D262–D266 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh021

The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology

Evelyn Camon*, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez and Rolf Apweiler

European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received August 15, 2003; Accepted August 22, 2003

ABSTRACT

The Gene Ontology Annotation (GOA) database (http://www.ebi.ac.uk/GOA) aims to provide high-quality electronic and manual annotations to the UniProt Knowledgebase (Swiss-Prot, TrEMBL and PIR-PSD) using the standardized vocabulary of the Gene Ontology (GO). As a supplementary archive of GO annotation, GOA promotes a high level of integration of the knowledge represented in UniProt with other databases. This is achieved by converting UniProt annotation into a recognized computational format. GOA provides annotated entries for nearly 60 000 species (GOA-SPTr) and is the largest and most comprehensive open-source contributor of annotations to the GO Consortium annotation effort. By integrating GO annotations from other model organism groups, GOA consolidates specialized knowledge and expertise to ensure the data remain a key reference for up-to-date biological information. Furthermore, the GOA database fully endorses the Human Proteomics Initiative by prioritizing the annotation of proteins likely to benefit human health and disease. In addition to a non-redundant set of annotations to the human proteome (GOA-Human) and monthly releases of its GO annotation for all species (GOA-SPTr), a series of GO mapping files and specific cross-references in other databases are also regularly distributed. GOA can be queried through a simple user-friendly web interface or downloaded in a parsable format via the EBI and GO FTP websites. The GOA data set can be used to enhance the annotation of particular model organism or gene expression data sets, although increasingly it has been used to evaluate GO predictions generated from text mining or protein interaction experiments. In 2004, the GOA team will build on its success and will continue to supplement the functional annotation of UniProt and work towards enhancing the ability of scientists to access all available biological information. Researchers wishing to query or contribute to the GOA project are encouraged to email: goa@ebi.ac.uk.

INTRODUCTION

The UniProt Knowledgebase (1) [which includes Swiss-Prot (2), TrEMBL (2) and PIR-PSD (3)] is the world's most highly annotated protein sequence database, having archived and annotated more than a million proteins through a combination of manual and electronic techniques. Over the next few years, it is estimated that this figure will increase to over 4 million proteins, the majority of which will lack biochemical and functional characterization. To prepare for this increase, those involved in bioinformatics have responded by developing new protocols for the capture, sharing and analysis of the functional annotation of the various data sets held. One way in which to maximize the annotation of these data while safeguarding quality, is to draw on the expertise of in-house and specialist community resources.

Successful integration is reliant on each database using the same language to characterize proteins and to distribute data in parsable formats. In this regard, one of the most important and well-used ontologies within the bioinformatics community is the Gene Ontology (GO) (4). GO is a dynamic controlled vocabulary of over 16 000 terms used to describe molecular function, process and location of action of a protein in a generic cell. The success of GO is based largely on its open source approach and the involvement throughout its development of various biological communities rich in expertise. Now 7 years on, the hype around GO has not lessened. The GO Consortium continues to develop strategies to improve the GO data set by staying abreast of future opportunities for integration with other useful open biological ontologies (OBOs) (5).

In support of standardized nomenclature, the UniProt group became a member of the GO Consortium annotation effort in 2001. It initiated the Gene Ontology Annotation (GOA) project (6,7) to provide assignments of GO terms to all well characterized proteins and in particular to that of the human proteome.

The initial aims and objectives of the GOA project have already been achieved. GOA has organized, shared and integrated protein knowledge using the GO structured vocabulary. This was facilitated with the help of UniProt curation, which led to the successful mapping to GO from existing references, resources and publications. Initially, GOA focused on producing GO annotations and improving update cycles. More recently, GOA has successfully supplemented the GOA-SPTr data set with annotations from GO Consortium

*To whom correspondence should be addressed. Tel: +44 1223 494465; Fax: +44 1223 494468; Email: goa@ebi.ac.uk initiators, Mouse Genome Database (MGD) (8), FlyBase (9) and *Saccharomyces* Genome Database (SGD) (10). In 2004, the GOA group will report on the manual evaluation of electronically extracted GO terms from literature as part of the BioCreative competition. It is also hoped that the information in GOA will accelerate the discovery of proteins of pharmaceutical interest.

GO ANNOTATION PROCESS

High-quality GO annotations (GOA) are generated through a combination of electronic and manual techniques, the latter of which employs a team of skilled biologists.

ELECTRONIC GO ANNOTATION

The large-scale assignment of GO terms to UniProt entries has been made possible by successfully converting a proportion of the pre-existing knowledge held within the flat files into GO terms (7). For example, UniProt description lines (DE) may contain Enzyme Commission (EC) numbers. Using an existing mapping of EC numbers to the GO molecular function ontology (ec2go) and a mapping of protein accession numbers to EC numbers, GOA can produce a UniProt to GO association. In a similar fashion the GOA group maintains a Swiss-Prot keyword to GO mapping (spkw2go). This mapping file is routinely used to generate a large number of annotations to GO process, function and component ontologies (see contents of current release on the GOA home page).

Bi-directional database cross-references also help to integrate GO annotations. For example, the majority of UniProt entries will cross-reference an InterPro identification number and vice versa. InterPro is a key database maintained at the EBI (11,12). It provides an integrated documentation resource for proteins, families and domains. A single InterPro entry provides comprehensive annotation describing a set of related proteins, some of which may have identical functions, be involved in the same processes and act in the same locations. During the curation of each InterPro entry, high-level GO terms are manually curated, based on a review of the literature available on the related proteins. This annotation is used to generate an InterPro2go mapping and also serves as a biological summary in the InterPro entry. So far, the application of the InterPro2go mapping in the electronic assignment of GO terms to gene products has produced the most coverage in the GOA data set (see contents of current release on the GOA home page). Both spkw2go and InterPro2go mappings are maintained in-house and distributed on the GO and EBI FTP sites on a regular basis. To support interoperability, InterPro2go has been used to generate GO mappings to its member databases (see Table 1) and these also are available for download.

The GO assignments are released monthly, in accordance with a GO Consortium agreed format, within a ‘gene association file’. As the mapping files used by GOA are manually curated, GOA is confident that its electronic annotation is of a high standard. Despite this, it is important that users have the ability to distinguish electronic from manually verified GO annotation. For this reason, the certainty of each GOA association is supported by annotating to one of 10 Consortium agreed evidence codes. Electronically generated associations are labelled as ‘inferred from electronic annotation’ or IEA. Proteins assigned this code are ‘likely’ to be involved in a particular GO activity.

MANUAL GO ANNOTATION

The large-scale assignment of GO terms to UniProt proteins using electronic methods is a fast and efficient way of associating high-level terms to a large number of proteins. However, to provide more reliable and specific annotation, the GOA project also makes use of manual curation using information extracted from published scientific literature (7). This process is slower than the use of electronic techniques but provides more accurate information as all annotation is validated by a team of skilled biologists. GOA recommends that users wishing to analyse GO annotation understand how GO is arranged and how GO assignments are made. Guidelines for GO annotation have been detailed before (10) and are published on the GO home page (http://www.geneontology.org/). Each assigned term is associated with a GO experimental evidence code (see GO home page for details of evidence codes) and a PubMed ID, which allows users to track the exact literature source and type of experiment used to support the annotation.

Priority is given in the GOA project to the annotation of data from the human proteome. This complements the efforts of the other consortium members as no other member freely provides human-specific data. The GOA project also assigns GO terms to proteins from a wide range of other species. There are almost 60 000 different species represented in the UniProt databases. Approximately 500 of these have already had GO terms manually assigned and this number continues to increase daily. New terms are requested as required to adequately describe the many species in UniProt, thus enhancing the GO ontologies and extending their scope.

UTILITY OF GO ANNOTATION

Manual GO annotation generates high-quality reliable information that is more accurate than electronic annotation. It also allows comparisons to be made with new annotation approaches and is an important tool for validation of these methods. However, manual annotation is time consuming and dependent on skilled biologists capable of extracting key information from the published literature.

In view of this, greater emphasis has been placed by the bioinformatics community on the development of new automatic annotation techniques, such as automated information extraction and the conversion of this knowledge into the GO vocabulary (9,17–20). This has resulted in a variety of GO prediction servers with varying abilities to interpret accurately the subtleties of the scientific natural language as well as GO structure, mappings and annotation styles (see GO Tools list on GO home page). To assess these information extraction techniques and allow users to apply the methods judiciously, the BioLINK group (http://www.pdg.cnb.uam.es/BioLINK/) organized the BioCreative (Critical Assessment of Information Extraction systems in Biology) competition. In collaboration with BioLINK, GOA provided one of the gold standard training and test sets of GO annotation. UniProt curators also took part in manual verification of GO terms

Table 1. Summary of sites for downloading or browsing the GOA data set

Resource	Description
Web-based tools	
QuickGO	A fast web-based EBI browser with access to core GO data and up-to-date electronic and manual UniProt GO annotations. Example GO term view: http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0006915. Example GOA view: http://www.ebi.ac.uk/ego/GSearch?query=P51587&mode=protein.
SRS	This Sequence Retrieval System (SRS) can be used to search GO annotation directly in the GOA database or in the UniProt Knowledgebase (Swiss-Prot, TrEMBL, PIR-PSD). Users can also search the GO Consortium repository and, via database links, perform complicated queries or link Blast outputs to GO annotations (http://srs.ebi.ac.uk).
Proteome Analysis pages	GO annotations have been produced for classification of proteins belonging to each complete proteome. A slimmed down version of GO (GO-slim), representing high-level GO terms, summarizes the biological attributes of a proteome (http://www.ebi.ac.uk/proteome). Users can compare proteomes with GO-slim, or simply find out what proportion of a proteome is involved in e.g. 'transport'. Annotations to GO-slim can also be downloaded.
InterPro	GO annotations made by InterPro curators are visible directly in InterPro entries or can be downloaded. Example entry: http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR000402.
AmiGO	This GO Consortium browser provides access to core GO data and released GOA data (http://www.godatabase.org/cgi-bin/go.cgi).
Ensembl	Ensembl is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources. Ensembl annotates known genes and predicts new ones, with functional annotation from many resources including GOA. Using the Ensembl GO view, users can find the location of human genes mapped to a particular GO term. Example entry: http://www.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000139618.
Genome KnowledgeBase	The Genome KnowledgeBase (GK) collaboration between the EBI and Cold Spring Harbor Laboratory, uses and contributes to the GOA project by curating GO terms to proteins involved in key pathways and reactions in human biology. [http://www.genomeknowledge.org/].
Alternative Splicing Database	GOA facilitates querying of the Alternative Splicing Database. The query page that links to the GO browser is at http://www.ebi.ac.uk/asd-srv/altextron/altsplice_hum/servlet/AltExtron.
LocusLink	In 2003, LocusLink replaced the human GO annotation supplied by the former Proteome group [now Incyte HumanPSD (13)] with that being provided by GOA-Human. (http://www.ncbi.nlm.nih.gov/LocusLink/).
Downloads	
GOA association file	This is a tab-delimited file of associations between gene products and GO terms which can be downloaded to supplement annotation of specialized data sets or validate automated ways of deriving information about gene function. Two GOA association files are currently produced: the GOA-Human file contains GO annotations for all proteins in the non-redundant human proteome set; the GOA-SPTr file contains GO annotations for all proteins in UniProt Knowledgebase. Both files can be accessed from the EBI FTP server (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa).
GOA Xref file	This is a file of cross references (xrefs.goa.gz) that displays the relationship between the entries in the GOA-Human data set with other databases using the International Protein Index (IPI) (2), as well as the nucleotide sequence databases (14), HUGO (15), LocusLink (16) and Refseq. This file supports the integration of GOA with other data sets. (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN).
UniProt Knowledgebase	GO annotation is visible in UniProt Knowledgebase (Swiss-Prot and TrEMBL, PIR-PSD) flat files, and these can be downloaded from: (http://www.ebi.ac.uk/uniprot). Note: In Swiss-Prot entries, only the manually annotated information is displayed. To view the complete GO annotation for a Swiss-Prot entry, you should download or browse the master copy of the data in the GOA association files. TrEMBL flat files display electronically derived GO annotation only when manual annotation is not available.
Annotations to GO-slim	To assist the analysis of GO annotation by specialist communities, GOA releases a mapping of its annotation to high-level GO-slim terms: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/.
Manual mappings to GO	The manual mappings of InterPro domains and families, Swiss-Prot keywords and EC numbers to GO terms can be directly downloaded to enhance the functional annotation of mass spectrometry or microarray data. These mappings are shared on the EBI FTP site: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/external2go/.
Electronic mappings to GO	Using InterPro2go a set of electronic GO mappings have been generated to other InterPro Consortium databases (e.g. SMART, PRINTS, ProDom, PROSITE and TIGRFAM). To promote GO integration, these are shared on the EBI and GO FTP sites (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/external2go/).

mined from the literature, which were corroborated by evidence from the text. In spring 2004, the results of the competition will be announced and it is anticipated that those techniques with the potential to assist GOA in the accurate prediction of deep-level GO terms will be highlighted and may supercede current electronic strategies.

The core objective of GOA is to provide high-quality GO annotation to supplement and improve the interoperability and querying of external and in-house knowledgebases. Increasingly, it is being used to predict the function and biological roles of new gene products, and to reclassify and model relationships between known proteins (19,20). For example, genes coding for gene products that are often involved in the same biological process have a likelihood of being regulated in a coordinated manner. For these reasons, public repositories of microarray data such as ArrayExpress (21) actively encourage users to provide GO annotation in the array design, to facilitate clustering of data according to the GO ontologies and to allow cross-platform data comparisons. GO annotation provides a link between biological knowledge and gene expression profiles. When combined with statistical analysis, the GOA data set is a useful resource for building pathways and can help facilitate microarray probe design to create more focused arrays (e.g. neurofunction arrays) (22,23). It is hoped that this use of GOA will assist researchers in identifying quickly new proteins of pharmaceutical interest based on their functional similarities. In addition to microarray data analysis tools, GOA is also incorporated into evolutionary studies, particularly when correlating structure–function relationships (24). These examples help to demonstrate the potential that exists for the use of GOA in the development and validation of tools that try to accurately predict GO annotation.

GOA can also be used to answer specific biological problems. As GO represents a universal set of curated keywords, many users wish to retrieve all possible annotations to a high-level GO term in a candidate-based approach. According to GO philosophy, every child term inherits the meaning of all of their parent terms. As such, every annotation to a child term should be true for every parent of that child; this is called the 'true path rule'. If a user wanted to analyse all proteins involved in the process of transcription they would have to retrieve all proteins annotated to the GO term for 'transcription' (GO:0006350) and the children of these GO terms. Retrieving the annotation to the children and parent GO term is possible via SRS (25) but requires prior knowledge of this powerful retrieval system.

Another way of performing the query is to use the protein assignments to a set of GO-slim terms. Essentially, GO-slim is a list of high-level GO terms that cover the main aspects of each of the three GO ontologies. As each community has different needs, a variety of GO-slim files have been archived on the GO home page by Consortium members (ftp://ftp.geneontology.org/pub/go/GO_slims/). GOA has created its own GO-slim (goslim_goa.2002) to summarize the GO annotation of each completed proteome on the Proteome Analysis pages (26, Table 1). As an additional service, this mapping of GO annotation to both GOA (goslim_goa.2002) and a generic set of GO-slim terms (generic.0208) is available for download on the EBI FTP site (Table 1). From there, users can download all possible annotations to the GO slim term for transcription 'GO:0006350'. Users wishing to use a different set of GO-slim terms are advised to use the map2slim.pl script archived on the Berkley Drosophila Genome Project (BDGP) home page (http://www.fruitfly.org/developers/src/go.dev/apps/query-utils/). This script uses the GO MySQL database and requires prior knowledge of Perl API.

DATABASE ACCESS

There are various ways of accessing and searching GOA project data (7). In addition to several web-based browsers, GOA files and mappings can be downloaded (Table 1). GOA is also cross-referenced in numerous databases including the UniProt Knowledgebase, Ensembl (27), LocusLink (16), EMBL/DDBJ/GenBank databases (14) and the Nuclear Protein Database (28).

FUTURE PERSPECTIVES

The GOA group will work towards extending its annotation coverage of the UniProt Knowledgebase and further promote database interoperability and utility of GO in research. The group is also actively recruiting new curators within the UniProt Consortium in support of the manual annotation of the dataset and the integration of GO assignments made by other specialised disease or GO Consortium databases. It is envisaged that GOA will contribute to the functional curation of the interaction and alternative splicing databases (IntAct, ASD). GOA will also facilitate the annotation of GO with other open biological ontologies (5; http://obo.sourceforge.net/).

To further develop electronic annotation techniques, GOA will apply new GO mappings [e.g. a manual mapping of GO terms to microbial families in HAMAP (29)] and stay abreast of new text mining and information extraction technologies. More complicated GO queries will be facilitated by upgrades to in-house retrieval systems.

HOW TO CONTRIBUTE TO THE GOA PROJECT

GO annotations are an extensive and valuable resource that the team is pleased to make freely available. The success and accuracy of GOA relies on the support from biological communities who use and share GO annotation and on our commitment to keep electronic and manual data sets up to date. GOA actively encourages all users to improve this resource, by communicating via email (goa@ebi.ac.uk) any GO annotations that require updating. In this scenario, please provide the accession number of the entry, along with the source and date of annotation retrieval. As GOA integrates GO annotation from other Consortium members, users can verify the original database source of manual annotation by viewing the attribution column of the association file. It is worth noting that GOA cannot be synchronized with all databases that display its annotation. Please check for current GOA version numbers on the home page.

GOA especially encourages specialist groups also interested in fast-tracking the annotation of the human proteome, to initiate collaborations with us, that we might cross-reference and share annotations and freely provide to our scientists as complete a resource as possible.

ACKNOWLEDGEMENTS

We especially thank all colleagues worldwide who have contributed to the information content of the GOA data set and to the development of annotation and retrieval tools. We would also like to acknowledge Philippe Rocca-Serra, Susanna-Assunta Sansone (ArrayExpress) and Michael McWhinney for their editorial contributions to this manuscript. The GOA project is grateful for the support of Grants QRLT-2001-00015 and QLRI-2000-00981 of the European Commission and a supplementary NIH grant, 1R01HGO2273-01.

REFERENCES

1. Butler,D. (2002) NIH pledges cash for global protein database. *Nature*, **419**, 101.
2. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
3. Wu,C.H., Huang,H., Yeh,L.S. and Barker,W.C. (2003) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.
4. The Gene Ontology Consortium. (2000) Gene ontology tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
5. Harris,M.A. and Parkinson,H. (2003) Conference Report: Standards and ontologies for functional genomics: towards unified ontologies for biology and biomedicine. *Comp. Funct. Genomics*, **4**, 116–120.
6. Camon,E., Barrell,D., Brooksbank,C., Magrane,M. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project—application of

GO in SWISS-PROT, TrEMBL and InterPro. *Comp. Funct. Genomics*, **4**, 71–74.

7. 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.
8. Hill,D.P., Blake,J.A., Richardson,J.E. and Ringwald,M. (2002) Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.*, **12**, 1982–1991.
9. King,O.D., Foulger,R.E., Dwight,S.S., White,J.V. and Roth,F.P. (2003) Predicting gene function from patterns of annotation. *Genome Res.*, **13**, 896–904.
10. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
12. Biswas,M., O'Rourke,J.F., Camon,E., Fraser,G., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N. *et al.* (2002) Applications of InterPro in protein annotation and genome analysis. *Brief. Bioinform.*, **3**, 285–295.
13. Hodges,P.E., Carrico,P.M., Hogan,J.D., O'Neill,K.E., Owen J.J., Mangan,M., Davis,B.P., Brooks,J.E. and Garrels,J.I. (2002) Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics. *Nucleic Acids Res.*, **30**, 137–141.
14. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
15. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–80.
16. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
17. Stevens,R., Goble,C.A. and Bechhofer,S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.*, **1**, 398–414.
18. Xie,H., Wasserman,A., Levine,Z., Novik,A., Grebinskiy,V., Shoshan,A. and Mintz,L. (2002) Large-scale protein annotation through gene ontology. *Genome Res.*, **12**, 785–794.
19. Schug,J., Diskin,S., Mazzarelli,J., Brunk,B.P. and Stoeckert,C.J.,Jr (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.*, **12**, 648–655.
20. Lagreid,A., Hvidsten,T.R., Midelfart,H., Komorowski,J. and Sandvik,A.K. (2003) Predicting gene ontology biological process from temporal gene expression patterns. *Genome Res.*, **13**, 965–979.
21. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
22. Zhong,S., Li,C. and Wong,W.H. (2003) ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res.*, **31**, 3483–3486.
23. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
24. Shakhnovich,B.E., Dokholyan,N.V., DeLisi,C. and Shakhnovich,E.I. (2003) Functional fingerprints of folds: evidence for correlated structure–function evolution. *J. Mol. Biol.*, **326**, 1–9.
25. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **8**, 1149–1150.
26. Pruess,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I., Servant,F. *et al.* (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res.*, **31**, 414–417.
27. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2002) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–54.
28. Dellaire,G., Farrall,R. and Bickmore,W.A. (2003) The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res.*, **31**, 328–330.
29. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **1**, 49–58.