Nucleic Acids Research, 2003, Vol. 31, No. 1 315–318
DOI: 10.1093/nar/gkg046

The InterPro Database, 2003 brings increased coverage and new features

Nicola J. Mulder[1,*], Rolf Apweiler[1], Teresa K. Attwood[3], Amos Bairoch[4], Daniel Barrell[1], Alex Bateman[2], David Binns[1], Margaret Biswas[5], Paul Bradley[1,3], Peer Bork[6], Phillip Bucher[7], Richard R. Copley[8], Emmanuel Courcelle[9], Ujjwal Das[1], Richard Durbin[2], Laurent Falquet[7], Wolfgang Fleischmann[1], Sam Griffiths-Jones[2], Daniel Haft[10], Nicola Harte[1], Nicolas Hulo[4], Daniel Kahn[9], Alexander Kanapin[1], Maria Krestyaninova[1], Rodrigo Lopez[1], Ivica Letunic[6], David Lonsdale[1], Ville Silventoinen[1], Sandra E. Orchard[1], Marco Pagni[7], David Peyruc[9], Chris P. Ponting[11], Jeremy D. Selengut[10], Florence Servant[1], Christian J. A. Sigrist[4], Robert Vaughan[1] and Evgueni M. Zdobnov[6,12]

[1]EMBL Outstation—European Bioinformatics Institute and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [3]School of Biological Sciences and Department of Computer Science, The University of Manchester, Manchester, UK, [4]Swiss Institute for Bioinformatics, Geneva, Switzerland, [5]ViaLactia Biosciences, Newmarket Auckland, New Zealand, [6]Biocomputing Unit EMBL, Heidelberg, Germany, [7]Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland, [8]Wellcome Trust Centre for Human Genetics, Oxford, UK, [9]CNRS/INRA, Toulouse, France, [10]The Institute for Genomic Research, MD, USA, [11]MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, UK and [12]EMBL, Heidelberg, Germany

Received September 16, 2002; Revised and Accepted October 2, 2002

ABSTRACT

InterPro, an integrated documentation resource of protein families, domains and functional sites, was created in 1999 as a means of amalgamating the major protein signature databases into one comprehensive resource. PROSITE, Pfam, PRINTS, ProDom, SMART and TIGRFAMs have been manually integrated and curated and are available in InterPro for text- and sequence-based searching. The results are provided in a single format that rationalises the results that would be obtained by searching the member databases individually. The latest release of InterPro contains 5629 entries describing 4280 families, 1239 domains, 95 repeats and 15 post-translational modifications. Currently, the combined signatures in InterPro cover more than 74% of all proteins in SWISS-PROT and TrEMBL, an increase of nearly 15% since the inception of InterPro. New features of the database include improved searching capabilities and enhanced graphical user interfaces for visualisation of the data. The database is available via a webserver (http://www.ebi.ac.uk/interpro) and anonymous FTP (ftp://ftp.ebi.ac.uk/pub/databases/interpro).

BACKGROUND

Protein signature databases, based on several different methods, have evolved with the need for efficient automatic methods of protein sequence classification and characterisation. In 1999, the major signature databases PROSITE (1), PRINTS (2), Pfam (3) and ProDom (4) formed a Consortium and agreed to integrate their data into a new database that became known as InterPro (5). Subsequently SMART (6) and TIGRFAMs (7) have joined the Consortium. The Consortium has agreed on the free availability and distribution of the data and protein sequence search methods, and free, efficient flow of information between the member databases and InterPro, as well as among themselves.

Signatures from the member databases are integrated manually at regular intervals by a team of biologists, whose role is also to annotate the new or existing entries. Each InterPro entry is described by one or more signatures, corresponding to a biologically meaningful family, domain, repeat or PTM. Two types of relationships can exist between InterPro entries: the parent/child and contains/found in relationship. Parent/child relationships are used to describe a common ancestry between entries whereas the contains/found in relationship generally refers to the presence of genetically mobile domains. All hits of the protein signatures in InterPro against a composite of the SWISS-PROT and TrEMBL databases (8) (SPTR) are precomputed. The matches are

*To whom correspondence should be addressed. Tel: +44 1223 494602; Fax: +44 1223 494468; Email: mulder@ebi.ac.uk

Table 1. Summary of the statistics for all InterPro releases

Release	Date	Entries	SPTR coverage (%)	New feature
1.0	March 2000	2990	60.2	PROSITE, PRINTS, Pfam
2.0	October 2000	3204	63.9	ProDom included
3.0	March 2001	3875	73.3	SMART included
4.0	November 2001	4691	72.2	TIGRFAMs included
5.0	May 2002	5312	74.0	General updates[a]
Currently	September 2002	5629	74.0	General updates[a]

[a]Updates to include new member database releases.

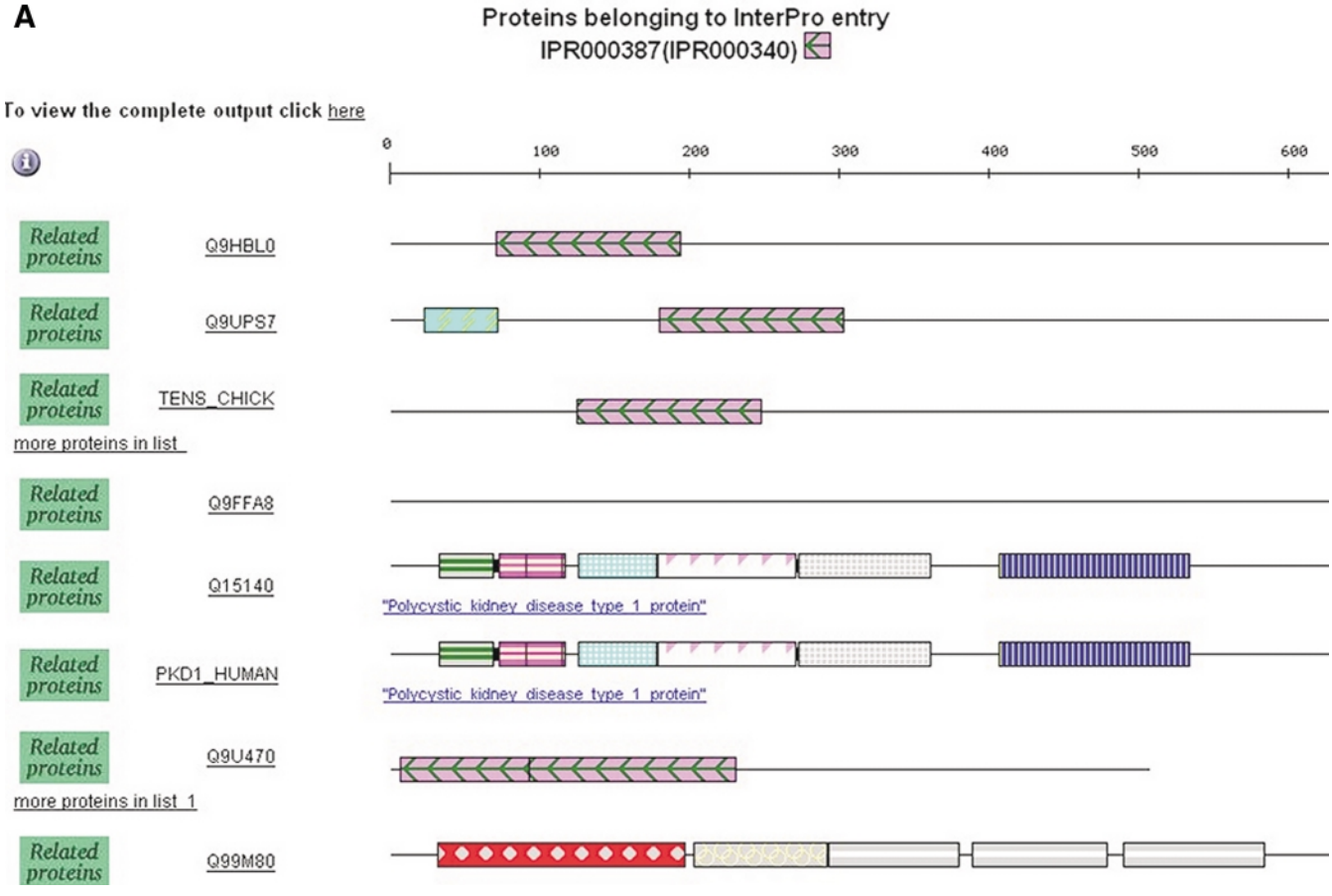
available for viewing in each InterPro entry in different formats including a match table, a detailed graphical view and a condensed graphical view.

There have been a number of improvements to the InterPro database since its inception, including increased coverage, additional features of the search tools, and a new look web interface. These are described in more detail below.

MORE ENTRIES AND INCREASED COVERAGE

The first official release of InterPro in October 1999 contained 2990 entries and covered 60.2% of all SPTR protein sequences. The latest release of the database contains 5629 entries, an increase of 2639 entries, or a doubling in just 3 years. A summary of the InterPro release and the coverage of the signatures in SPTR are shown in Table 1. On average, there has been an increase of 500–600 new entries per release, which does not necessarily correspond with the number of new signatures, since many may overlap with existing entries represented by other member databases.

The coverage of SPTR by InterPro signatures has increased by nearly 15%, a significant figure considering that the SPTR databases themselves have increased from 279 794 to 734 448 protein sequences over the same period of time. There may be



an overlap in coverage by entries which are 'children' of or 'found in' other entries, so a protein may hit several entries. The coverage of InterPro in complete proteomes ranges from 64% to 74% in eukaryotes, with a coverage of 73.5% of the non-redundant human proteome, and averages ~66–68% in prokaryotes, with some having a coverage of up to 75%. Mostly a hit to InterPro provides useful functional information, however, there are ~370 entries that describe 'proteins of unknown function' and hence prevent inference of function. However, these entries do group related proteins and if one protein in the entry is biochemically characterised then this may shed light on the function of the related proteins.

NEW FEATURES

Several new features have been introduced into InterPro since the last publication in this journal in 2000. On the annotation side, InterPro entries have been mapped to Gene Ontology (GO) (10) terms where a term applies to all proteins matching that entry. Not all entries can be mapped due to low specificity in function or process, but for those that can this provides a powerful tool for automatic large scale annotation of proteins to GO terms. Currently, 4102 InterPro entries have been mapped to 1899 unique GO terms, which results in automatic GO assignment to 405 684 unique proteins in SPTR.

A notable improvement in InterPro has been in the searching capabilities. The sequence search package, InterProScan (11), has been extended to include all new member databases and data, and the Perl stand-alone version has additional features, including allowance for GO annotation, and the potential to plug in the transmembrane and signal peptide prediction programs TMHMM (12) and SignalP (13) respectively. InterProScan is available for interactive as well as email sequence submissions. Additional files, for example a list of all InterPro entries, a list of InterPro to GO mappings and a summary of all protein matches are now available on the FTP site. The text search capabilities have been extended to both a simple text search and an SRS-based (14) search facility for more complex queries.

InterPro has developed an improved user interface for visualisation of the protein matches in a condensed graphical view derived from the ProDom graphical interface (4). The consensus domain boundaries are computed, and the resulting protein matches are combined rather than each signature being displayed (Fig. 1A and B). Parent/child related InterPro entries are collapsed into one line, while domain entries are shown on separate line, thereby providing a simple view of family and domain composition. From this view, all proteins sharing a

Figure 1. (Opposite and above). New graphical user interface for viewing protein matches of a particular InterPro entry. (**A**) Graphical view of representative list of proteins matching IPR000340, in which consensus domain boundaries have been computed for the domain line, and parent and children entries have been collapsed into one family line. This enables the family and domain composition information to be seen at a glance. (**B**) From the "more proteins in list" link in view a.) it is possible to show all proteins sharing a common domain architecture These protein sequences can then be retrieved or their alignments can be visualised.

common domain architecture can be grouped, and the sequences aligned and visualised using Jalview (http://www.ebi.ac.uk/~michele/jalview/) or DisplayFam (15). Recently, the general web interface for InterPro has been developed, and changes reflect style changes to the EBI web server. A useful addition to the pages is the option to display them as simple HTML, a printer-friendly version, XML and the default view with or without the menu.

DISCUSSION

The amalgamation of the major protein signature databases into InterPro has proven to be an enormous success, and has produced a powerful tool for protein sequence analysis and characterisation. The tools and data have numerous applications described in more detail elsewhere (16), and InterPro has been the tool of choice for the annotation of new genomes, including the human genome (17). Future plans involve integration of the next database, PIR superfamilies (18), which facilitate protein family information retrieval, identification of domain and family relationships and classification of multi domain proteins. In addition, there are plans for expansion into the field of protein secondary and tertiary structure. Protein structure information is vital in understanding protein function and evolutionary relationships. A project has been initiated to rationalise the data of SCOP (Structural Classification of Proteins) (19), CATH (Class, Architecture, Topology, Homology) (20), and SWISS-MODEL 3D structure homology models (21) with that of InterPro. This integration will enhance the capability of the database in the field of protein classification and characterisation and make the database, a true integrated resource for complete protein sequence and structure information.

The InterPro database is available via a webserver (http://www.ebi.ac.uk/interpro) and anonymous FTP (ftp://ftp.ebi.ac.uk/pub/databases/interpro).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENT

The InterPro project is supported by the ProFuSe grant (no. QLG2-CT-2000-00517) of the European Commission.

REFERENCES

1. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J.A., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
2. Attwood,T.K., Blythe,M.J., Flower,D.R., Gaulton,A., Mabey,J.E., Maudling,N., McGregor,L., Mitchell,A.L., Moulton,G., Paine,K. *et al.* (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.*, **30**, 239–241.
3. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.*, **30**, 276–280.
4. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
5. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
6. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
7. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.
8. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
9. Doerks,T., Copley,R.R., Schultz,J., Ponting,C.P. and Bork,P. (2002) Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.*, **12**, 47–56.
10. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
11. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
12. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
13. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
14. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
15. Corpet,F., Gouzy,J. and Kahn,D. (1999) Browsing protein families via the 'Rich Family Description' format. *Bioinformatics*, **15**, 1020–1027.
16. 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.
17. The International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
18. Wu,C.H., Xiao,C., Hou,Z., Huang,H. and Barker,W.C. (2001) *i*ProClass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res.*, **29**, 52–54.
19. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
20. Pearl,F.M., Lee,D., Bray,J.E., Buchan,D.W., Shepherd,A.J. and Orengo,C.A. (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Sci.*, **11**, 233–244.
21. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modelling. *Electrophoresis*, **18**, 2714–2723.