D224–D228 *Nucleic Acids Research, 2007, Vol. 35, Database issue*
doi:10.1093/nar/gkl841

New developments in the InterPro database

Nicola J. Mulder[1,*], Rolf Apweiler[1], Teresa K. Attwood[3], Amos Bairoch[4,5], Alex Bateman[2], David Binns[1], Peer Bork[6], Virginie Buillard[4], Lorenzo Cerutti[4], Richard Copley[7], Emmanuel Courcelle[8], Ujjwal Das[1], Louise Daugherty[1], Mark Dibley[9], Robert Finn[2], Wolfgang Fleischmann[1], Julian Gough[10], Daniel Haft[11], Nicolas Hulo[4], Sarah Hunter[1], Daniel Kahn[12], Alexander Kanapin[1], Anish Kejariwal[13], Alberto Labarga[1], Petra S. Langendijk-Genevaux[4], David Lonsdale[1], Rodrigo Lopez[1], Ivica Letunic[6], Martin Madera[14] , John Maslen[1], Craig McAnulla[1], Jennifer McDowall[1], Jaina Mistry[2], Alex Mitchell[1,3], Anastasia N. Nikolskaya[15], Sandra Orchard[1], Christine Orengo[9], Robert Petryszak[1], Jeremy D. Selengut[11], Christian J. A. Sigrist[4], Paul D. Thomas[13], Franck Valentin[1], Derek Wilson[14], Cathy H. Wu[15] and Corin Yeats[9]

[1]EMBL Outstation—European Bioinformatics Institute and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [3]Faculty of Life Sciences and School of Computer Science, University of Manchester, Manchester, UK, [4]Swiss Institute for Bioinformatics, Geneva, Switzerland, [5]Department of Structural Biology and Bioinformatics, University of Geneva, Switzerland, [6]Biocomputing Unit EMBL, Heidelberg, Germany, [7]Wellcome Trust Centre for Human Genetics, Oxford, UK, [8]CNRS/INRA, Toulouse, France, [9]Biochemistry and Molecular Biology Department, University College London, University of London, UK, [10]Genomic Sciences Centre, RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Japan, [11]The Institute for Genomic Research, Rockville, MD, USA, [12]Laboratoire de Biométrie et Biologie Evolutive and INRIA HELIX Project, University Lyon 1, France, [13]Evolutionary Systems Biology Group, SRI International, Menlo Park, CA, USA, [14]MRC Laboratory of Molecular Biology, Cambridge, UK and [15]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA

Received September 5, 2006; Revised and Accepted October 6, 2006

ABSTRACT

InterPro is an integrated resource for protein families, domains and functional sites, which integrates the following protein signature databases: PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER. The latter two new member databases have been integrated since the last publication in this journal. There have been several new developments in InterPro, including an additional reading field, new database links, extensions to the web interface and additional match XML files. InterPro has always provided matches to UniProtKB proteins on the website and in the match XML file on the FTP site. Additional matches to proteins in UniParc (UniProt archive) are now available for download in the new match XML files only. The latest InterPro release (13.0) contains more than 13 000 entries, covering over 78% of all proteins in UniProtKB. The database is available for text- and sequence-based searches via a webserver (http://www.ebi.ac.uk/interpro), and for download by anonymous FTP (ftp://ftp.ebi.ac.uk/pub/databases/interpro). The InterProScan search tool is now also available via a web service at http://www.ebi.ac.uk/Tools/webservices/WSInterProScan.html.

INTRODUCTION

InterPro (1) incorporates the major protein signature databases into a single resource. These include: PROSITE (2), which uses regular expressions and profiles, PRINTS (3), which uses Position Specific Scoring Matrix-based (PSSM-based) fingerprints, ProDom (4), which uses automatic sequence clustering, and Pfam (5), SMART (6), TIGRFAMs (7), PIRSF (8), SUPERFAMILY (9), Gene3D (10) and PANTHER (11), all of which use hidden Markov models (HMMs). Table 1 shows the coverage of each of these member

*To whom correspondence should be addressed. Tel: +44 1223 494 602; Fax: +44 1223 494 468; Email: mulder@ebi.ac.uk
Present address:
Julian Gough, Unite de Bioinformatique Structurale, Institut Pasteur, Paris, France

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Coverage of protein sequences and amino acid residues for each member database

Member database	Number of methods in InterPro[a]	Total number of proteins hit by database	Total number of residues covered	Number of unique proteins hit by database[b]
Gene3D	1465	1 736 593	395 970 746	18 504
PANTHER	39 648	582 799	173 969 368	6355
PIRSF	1347	161 248	58 525 186	851
PRINTS	1900	645 272	55 137 257	3936
PROSITE patterns	1336	766 422	16 861 589	14 229
PROSITE profiles	632	763 334	153 498 831	2131
Pfam	8296	2 502 476	570 591 566	281 062
ProDom	3538	506 284	61 153 722	19 926
SMART	706	514 466	94 310 609	2252
SUPERFAMILY	1122	1 929 112	484 789 136	51 282
TIGRFAMs	2625	501 897	170 121 752	7306

[a]Not all the methods are integrated into InterPro entries, e.g. for PANTHER, but InterPro provides matches to them in the match XML file.
[b]This is the number of proteins hit by one database only.

databases. Protein signatures from these databases that describe the same family or domain, in terms of sequence positions and protein coverage are integrated into single InterPro entries, to which are added annotation and cross-references. Annotation includes an abstract, name, short name and GO terms (12) (where applicable). Cross-references are provided to specialized databases and protein structural information. All matches of the protein signatures contributed by member databases against the UniProt Knowledgbase (UniProtKB) (13) are calculated using the InterProScan software (14), which integrates the search algorithms from the member databases into a single package. The matches are available for viewing in various formats for each InterPro entry. The InterPro data are available for searching and retrieval via a web interface at http://www.ebi.ac.uk/interpro, and for download by anonymous FTP ftp://ftp.ebi.ac.uk/pub/databases/interpro.

InterPro is constantly being updated to keep up with the changing face of Bioinformatics. Two new member databases, PANTHER and Gene3D, have joined the InterPro consortium and their HMMs are being integrated. In addition, new database cross-references to CluSTr (15) and Pfam clans (5) have been included, and entries link to the IntAct molecular interaction database (16) where manually curated examples of domain–domain interactions are available. Proteins with 3D structures modelled by MODBASE (17) and SWISS-MODEL (18) have links to the structure predictions from the match graphical views. These links complement the experimentally determined structures in the protein data bank (PDB). The web interface has been extended for more advanced searching capabilities, and a web service is now available, providing programmatic access to InterProScan. In addition to UniProtKB, InterPro now provides matches to all proteins in the UniProt archive, UniParc, and these are currently available in XML format on the FTP site. The match XML files are also indexed in SRS to allow users to query the data within the SRS interface. The new features of InterPro are described in more detail below.

NEW FEATURES OF INTERPRO

Annotation

Two new member databases have been integrated into InterPro, PANTHER and Gene3D. PANTHER (http://www.pantherdb.org) (11) HMMs define protein families and subfamilies modelled on the divergence of specific functions within the families, which permits more accurate association with function based on ontology terms and pathways, as well as inference of amino acids important for functional specificity. PANTHER currently has high coverage of all families that contain at least one metazoan protein, including homologous proteins from all taxa. Consequently, coverage is very high for proteins found in animals and less so for other groups, such as plants, fungi and bacteria. The addition of PANTHER HMMs to InterPro is facilitating more fine-grained annotation of functionally and evolutionarily related subfamilies. Gene3D (http://cathwww.biochem.ucl.ac.uk:8080/Gene3D/) (10) is a library of HMMs that represent all proteins of known structure. The seed alignments for the models are derived from the proteins found within the homologous superfamily (H-level) classification level in CATH, which groups together domains that are thought to share a common ancestor. Gene3D models are being integrated to complement the SUPERFAMILY models that are based on SCOP superfamilies.

To further extend the publications section of InterPro entries, we have introduced the 'additional reading' field. This field lists any publications provided by the member databases for the methods associated with each InterPro entry, which are not directly referenced in the InterPro abstract. Additionally, a maximum of five references per entry are taken from the PDB when one or more of the proteins in the entry has had its structure determined. These references provide the user with additional publications to visit to find out more about the proteins in the entry, and also provide InterPro curators with a list of references to consult when updating abstracts.

The 'database links' field has been extended to include new links to CluSTr and Pfam clans. Table 2 lists the databases cross-referenced in InterPro and the number of entries containing these links. CluSTr (http://www.ebi.ac.uk/clustr) (15) is a database containing protein clusters from more than 368 organisms with completely sequenced genomes. The clustering is based on pairwise comparisons between the protein sequences. InterPro entries are linked to protein clusters only where at least 70% of the CluSTr members occur in the InterPro entry. Links to Pfam clan pages are now available in the database links field where applicable.

Table 2. Number of InterPro entries with cross-references to the databases InterPro provides links to

Database	Number of InterPro entries with links
UniProtKB	13 131
BLOCKS	6134
CAZy	119
COMe	204
IntEnz	2336
IUPHAR receptor	113
MEROPS	548
PANDIT	7702
PROSITE doc	1479
Pfam Clans	1544
CluSTr	6818
IntAct	135
GO	7131
MSDsite	1313
PDB	68 021
SCOP	6537
CATH	6212

A clan contains two or more Pfam families that have arisen from a single evolutionary origin, based on evidence from structure, function, profile–profile comparisons and whether the sequences are matched by more than one HMM. Clans were introduced to resolve the issue of Pfam HMMs overlapping on a sequence, as this is forbidden in the Pfam database. Clan information is used in post-processing of matches to remove these overlaps. The link from InterPro entries to clans provides a popup display of the Pfam clan name and all Pfam clan members with their corresponding InterPro accession numbers. These InterPro entries will not necessarily be related to each other through parent/child or contains/found in relationships.

Links to IntAct (http://www.ebi.ac.uk/intact/site/) (16), the molecular interaction database, have been incorporated into InterPro, providing manually curated examples of domain–domain interactions. IntAct incorporates protein–protein interaction data derived from the literature and direct submissions, and provides a query interface and modules to analyze the data. Links from InterPro to IntAct are provided at the level of individual UniProtKB accessions, and are restricted to 20 randomly chosen examples. There are currently 135 InterPro entries with links to 1180 IntAct entries, involving ~400 proteins. This number is likely to remain low, compared to the total number of interactions in IntAct, as these links are based on well curated domain interactions, rather than every protein–protein interaction.

New positional links are available for UniProtKB proteins to MODBASE (http://modbase.compbio.ucsf.edu/modbase-cgi-new/index.cgi) (17) and SWISS-MODEL (http://swissmodel.expasy.org/) (18). MODBASE is a database of 3D protein models calculated by comparative modelling using ModPipe, an automated modelling pipeline relying on programs, such as PSI-BLAST and MODELLER. MODBASE matches to protein sequences are shown in the detailed graphical view as yellow and white striped bars. SWISS-MODEL is a repository of annotated 3D protein structure models from the UniProtKB sequence database, and provides a protein structure homology modelling server. Matches to protein sequences are shown in the detailed graphical view as red and white striped bars. These cross-references, as well the other links to more than 30 different databases, increase the value of InterPro with respect to its interoperability and integration with other data sources.

Protein matches

Protein matches in InterPro are pre-calculated using the InterProScan software (14). InterProScan is a tool that combines different protein signature recognition methods of the InterPro member databases into one resource, and provides the corresponding InterPro accession numbers and GO annotation in the results. InterProScan can be used via a web interface or email server, which allows searching of a sequence against InterPro, or it can be installed and run locally for bulk searches. A new development has been the establishment of a web service for running single or multiple sequences through InterProScan. More information about the web service and example clients in Perl and Java for accessing the service is available from http://www.ebi.ac.uk/Tools/webservices/WSInterProScan.html. This service provides programmatic access to the tool for users who want to run bulk searches or use InterProScan as part of a pipeline.

Over the past two years, additional protein matches have become available in InterPro. Previously, InterPro matches were available only for UniProtKB proteins, but now InterPro provides additional matches to alternative splice products and UniParc proteins. Matches to splice variant sequences associated with UniProtKB accession numbers can be accessed through the 'protein with splice variants' link from the Matches field, and are available through the compact and detailed displays. The matches for the master sequence are shown at the top with the splice variant matches below them, so it is easy to identify where matches differ between isoforms. The splice variant sequences originate from UniProtKB, and of the 25 927 splice variants available, 24 268 have hits to a total of 3483 InterPro entries.

The UniProt archive (UniParc) is a repository of all protein sequences, with each unique sequence stored once. These sequences are then cross-referenced to the relevant databases, e.g. UniProtKB, and include data submitted from metagenomics projects. This repository contains ~7.5 million protein sequences, including UniProtKB proteins, and therefore the calculation of InterPro matches is slow. These calculations are ongoing, and the data provided incorporates the most up-to-date matches available at that point in time. Currently, there are just over 50 million InterPro matches to UniParc proteins. UniParc matches are not yet visible in InterPro entries, but are available in XML format from the FTP site and are searchable in SRS. An additional match XML file, match_complete.xml, is provided with each release, and contains UniProtKB sequence matches for all member database signatures, including those that have not yet been integrated into InterPro. This is to ensure that the public has access to all protein signature matches that have been calculated. All protein matches are updated on each major InterPro release (approximately every 3 months).

Web interface

The web interface has been extended to provide additional searching options. From the text search page (http://www.ebi.ac.uk/interpro/search.html) the user can search within

InterPro entries or protein matches. One can retrieve matches for a UniProtKB accession number by pasting the accession number in the search box and selecting 'Find protein matches'. This returns the matches in a combination of formats. The protein match views can also be selected in the Matches section of an InterPro entry, which provides options for displaying the matches in different tabular or graphical views. From any of these views, the user can then select a set of proteins by UniProtKB accession number(s) or InterPro accession, and can refine the set to show splice variants or proteins with known structure or both. Alternatively, the user can filter the protein set by taxonomy using the 'tax ID'. Once the protein set has been defined, the user can select the output display format from 'compact', 'detailed', 'architectures' or 'table', and can specify the order of proteins in the display by UniProtKB accession or identifier.

In addition to links to complete match lists, each InterPro entry page contains a taxonomy wheel showing the taxonomic range of proteins matching the entry. The numbers on the wheel for each taxonomic group are now 'clickable'. Clicking on a particular lineage returns only the protein matches for the selected taxonomy. In this view, the species are sorted and displayed alphabetically and the lineage is shown at the top. The numbers on the phylogeny show the number of proteins associated with each taxonomic group that match the entry.

DISCUSSION

InterPro now integrates protein signatures from 10 different member databases, and links >20 additional resources, including UniProtKB, structural data and specialized protein family databases. It has proven its usefulness in the functional characterization of proteins, and is used by genome annotation projects (19–22) and individual researchers worldwide. In the last year, the InterPro website received ~3 million hits per month from up to 35 000 unique hosts. Through the mapping of InterPro entries to GO terms, InterPro contributes the majority of annotations of proteins to GO terms. Approximately 68% of all UniProtKB proteins are annotated with GO terms from a combination of manual annotation and the use of mappings, such as InterPro2GO, Swiss-Prot keyword2GO, etc. InterPro2GO alone provides GO annotations for 61% of UniProtKB proteins, thus accounting for a significant proportion of the total number of annotations currently available. These GO mappings are also available via InterProScan, which facilitates GO annotation to query proteins. The current release of InterPro contains more than 13 000 entries, with its signatures covering over 78% of UniProtKB proteins. The integration of new protein signatures from the existing and new member databases will continue to increase the coverage, as well as the depth, of InterPro.

The InterPro database will continue to develop and increase its functionality. Future plans include the provision of protein match views for UniParc matches, facilitating the searching and browsing of InterPro entries by function, and the provision of data for unintegrated protein signatures via the InterPro web interface. Integration of signatures into InterPro entries and subsequent annotation of the entries is done manually and is thus of high-quality, but is time-consuming. In order to make the signatures awaiting integration available to the public via the web interface, new entries will be created automatically for the unintegrated signatures and will be searchable by their member database accession numbers. The protein matches will be available in the same format as match views from InterPro entries so that the user can see how the new signature relates to existing entries. These new features will increase the usefulness of this already popular high-quality resource.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Steffen Schulze-Kremer and the HLRN staff for their continued and valuable assistance. InterPro is funded in part by the MRC e-family grant number G0100305. A large proportion of HMMER-based calculations are performed on the IBMP690 Supercomputer at HLRN. Funding to pay the Open Access publication charges for this article was provided by the European Bioinformatics Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
2. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
3. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
4. Bru,C., Courcelle,E., Carrere,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.
5. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
6. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
7. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
8. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
9. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
10. Yeats,C., Maibaum,M., Marsden,R., Dibley,M., Lee,D., Addou,S. and Orengo,C.A. (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res.*, **34**, D281–D284.
11. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O., Campbell,M.J. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**, D284–D288.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

13. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
14. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
15. Petryszak,P., Kretschmann,E., Wieser,D. and Apweiler,R. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.
16. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
17. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.
18. Kopp,J. and Schwede,T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.*, **34**, D315–D318.
19. The International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
20. Kawaji,H., Schonbach,C., Matsuo,Y., Kawai,J., Okazaki,Y., Hayashizaki,Y. and Matsuda,H. (2002) Exploration of novel motifs derived from mouse cDNA sequences. *Genome Res.*, **12**, 367–378.
21. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
22. Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., LiP,W., Apweiler,R., Fleischmann,W. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.