. URL: http://www.godatabase.org/cgi-bin/amigo/go.cgi
32. URL: http://www.ebi.ac.uk/ego/
33. URL: http://www.geneontology.org/GO.tools.browsers.shtml
34. URL: http://www.geneontology.org/GO.consortiumlist.shtml
35. URL: http://www.geneontology.org/GO.interests.shtml
36. URL: https://sourceforge.net/tracker/?atid=440764&group_id=36855&func=browse/
37. URL: http://www.geneontology.org/GO.requests.shtml
38. URL: http://obo.sourceforge.net/