Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3799–3803***
DOI: 10.1093/nar/gkg555

OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms

Günther Zehetner*

Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received February 14, 2003; Revised and Accepted 31 March, 2003

ABSTRACT

OntoBlast allows one to find information about potential functions of proteins by presenting a weighted list of ontology entries associated with similar sequences from completely sequenced genomes identified in a BLAST search. It combines, in a single analysis step, the search for sequence similarities in several species with the association of information stored in ontologies. From each identified ontology term a list of genes, which share the functional annotation, can be retrieved. The OntoBlast function is an integral part of the 'Ontologies TO GenomeMatrix' tool which provides an alternative entry point from ontology terms to the Genome–Matrix database. OntoBlast's web interface is accessible on the 'Ontologies TO GenomeMatrix Gate' page at http://functionalgenomics.de/ontogate/.

INTRODUCTION

The integration of sequence data with information from functional analyses of genes is an important and challenging task. Functional annotations of sequences allow first insights into the processes in which a gene product might be involved. A possible way to provide such an annotation is the association of a gene or protein sequence with predefined terms describing known characterised functions. A widely used structured vocabulary of this type is the Gene Ontology (GO) resource (1), which consists of three ontologies describing molecular functions, biological processes and cellular components.

A growing number of associations of genes, gene products and database identifiers to GO terms are readily available via the internet either from the GO website (http://www.geneontology.org/#indices, http://www.geneontology.org/#annotations), the GO Annotation@EBI (2) (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/) or other external species specific databases. There are also several tools which permit to browse and search the GO ontologies and to display associated entries in external databases either using one of many GO browsers (http://www.geneontology.org/#tools) or the SRS service (http://srs.ebi.ac.uk/). A few tools allow the use of GO terms to locate associated human and/or mouse genes (CGAP GO Browser: http://cgap.nci.nih.gov/Genes/GOBrowser/ or the EP GO Browser: http://ep.ebi.ac.uk/EP/GO/) to identify relations between GO terms and diseases [Genes2Diseases (3): http://www.bork.embl-heidelberg.de/g2d/] or to use protein database accession numbers to retrieve the corresponding GO terms (ProToGO: http://www.protogo.cs.huji.ac.il/).

Recently BLAST servers have been made available which combine their search results directly with annotations from GO. One such service is the GOst software tool which can be accessed from the AmiGO browser (http://www.godatabase.org/cgi-bin/go.cgi). Another BLAST server which retrieves automatically associated GO terms is GOblet (http://goblet.molgen.mpg.de/) developed as a project within the NGFN (Nationales Genomforschungsnetz) in Germany.

The primary purpose of OntoBlast (OB) is not to provide just a value-added BLAST server, but to generate a list of ontology terms associated with the query sequence which serve as entry points linking to the Genome-Matrix (GM, http://genome-matrix.org), a multi species, gene-region database which was introduced at the GM2002 meeting in Shanghai (4) (http://hgm2002.hgu.mrc.ac.uk/Abstracts/Publish/WorkshopPosters/WorkshopPoster01/hgm0023.htm) and will be described elsewhere (A. Hewelt *et al.*, in preparation). Using these links as part of the 'Ontologies TO GenomeMatrix' tool, it is possible to identify, in a second step, genes which are related to the original query sequence, not by structural similarity, but by sharing functional annotations. The information accessible from the GM database can than assist in the analysis and evaluation of the proposed associations between sequence and functions.

Sequence similarity is a frequently used feature to generate annotations and has also been used, together with protein domain analysis, in systematic protein annotation projects working with the Gene Ontology (5,6).

MATERIALS AND METHODS

Source data

Genome data from nine species have been prepared for this tool and correspond to the datasets used for the GM project as shown in Table 1.

*Tel: +49 (0)3084131357; Fax: +49 (0)3084131384; Email: zehetner@molgen.mpg.de

Table 1.

Species	Data source
Homo sapiens	ENSEMBL (7,8), http://ensembl.org/Homo_sapiens/
Mus musculus	ENSEMBL, http://ensembl.org/Mus_musculus/
Caenorhabditis elegans	WormBase (9), http://wormbase.org/
Drosophila melanogaster	FlyBase (10), http://www.flybase.org/
Rattus norvegicus	Ratmap, http://ratmap.gen.gu.se/ RGD (11), http://rgd.mcw.edu/
Saccharomyces cerevisiae	SGB (12), http://genome-www.stanford.edu/Saccharomyces/
Schizosaccharomyces pombe	GeneDB, http://www.sanger.ac.uk/Projects/S_pombe/
Plasmodium falciparum	PlasmoDB (13), http://plasmodb.org/
Anopheles gambiae	ENSEMBL, http://ensembl.org/Anopheles_gambiae/

Amino acid sequences have been extracted from SWISS-PROT (14) (http://www.ebi.ac.uk/swissprot/), TrEMBL (http://www.ebi.ac.uk/trembl/) and species specific databases. At present the three ontologies from Gene Ontology are implemented (data have been downloaded from the GO ftp site ftp://ftp.geneontology.org/), but it is planned to add other vocabularies like the EC enzyme classification. GO terms listed after a BLAST search are linked to GM entries via categorised tables and sorted gene lists.

This information, linking each ontology term (and its parents) to all associated genes (including homologs), as well as to the corresponding amino acid sequences, is pre-calculated on a regular basis to ensure an up-to-date and fast display. About 58 000 such data files each, for the cumulative and non-cumulative version, are produced.

Web interface

The main part of the user interface consists of two frames. The 'search frame' displays the form in which the query sequence and parameters for the BLAST search can be entered and later the actual BLAST result. The 'list frame' shows the weighted list of ontology terms. Selecting such a term displays the table with the GM entry links in the search frame.

The query sequence can be entered as simple amino acid sequence or in FASTA format. Either all or only databases from selected species can be searched. The ontology terms in the list frame are grouped by the different ontologies they belong to. Each entry consists of a number of dots indicating the terms position (depth) in the hierarchy of the ontology, followed by the term name (forming a link to the corresponding pre-calculated table which serves as entry point to the GM and also shows the complete upwards branch of the ontology tree). Underneath the term a list of gene names, which are part of the BLAST result and are associated to that term, as well as the weighting number for this term, are shown. Each name forms a link to the corresponding entry position in the BLAST result, allowing to look-up its E-value and with that the degree of similarity to the query sequence. The weighting numbers are used to sort the term list and provide a very simplistic way to judge the likely quality of the potential association to the query sequence. These numbers are calculated by multiplying the E-values of all sequences in the BLAST result associated with the term in question. They give an indication of how strong the evidence for a term is, relative to other terms. The lower the number, the stronger the sequence similarities and more trustworthy the association. Other important factors are the absolute number of different genes associated with the term and from how many different species they originate.

The standard BLAST output in the search frame shows the gene names as direct links to summary information pages in their respective source databases (which provides an easy possibility to see alternative names and known biological information for each gene) and is followed by the species name in brackets (except for mouse and human genes which can be identified by their ENSEMBL ids). Sequence alignments are only shown if selected in the search form, to speed up the display. If alignments are included the score (bits) numbers for those results are links and can be used to jump to the corresponding display.

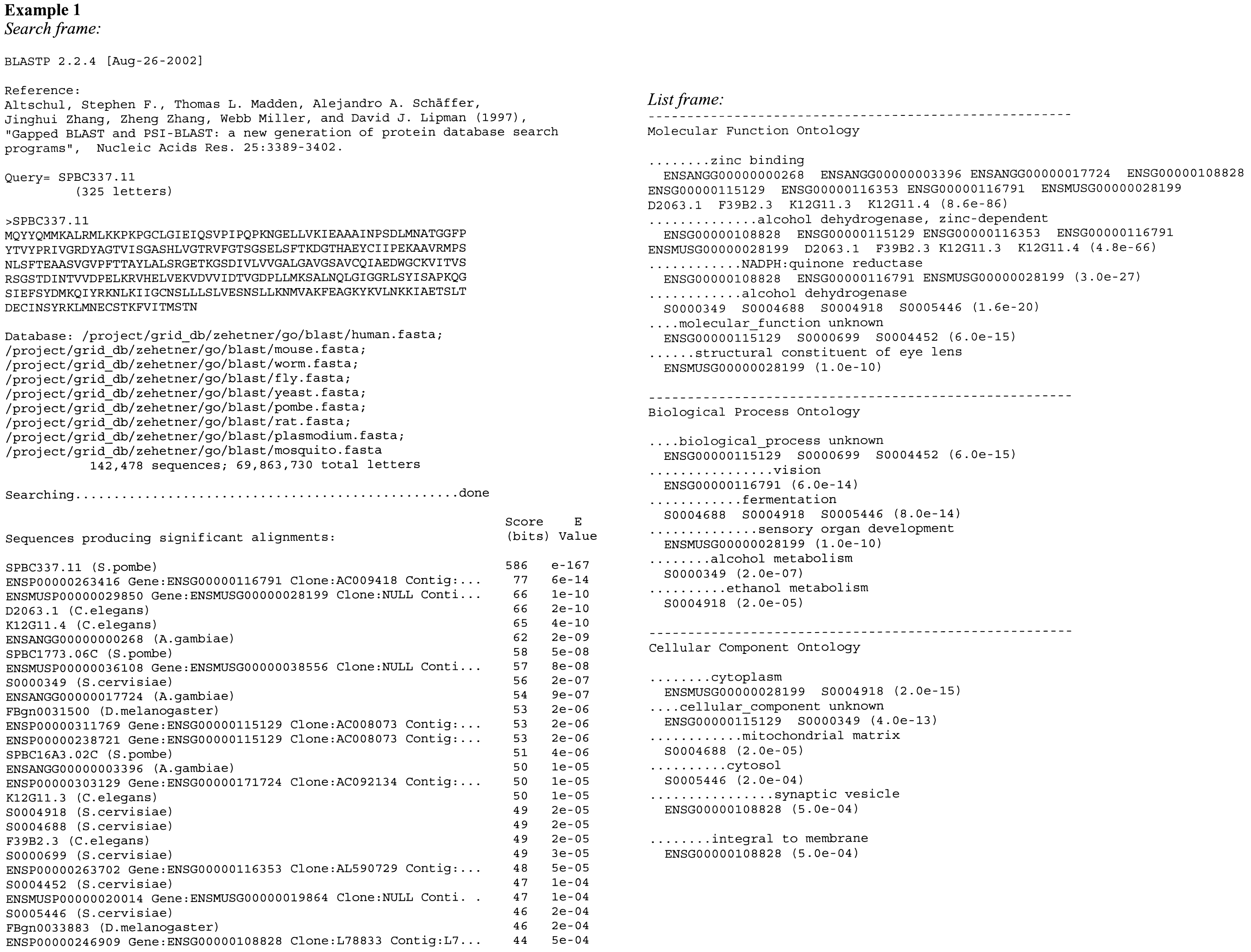
RESULTS

Example searches

The search only requires to paste an amino acid sequence into the query field. The default settings (search all databases, E-value threshold 0.001, cumulative mode off) should usually be kept for the first search. If the cumulative mode is on, the whole trees with all parent terms of the matching GO terms are included, but this can easily obscure the listing. If no matching sequences are found, the E-value threshold can be increased, although this also increases the likelihood of false positive results. After clicking the BLAST button the BLAST search is performed and the two result frames are displayed after a short time, depending on the length of the query sequence and the free capacity on the server (BLAST version 2.2.4 with default query sequence filtering).

In order to check the reliability of the function, a large number of amino acid sequences have been used for searching and the results have been compared with known data. Sequences have been chosen, for which the underlying gene has no direct association to GO terms recorded, but has links to either InterPro (15) entries or other information, suggesting a certain functionality of the gene product. The GO terms found by this tool can therefore not originate directly from the query sequence, but only from other sequences which show similarities to it.

Figure 1 shows the content of the two result frames after searching with the *Schizosaccharomyces pombe* sequence SPBC337.11. The GeneDB entry for this gene does not provide any GO ids, but suggests as related function 'Zinc-containing alcohol dehydrogenase superfamily' (InterPro, IPR002085) and 'Zinc-binding dehydrogenase' (Pfam). The BLAST search result shows 26 sequences (E-value set to default 0.001) originating from seven different species. The weighted list of GO terms clearly confirms with the GeneDB functions. Eleven matching sequences from four species are associated with the molecular function ontology term 'zinc binding' and nine sequences from three species with the term 'alcohol dehydrogenase, zinc dependent' (these two GO terms



Example 1
Search frame:

```
BLASTP 2.2.4 [Aug-26-2002]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= SPBC337.11
         (325 letters)

>SPBC337.11
MQYYQMMKALRMLKKPKPGCLGIEIQSVPIPQPKNGELLVKIEAAAINPSDLMNATGGFP
YTVYPRIVGRDYAGTVISGASHLVGTRVFGTSGSELSFTKDGTHAEYCIIPEKAAVRMPS
NLSFTEAASVGVPFTTAYLALSRGETKGSDIVLVVGALGAVGSAVCQIAEDWGCKVITVS
RSGSTDINTVVDPELKRVHELVEKVDVVIDTVGDPLLMKSALNQLGIGGRLSYISAPKQG
SIEFSYDMKQIYRKNLKIIGCNSLLLSLVESNSLLKNMVAKFEAGKYKVLNKKIAETSLT
DECINSYRKLMNECSTKFVITMSTN

Database: /project/grid_db/zehetner/go/blast/human.fasta;
/project/grid_db/zehetner/go/blast/mouse.fasta;
/project/grid_db/zehetner/go/blast/worm.fasta;
/project/grid_db/zehetner/go/blast/fly.fasta;
/project/grid_db/zehetner/go/blast/yeast.fasta;
/project/grid_db/zehetner/go/blast/pombe.fasta;
/project/grid_db/zehetner/go/blast/rat.fasta;
/project/grid_db/zehetner/go/blast/plasmodium.fasta;
/project/grid_db/zehetner/go/blast/mosquito.fasta
           142,478 sequences; 69,863,730 total letters

Searching..................................................done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

SPBC337.11 (S.pombe)                                              586   e-167
ENSP00000263416 Gene:ENSG00000116791 Clone:AC009418 Contig:...     77   6e-14
ENSMUSP00000029850 Gene:ENSMUSG00000028199 Clone:NULL Conti...     66   1e-10
D2063.1 (C.elegans)                                                66   2e-10
K12G11.4 (C.elegans)                                               65   4e-10
ENSANGG00000000268 (A.gambiae)                                     62   2e-09
SPBC1773.06C (S.pombe)                                             58   5e-08
ENSMUSP00000036108 Gene:ENSMUSG00000038556 Clone:NULL Conti...     57   8e-08
S0000349 (S.cervisiae)                                             56   2e-07
ENSANGG00000017724 (A.gambiae)                                     54   9e-07
FBgn0031500 (D.melanogaster)                                       53   2e-06
ENSP00000311769 Gene:ENSG00000115129 Clone:AC008073 Contig:...     53   2e-06
ENSP00000238721 Gene:ENSG00000115129 Clone:AC008073 Contig:...     53   2e-06
SPBC16A3.02C (S.pombe)                                             51   4e-06
ENSANGG00000003396 (A.gambiae)                                     50   1e-05
ENSP00000303129 Gene:ENSG00000171724 Clone:AC092134 Contig:...     50   1e-05
K12G11.3 (C.elegans)                                               50   1e-05
S0004918 (S.cervisiae)                                             49   2e-05
S0004688 (S.cervisiae)                                             49   2e-05
F39B2.3 (C.elegans)                                                49   2e-05
S0000699 (S.cervisiae)                                             49   3e-05
ENSP00000263702 Gene:ENSG00000116353 Clone:AL590729 Contig:...     48   5e-05
S0004452 (S.cervisiae)                                             47   1e-04
ENSMUSP00000020014 Gene:ENSMUSG00000019864 Clone:NULL Conti. .     47   1e-04
S0005446 (S.cervisiae)                                             46   2e-04
FBgn0033883 (D.melanogaster)                                       46   2e-04
ENSP00000246909 Gene:ENSG00000108828 Clone:L78833 Contig:L7...     44   5e-04
```

List frame:

```
------------------------------------------------------
Molecular Function Ontology

........zinc binding
  ENSANGG00000000268  ENSANGG00000003396 ENSANGG00000017724  ENSG00000108828
ENSG00000115129  ENSG00000116353 ENSG00000116791  ENSMUSG00000028199
D2063.1  F39B2.3  K12G11.3  K12G11.4 (8.6e-86)
..............alcohol dehydrogenase, zinc-dependent
  ENSG00000108828  ENSG00000115129 ENSG00000116353  ENSG00000116791
ENSMUSG00000028199  D2063.1  F39B2.3 K12G11.3  K12G11.4 (4.8e-66)
............NADPH:quinone reductase
  ENSG00000108828  ENSG00000116791 ENSMUSG00000028199 (3.0e-27)
............alcohol dehydrogenase
  S0000349  S0004688  S0004918  S0005446 (1.6e-20)
....molecular_function unknown
  ENSG00000115129  S0000699  S0004452 (6.0e-15)
......structural constituent of eye lens
  ENSMUSG00000028199 (1.0e-10)

------------------------------------------------------
Biological Process Ontology

....biological_process unknown
  ENSG00000115129  S0000699  S0004452 (6.0e-15)
................vision
  ENSG00000116791 (6.0e-14)
............fermentation
  S0004688  S0004918  S0005446 (8.0e-14)
..............sensory organ development
  ENSMUSG00000028199 (1.0e-10)
........alcohol metabolism
  S0000349 (2.0e-07)
..........ethanol metabolism
  S0004918 (2.0e-05)

------------------------------------------------------
Cellular Component Ontology

........cytoplasm
  ENSMUSG00000028199  S0004918 (2.0e-15)
....cellular_component unknown
  ENSG00000115129  S0000349 (4.0e-13)
............mitochondrial matrix
  S0004688 (2.0e-05)
..........cytosol
  S0005446 (2.0e-04)
................synaptic vesicle
  ENSG00000108828 (5.0e-04)

........integral to membrane
  ENSG00000108828 (5.0e-04)
```

Figure 1. The content of the two result frames after searching with the *S.pombe* sequence SPBC337.11.

Table 2. Results of the search with *S.pombe* gene SPBC146.09c (SWISS-PROT/TrEMBL ID: Q9Y802; E-value: 0.001)

BLAST result	GO associations	Known information
16 sequences from six different species	*Molecular Function* Electron transport (6/3/4.7e−54)	GeneDB/*S.pombe*: no GO Ids InterPro: electron transport (GO:0006118)

The only GO term found with OB is identical with the GO term indicated in the InterPro database.

are the same terms which are provided by the InterPro entry IPR002085).

The association to 'NADPH:quinone reductase' is also supported by the InterPro description, which mentions that this family includes NADP-dependent quinone oxidoreductase. It also states that the enzyme has been recruited as an eye lens protein in some species. This correlates with the associations to 'structural constituent of eye lens' (molecular function ontology), as well as 'vision' and 'sensory organ development' (biological process ontology).

Tables 2–5 show summary results of further searches with various sequences. Many more searches have been performed which are not shown but gave similar positive results. Relevant information found in species specific databases or InterPro are indicated to allow a comparison with the search results obtained with OB. In the column labelled 'BLAST result' the number of similar sequences is shown (using the indicated E-value) together with the number of different species they originated from. The column labelled 'GO associations' lists all GO terms associated with the sequences shown in the

Table 3. Results of the search with *D.melanogaster* gene FBgn0004395 (SWISS-PROT/TrEMBL ID: Q960U9; E-value: 0.001)

BLAST result	GO associations	Known information
36 sequences from seven different species	*Molecular Function* Nucleic acid binding (2/1/2.4e−102), apoptosis inhibitor (5/3/4.2e−23), apoptosis regulator (1/1/1.0e−04), cytoskeletal protein binding (1/1/5.0e−4) *Biological Process* Anti-apoptosis (5/3/4.2e−23), apoptosis (2/1/7.0e−10), cell surface receptor linked signal transduction (1/1/1.0e−4), neurogenesis (1/1/5.0e−4), cell motility (1/1/5.9e−4) *Cell Component* Intracellular (3/2/7.0e-14), cytosol (1/1/2.0e−5), non-muscle myosin (1/1/5.0e−4), cytoskeleton (1/1/5.0e−4)	FlyBase: no GO Ids, similar genes in four species, three involve apoptosis inhibitor InterPro information for Zinc finger domain contained in protein: nucleic acid binding (GO: 0003676)

The most significant match is the same as in InterPro, followed by the apoptosis related terms.

Table 4. Results of the search with *D.melanogaster* gene FBgn0010292 (SWISS-PROT/TrEMBL ID: P51406; E-value: 0.001)

BLAST result	GO associations	Known information
10 sequences from seven different species	*Molecular Function* Molecular_function unknown (2/2/5.0e−175) Cell adhesion molecule (1/1/1.0e−106) *Biological Process* Cell adhesion (2/2/1.0e−211) Pregnancy (1/1/1.0e−106) Larval development (sensu Nematoda) (1/1/8.0e−94) Growth (1/1/8.0e−94) Cell growth and/or maintenance (1/1/5.0e−70) *Cell Component* Cytoplasm (1/1/1.0e−106) Cellular_component unknown (1/1/1.0e−105) Nucleus (1/1/5.0e−70)	FlyBase: no GO Ids, similar genes in five species, belonging to the bystin family

The found GO terms fit to the role of the suggested protein family. Bystin is involved in embryo implantation, it mediates cell adhesion between trophoblastic cells and endometrial epithelial cells at the respective apical cell membranes.

Table 5. Results of the search with *C.elegans* gene B0303.11 (SWISS-PROT/TrEMBL ID: P34261; E-value: 0.001)

BLAST result	GO associations	Known information
25 sequences from six different species	*Molecular Function* Amino acid-polyamine transporter (7/3/9.4e−174) Transporter (5/2/4.7e−165) Sodium:chloride symporter (2/2/2.4e−114) Carrier (4/1/8.4e−85) Sodium:chloride/potassium:chloride symporter (2/2/4.2e−64) ATP binding (1/1/2.0e−56) Cation:chloride cotransporter (1/1/1.0e−5) Protein binding (1/1/1.0e−5) Potassium:chloride symporter (1/1/7.0e−5) *Biological Process* Amino acid transport (7/3/9.4e−174) Ion transport (5/2/4.7e−165) Small molecule transport (3/1/3.9e−111) Transport (4/1/8.4e−85) Regulation of cell volume (1/1/7.0e−5) *Cell Component* Membrane (7/3/1.0e−231) Integral to plasma membrane (6/2/4.7e−137) Membrane fraction (4/2/7.8e−115) Integral to membrane (3/2/1.2e−59) Plasma membrane (1/1/7.0e−59)	WormBase: no GO Ids, contains similarity to renal Na-K-Cl cotransporter isoform A and amino acid permease

Most of the found GO terms relate to the predicted transporter functions, which is also consistent with the membrane related cell component terms.

previous column (the numbers in brackets show the number of sequences associated with the GO term, followed by the number of different species they belong to, followed by the weighting number of the term).

CONCLUSION

OntoBlast provides a quick and simple way to test if a potential function can be predicted for an unknown sequence, if similar sequences associated with an ontology entry can be found in any of the searchable species databases and which other genes share those ontology terms. While results can certainly contain a number of false positive ontology terms, which have been selected by insignificant sequence similarities, the examples show, that if a positive functional correlation between similar sequences exists, it can be highlighted by the OB tool. A large number of sequences from all included species with known function have been tested in order to check the reliability and specificity of the returned results. Many functions suggested by the resulting ontology terms showed a clear and correct correlation to the described known protein information. In general, sequence similarities with an E-value $<1.0e{-}4$ gave reasonable assignments to ontology terms. Replacing the very simple weighting number, which is now used to sort the ontology term list, by a more sophisticated mechanism, including the statistical likelihood of associating certain GO terms with genes from certain species, would probably allow to distinguish even more clearly between significant and random assignments. The simultaneous comparison to all genes from nine species gives the advantage to often find assignments supported by hits from two, three or more species.

Further analysis is greatly assisted by direct links from the resulting ontology terms to all associated genes in the Genome–Matrix database (including their surrounding gene regions linked to ortholog genes from other species), providing a unique and direct access to a large collection of relevant structural and functional information from many disperse data sources.

ACKNOWLEDGEMENTS

The Genome–Matrix is a project of the Vertebrate Genomics department (Head Hans Lehrach) of the Max-Planck-Institute for Molecular Genetics, Berlin (MPI) in cooperation with the Resource Centre of the German Human Genome Project gGmbH, Berlin (RZPD). Human and mouse specific matrixes have been compiled by Steffen Hennig (MPI), ortholog data calculated by Alia Ben Kahla (MPI), GUI and database programming by Andreas Hewelt, Qing Dong and Ram Narang (RZPD).

REFERENCES

1. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
2. Camon,E. Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Dinn,T., Maslen,J., Cox,A. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL and InterPro. *Genome Res.*, **13**, 662–672.
3. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet.*, **31**, 316–319.
4. Hewelt,A., Ben Kahla,A., Hennig,S., Nagel,A., Himmelbauer,H., Zehetner,G., Haas,S., Vingron,M., Yaspo,M.L. and Lehrach,H. (2002) The GenomeMatrix Information Retrieval System, Poster Abstracts of HGM2002 (Human Genome Meeting, April 14–17, 2002, Shanghai, China). Genome Informatics and Annotation, Abstract 23.
5. 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.
6. 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.
7. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
8. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
9. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al.* (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
10. The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
11. Steen,R.G., Kwitek-Black,A.E., Glenn,C., Gullings-Handley,J., Van Etten,W., Atkinson,O.S., Appel,D., Twigger,S., Muir,M., Mull,T. *et al.* (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.*, **9**, AP1–AP8.
12. Weng,S., Dong,Q., Balakrishnan,R., Christie,K., Costanzo,M., Dolinski,K., Dwight,S.S., Engel,S., Fisk,D.G., Hong,E. *et al.* (2003) Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.
13. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al.* (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
14. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., 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.
15. 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.