The UniProt-GO Annotation database in 2011

Emily C. Dimmer[1,*], Rachael P. Huntley[1], Yasmin Alam-Faruque[1], Tony Sawford[1], Claire O'Donovan[1], Maria J. Martin[1], Benoit Bely[1], Paul Browne[1], Wei Mun Chan[1], Ruth Eberhardt[1], Michael Gardner[1], Kati Laiho[1], Duncan Legge[1], Michele Magrane[1], Klemens Pichler[1], Diego Poggioli[1], Harminder Sehra[1], Andrea Auchincloss[2], Kristian Axelsen[2], Marie-Claude Blatter[2], Emmanuel Boutet[2], Silvia Braconi-Quintaje[2], Lionel Breuza[2], Alan Bridge[2], Elizabeth Coudert[2], Anne Estreicher[2], Livia Famiglietti[2], Serenella Ferro-Rojas[2], Marc Feuermann[2], Arnaud Gos[2], Nadine Gruaz-Gumowski[2], Ursula Hinz[2], Chantal Hulo[2], Janet James[2], Silvia Jimenez[2], Florence Jungo[2], Guillaume Keller[2], Phillippe Lemercier[2], Damien Lieberherr[2], Patrick Masson[2], Madelaine Moinat[2], Ivo Pedruzzi[2], Sylvain Poux[2], Catherine Rivoire[2], Bernd Roechert[2], Michael Schneider[2], Andre Stutz[2], Shyamala Sundaram[2], Michael Tognolli[2], Lydie Bougueleret[2], Ghislaine Argoud-Puy[3], Isabelle Cusin[3], Paula Duek-Roggli[3], Ioannis Xenarios[2,4] and Rolf Apweiler[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Swiss-Prot group, [3]CALIPHO group, Swiss Institute of Bioinformatics, CMU, 1 Michel Servet, 1211 Geneva 4 and [4]Vital-IT Group, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland

ABSTRACT

The GO annotation dataset provided by the UniProt Consortium (GOA: http://www.ebi.ac.uk/GOA) is a comprehensive set of evidenced-based associations between terms from the Gene Ontology resource and UniProtKB proteins. Currently supplying over 100 million annotations to 11 million proteins in more than 360 000 taxa, this resource has increased 2-fold over the last 2 years and has benefited from a wealth of checks to improve annotation correctness and consistency as well as now supplying a greater information content enabled by GO Consortium annotation format developments. Detailed, manual GO annotations obtained from the curation of peer-reviewed papers are directly contributed by all UniProt curators and supplemented with manual and electronic annotations from 36 model organism and domain-focused scientific resources. The inclusion of high-quality, automatic annotation predictions ensures the UniProt GO annotation dataset supplies functional information to a wide range of proteins, including those from poorly characterized, non-model organism species. UniProt GO annotations are freely available in a range of formats accessible by both file downloads and web-based views. In addition, the introduction of a new, normalized file format in 2010 has made for easier handling of the complete UniProt-GOA data set.

INTRODUCTION

The UniProt Knowledgebase (UniProtKB; http://www.uniprot.org) (1) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation (1). Included among the wealth of annotation data are detailed Gene Ontology annotation statements, created in collaboration with the Gene Ontology project (http://www.geneontology.org) (2). The UniProt Consortium is a central member of the Gene Ontology Consortium, an initiative founded in 1998 to develop and use a set of structured, controlled vocabularies to represent three aspects of biology carried out by gene products from any organism. Terms within the GO describe those molecular functions and biological

processes that gene products carry out and the subcellular locations in which they are located.

This article will report the developments of the UniProt GO annotation data set since its last description in the NAR database issue in 2009 (3).

OVERVIEW

What are GO annotations?

A GO annotation is an evidenced-based association between a gene product identifier (in this instance, a UniProtKB accession) and a Gene Ontology term. Two additional and essential components of a GO annotation are the presence of a reference that supports the association, and an evidence code, used to indicate the type of evidence for the association in the cited reference. Manual GO annotations are created by trained biocurators who critically curate data from published, peer-reviewed scientific literature. A single well-characterized gene product can be annotated to multiple GO terms located at different levels in the three GO hierarchies, according to the data presented in different scientific references.

GO annotation sets are now widely used by the scientific community to aid the analysis of large proteomic and genomic data sets; patterns of GO term attribution are measured to assist in hypothesis generation for the biological events occurring behind experimental results, to validate experimental methods or to provide a broad overview of the principal functional characteristics of a genome or proteome.

Manual GO annotation developments

The UniProt resource benefits from the wide range of manual curation activities of its curators, who interpret and integrate data from a wide variety of sources. Manual GO annotation is a core activity for UniProt curators (www.uniprot.org/docs/sop_manual_curation .pdf), who supply data for a wide range of species obtained from curation of primary experimental literature. UniProt concentrates its manual annotation effort on entries from model organisms, providing high-quality annotation to proteins across a diverse set of taxonomic groups, including mammalian, fungal, plant, insect, bacteria, nematode and viral species. This UniProt GO annotation effort is supported by a group of specialized UniProt curators whose primary role is to support, annotate and develop this GO annotation data set. In addition, these specialized curators are responsible for detailed GO annotation sets for human targets prioritized by the GO Consortium as well as extensive annotation of proteins associated with the mammalian renal system and apoptosis processes (4,5).

The UniProt GO annotations are supplemented with those from 36 external groups [annotations from the PAMGO (6), EcoCyc (7), EcoWiki (8), JCVI (9) and CGD (10) have been added to the data set since 2009], ensuring that all manual annotations from GO Consortium members that can be mapped to a UniProtKB accession are included in the UniProt data set. Integration of GO annotations from other sources involves extensive efforts to ensure external identifiers are appropriately applied and that the annotations integrated are of the required semantic and syntactic quality. This integration work is a highly collaborative effort between UniProt and the external curation resources. Such integration efforts can be exemplified by work in the last year to include 43 different gene, protein and chemical identifier types [such as WormBase (11), ChEBI (12) and EcoCyc (7) identifiers] in the annotation format to provide a more complete display of external manual annotations.

Through its GO curation tool and annotation release pipeline, UniProt has also been able to provide support to smaller, expert curation groups, ensuring that high-quality curation can be carried out without the need for every group to be involved in expensive software development activities. In this manner five curation groups [AgBase (13), BHF-UCL (4), dictyBase (14) as well as annotation efforts in the Systems Biology departments of Tufts University, Massachusetts, USA and the Norwegian University of Science and Technology] are currently supported by UniProt, with the annotations created being attributed to the supported group but integrated directly into the UniProt GO annotation set and annotation views.

Automatic GO annotation developments

UniProt GO annotation has increased in line with the growing number of sequences and annotations available in UniProtKB. Supplying a first pass of high-quality functional statements to such a rapidly increasing sequence set can only be attempted with the application of automatic GO annotation prediction methods. Such prediction methods are highly valuable; they are often the only way in which proteins from non-model organisms, without funding to support manual GO annotation work, can be supplied with any functional annotation, and whereby all species can benefit from a greater consistency in the annotation data set.

UniProt is the main supplier of electronic GO annotation predictions to the GO Consortium via close collaboration with independent annotation efforts in the InterPro (15), Ensembl Genomes (16) and Ensembl (17) projects. At their core, all methods include a manual annotation effort to ensure appropriate automatic propagation of GO annotations between proteins. A full description of the different automatic annotation methods is described elsewhere (3). However in brief, annotation prediction pipelines differently exploit existing cross-references, keywords from external controlled vocabularies, protein sequence signatures that indicate the involvement of a protein set in carrying out a particular function, and gene orthology data.

In June 2011, a new electronic annotation pipeline was implemented that supplies GO annotations to the UniProt data set, provided by Gramene and the EnsemblPlants resource (16). This pipeline produces annotation predictions by projecting manual GO annotations from *Arabidopsis thaliana* or *Oryza sativa* onto proteins from one or more target species based on gene orthology

data. This is an important addition to an existing Ensembl automatic annotation method that applies orthology data from the Ensembl Compara annotation pipeline to project GO annotations between selected vertebrates (17). The first release of the EnsemblPlants data set in August 2011 supplied almost 230 000 annotations to over 50 000 proteins covering 16 species, including poplar, maize, sorghum, grape and *Physcomitrella*.

Figure 1 summarizes the data flows into the UniProt GO annotation program as well as the exported GO annotation products.

Improvements to annotation quality

To ensure that UniProt can confidently supply high-quality GO annotation statements, an extensive set of quality controls has been developed. Syntactic tests are applied to ensure that all identifiers supplied (sequence identifiers, GO identifiers, references) are valid. Semantic quality controls have also been increasingly applied over the last 2 years to ensure that the functional statement is sound, including tests to check whether the biological data provided is appropriate for the protein's taxonomic group [for instance, ensuring that pig proteins are not incorrectly annotated to a GO term that describes 'GO:0007629; flight behaviour'(18)], and that the annotation statement appropriately applies the expected values to ensure that meaningful data is supplied (for example, supplying the identity of the interacting entity when the term 'GO:0005488; binding' is applied). Such checks have been formalized in consultation with the GO Consortium and are embedded into the UniProt-GOA curation tool to prevent errors from being added into the data set when manual annotations are created and regularly applied to entire GO annotation set before each release (http://www.geneontology.org/GO .annotation_qc.shtml).

Greater efforts to supply high-quality protein–protein interaction data has also meant that the UniProt GO annotation file now supplies a reduced number of protein binding GO annotations from the IntAct protein interaction database (19). A subset of reliable interactions is now extracted from the IntAct data set, the export being determined using a simple scoring system and rule set, coupled to a score threshold that has been deliberately chosen to exclude interactions supported by only one experimental observation.
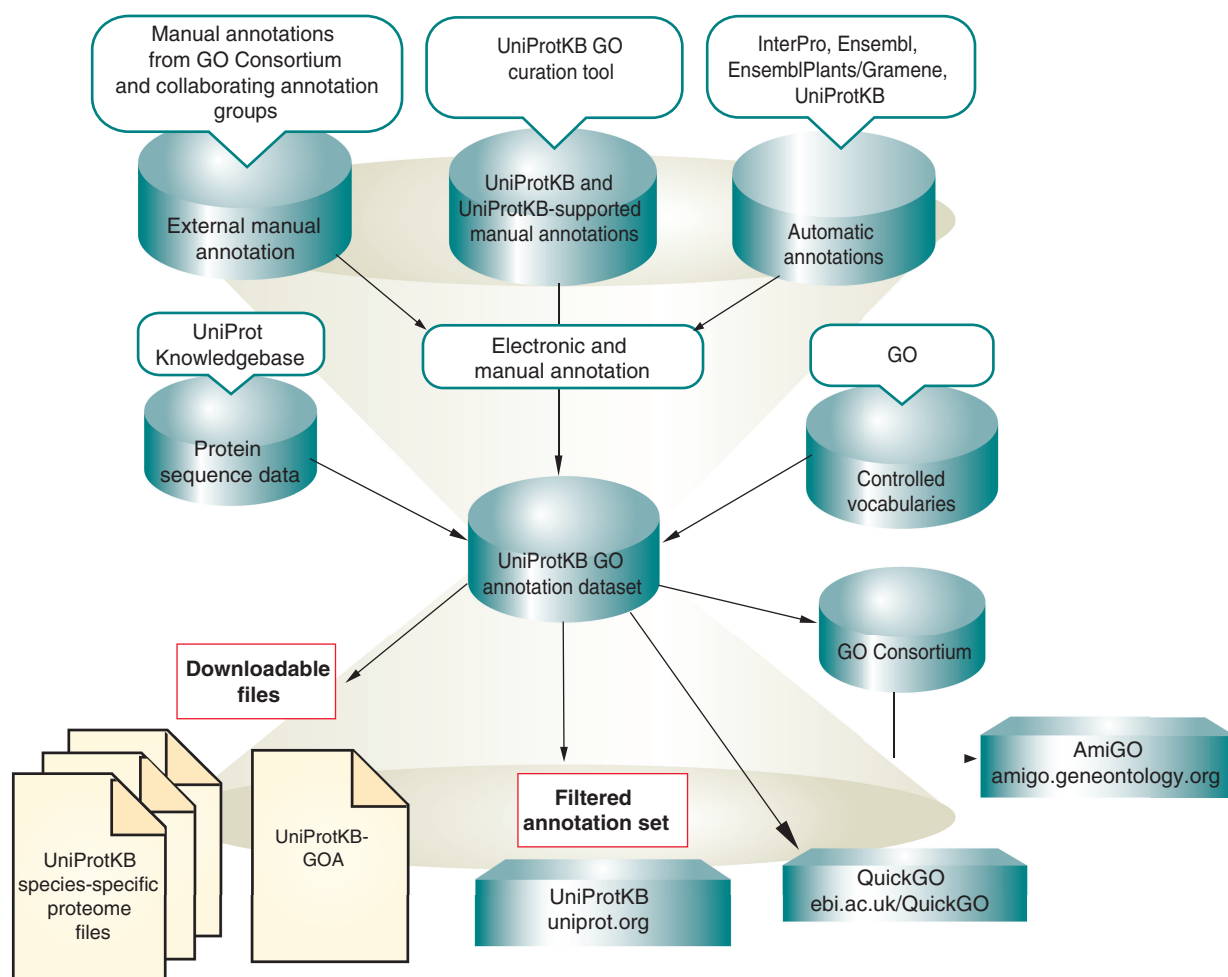

Figure 1. Sources and flow of data for the UniProt-GOA data set.

Improvements to annotation consistency

Work to improve the GO annotation set has not only involved the removal of incorrect GO annotations, but improvements to annotation consistency have also required creating annotations using inter-ontology relationships that now exist between the three GO ontologies. GO Consortium-designed annotation pipelines are run over the entire UniProt GO annotation set resulting in Biological Process annotations being inferred and automatically generated where a particular Molecular Function term, which has been asserted to always occur within the context of that specific process, has been applied in an annotation (an example would be the Molecular Function term: 'GO:0004672; protein kinase activity' always infers the Biological Process term 'GO:0006468; protein phosphorylation'). A similar method is applied to infer and create Cellular Component annotations using ontology relationships that indicate where a process is *always* carried out at a specific subcellular location. The generation of inferred annotations for the UniProt GO annotation set commenced in 2011. The regular update of these inferences ensure curators are assisted in creating a complete annotation set that has consistently applied terms from each of the three GO ontologies.

Annotation file developments

Changes to file composition. In December 2010, UniProt began to produce species-specific GO annotation gene association files using the UniProt complete proteomes sets (1). Previously, proteome sets defined by the now-defunct Integr8 (20) and IPI (21) projects were used by the UniProt-GOA project. A UniProt complete proteome is defined as the entire set of proteins expressed by an organism. Most UniProt complete proteome sets are based on the translation of a completely sequenced genome, including sequences from extra-chromosomal-elements such as organellar genomes. Some complete proteomes may also include translations of high-quality cDNAs where necessary (22). UniProt currently provides individual GO annotation files for 1764 complete proteome sets (20 September 2011 release).

Gene association file 2.0. The official GO Consortium exchange file format for GO annotations was updated in March 2010 to GAF2.0, to include an additional two columns (23): the 'Annotation Extension' and 'Gene Product Form ID' columns. UniProt-GOA has been supplying GO annotations in this new GAF2.0 file format since April 2010.

Annotation Extension field allows curators to add extra contextual information to enhance a functional or subcellular location statement. The field format is designed to accept OBO identifiers [e.g. Cell Type (24), GO or ChEBI (12)] or sequence identifiers, which are prefixed by descriptive relationship types. This format allows curators to supply detailed information as to the context of an annotated function or location for specific proteins. For example, an annotation describing the kinase activity of TGBR2 can now include data describing a specific substrate: 'has_substrate (P17813)'. Alternatively, an annotation demonstrating the protein Tid1-L binding to Hsc70 can include location information: 'occurs_in(GO:0005829; cytosol)'. The September 2011 UniProt GO annotation release included 36 000 annotations that have included such contextual information, and this number is expected to rapidly increase over the next year. *Gene Product Form ID* provides users with identifiers for specific variants of the gene product being annotated. This change has meant that the functions or locations of specific isoforms or post-translationally modified proteins can now be consistently included. The September 2011 UniProt GO annotation release contains 3900 annotations to specific protein forms. As an example, this type of annotation data is now able to state that curation of the paper indexed in PubMed as 19085255 provides experimental evidence that Ribosomal protein S6 kinase 1 (S6K1) isoforms are differently located in the cytoplasm and nucleus.

Details of the GAF2.0 annotation file format can be viewed on the GO Consortium website (http://www.geneontology.org/GO.format.gaf-2_0.shtml)

Gene product association data and gene product information files

Since June 2010, UniProt has additionally supplied GO annotations for the UniProtKB-GOA 'UniProt' gene association file in an alternative format. The Gene Product Association Data file (GPAD), only contains the necessary information required to describe a GO annotation, whereas the complementary Gene Product Information file (GPI) contains information that identifies and describes UniProtKB proteins. The primary advantage of using these two files instead of the GAF2.0 format is the reduced redundancy in the information supplied; the combined compressed size of the GPAD and GPI files is 212MB less that the gene association file format for the UniProt-GOA data set, therefore benefiting users who would like to make use of this data but have found the size of the GAF file too unwieldy for their needs.

The GPAD file additionally allows the inclusion of Evidence Code Ontology (ECO) identifiers, an ontology of experimental and other evidence statements (http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidence_code). For example, for annotations that have applied the GO experimental evidence code EXP, the GPAD file will display the equivalent ECO identifier for this code (ECO:0000269: experimental evidence used in manual assertion). The availability of annotations using ECO codes means that evidence categories can be grouped or expanded using the ECO ontology structure according to individual user needs.

DATA ACCESS

QuickGO

The QuickGO tool (http://www.ebi.ac.uk/QuickGO) (25), is the primary location for UniProt GO annotation data, where a full, weekly updated GO annotation set is

made available to view, filter and download. Even though UniProt GO annotations are fully cross-referenced in UniProtKB entries, the displayed GO annotation set on the UniProt web site is filtered due to the frequent redundancy of the GO terms supplied by different manual and automatic GO annotation efforts to the same protein.

Following a major redevelopment of QuickGO, the tool is now able to offer users extensive facilities in GO term and annotation searching, filtering, identifier mapping and downloading of tailor-made subsets of the GO annotations. This facility is particularly valuable as the full UniProtKB GO annotation file is now 17.6 GB in its uncompressed form.

Users can also browse the GO hierarchies with QuickGO, and the tool uniquely provides: 1. a Change Log detailing changes to the three GO ontologies (Figure 2), 2. information on those terms that are often co-annotated to a protein and 3. suggestions for terms to replace GO terms that have been made obsolete or secondary. The extensive data supplied both for the ontologies and the associated GO annotation set has additionally allowed QuickGO to provide a user-friendly means of creating and annotating GO slims.

Web services

All data supplied by QuickGO can be queried remotely, both for ontology and term information, as well as annotation data. Details on how to access QuickGO's web services are at http://www.ebi.ac.uk/QuickGO/WebServices.html

Data releases. The entire UniProt GO annotation data set is updated weekly and released into the QuickGO browser. Every weekly release includes a fully updated set of manual and electronic annotations. Sanity checks are performed immediately before release to ensure annotation consistency and an 'annotation blacklist' is applied to ensure swift removal of individual incorrect functional assertions from the automatic pipelines.

UniProt provides monthly releases of GO annotations in a 17-column, tab-delimited 'gene association file' (GAF2.0) format (described here: http://www.geneontology.org/GO.format.gaf-2_0.shtml). The 'UniProt' file is additionally supplied in the GPAD and GPI formats, as described in the 'Annotation file developments' section above.

Annotations files can be accessed from the ftp site; ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/

A full history of the changes that have been implemented by UniProt to the GO annotation sets can be read at the project's news page, here: http://www.ebi.ac.uk/GOA/news.html

SUMMARY

The UniProt-GOA data set is the largest and most comprehensive source of multi-species GO annotation freely available to the scientific community. By integrating new sources of both manual and automatic GO annotation and providing quality assessments on our entire data set, we continue to improve the number and quality of GO annotations for an increasing numbers of species.

Figure 2. The Change Log display in QuickGO. QuickGO displays ontology changes relating to the addition of new terms, term obsoletions, changes in definitions or synonyms, relationships between terms and cross-references.

The QuickGO tool has evolved into a powerful facility providing highly flexible filtering options for obtaining customized GO annotation data sets. Its popularity is increasing; QuickGO had an average of 24 750 unique visitors per month during 2011, up from 21 100 in 2010.

Annotation feedback is welcomed through the SourceForge annotation tracker (http://sourceforge.net/tracker/?group_id = 36855&atid = 605890).

REFERENCES

1. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
2. The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
3. 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.
4. Lovering,R.C., Dimmer,E.C. and Talmud,P.J. (2009) Improvements to cardiovascular gene ontology. *Atherosclerosis*, **205**, 9–14.
5. Alam-Faruque,Y., Dimmer,E.C., Huntley,R.P., O'Donovan,C., Scambler,P. and Apweiler,R. (2010) The renal gene ontology annotation initiative. *Organogenesis*, **6**, 71–75.
6. Torto-Alalibo,T., Collmer,C.W. and Gwinn-Giglio,M. (2009) The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. *BMC Microbiol.*, **9**, S1.
7. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muñiz-Rascado,L., Bonavides-Martinez,C., Paley,S., Krummenacker,M., Altman,T. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.
8. Hu,J.C., Karp,P.D., Keseler,I.M., Krummenacker,M. and Siegele,D.A. (2009) What we can learn about *Escherichia coli* through application of Gene Ontology. *Trends Microbiol.*, **17**, 269–278.
9. Davidsen,T., Beck,E., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K. and White,O. (2010) The comprehensive microbial resource. *Nucleic Acids Res.*, **38**, D340–D345.
10. Skrzypek,M.S., Arnaud,M.B., Costanzo,M.C., Inglis,D.O., Shah,P., Binkley,G., Miyasato,S.R. and Sherlock,G. (2010) New tools at the Candida Genome Database: biochemical pathways and full-text literature search. *Nucleic Acids Res.*, **38**, D428–D432.
11. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
12. de Matos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
13. McCarthy,F.M., Gresham,C.R., Buza,T.J., Chouvarine,P., Pillai,L.R., Kumar,R., Ozkan,S., Wang,H., Manda,P., Arick,T. *et al.* (2011) AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Res.*, **39**, D497–D506.
14. Gaudet,P., Fey,P., Basu,S., Bushmanova,Y.A., Dodson,R., Sheppard,K.A., Just,E.M., Kibbe,W.A. and Chisholm,R.L. (2011) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res.*, **39**, D620–D624.
15. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
16. Jaiswal,P. (2011) Gramene database: a hub for comparative plant genomics. *Methods Mol. Biol.*, **678**, 247–275.
17. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
18. Deegan,J.I., Dimmer,E.C. and Mungall,C.J. (2010) Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics*, **11**, 530.
19. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
20. Pruess,M., Kersey,P. and Apweiler,R. (2005) The Integr8 project– a resource for genomic and proteomic data. *In Silico Biol.*, **5**, 179–185.
21. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
22. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
23. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
24. Meehan,T.F., Masci,A.M., Abdulla,A., Cowell,L.G., Blake,J.A., Mungall,C.J. and Diehl,A.D. (2011) Logical development of the cell ontology. *BMC Bioinformatics*, **12**, 6.
25. Binns,D., Dimmer,E., Huntley,R., Barrell,D., O'Donovan,C. and Apweiler,R. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.