InterPro in 2011: new developments in the family and domain prediction database

Sarah Hunter, Philip Jones, Alex Mitchell, Rolf Apweiler, Teresa K. Attwood, Alex Bateman, Thomas Bernard, David Binns, Peer Bork, Sarah Burge, et al.

doi:10.1093/nar/gkr948

InterPro in 2011: new developments in the family and domain prediction database

Sarah Hunter[1,*], Philip Jones[1,*], Alex Mitchell[1,*], Rolf Apweiler[1], Teresa K. Attwood[2], Alex Bateman[3], Thomas Bernard[4], David Binns[1], Peer Bork[5], Sarah Burge[1], Edouard de Castro[6], Penny Coggill[3], Matthew Corbett[1], Ujjwal Das[1], Louise Daugherty[1], Lauranne Duquenne[4], Robert D. Finn[3], Matthew Fraser[1], Julian Gough[7], Daniel Haft[8], Nicolas Hulo[6], Daniel Kahn[4], Elizabeth Kelly[9], Ivica Letunic[5], David Lonsdale[1], Rodrigo Lopez[1], Martin Madera[7], John Maslen[1], Craig McAnulla[1], Jennifer McDowall[1], Conor McMenamin[1], Huaiyu Mi[10], Prudence Mutowo-Muellenet[1], Nicola Mulder[9], Darren Natale[11], Christine Orengo[12], Sebastien Pesseat[1], Marco Punta[3], Antony F. Quinn[1], Catherine Rivoire[6], Amaia Sangrador-Vegas[1], Jeremy D. Selengut[8], Christian J. A. Sigrist[6], Maxim Scheremetjew[1], John Tate[3], Manjulapramila Thimmajanarthanan[1], Paul D. Thomas[10], Cathy H. Wu[12], Corin Yeats[12] and Siew-Yit Yong[1]

[1]EMBL Outstation European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, [2]Faculty of Life Science and School of Computer Science, The University of Manchester, M13 9PL, Manchester, [3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK, [4]Pôle Rhône-Alpin de Bio-Informatique (PRABI) and Laboratoire de Biométrie et Biologie Evolutive; CNRS; INRA; Université de Lyon; Université Lyon 1, 69622 Villeurbanne, France, [5]European Molecular Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany, [6]Swiss Institute of Bioinformatics (SIB), CMU - Rue Michel-Servet 11211, Geneva 4, Switzerland, [7]Department of Computer Science, University of Bristol, Woodland Road, Bristol, BS8 1UB, UK, [8]J. Craig Venter Institute (JCVI), 9704 Medical Center Drive, Rockville, MD 20850, USA,[9]Computational Biology Unit, Institute of Infectious Disease and Molecular Medicine, University of Cape Town Health Sciences Campus, Anzio Road, Observatory 7925, South Africa, [10]University of Southern California, Los Angeles, CA 90089, USA, [11]Protein Information Resource (PIR), Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200 Washington, D.C. 20007, USA and[12]Structural and Molecular Biology Department, University College London, University of London, WC1E 6BT UK

Received October 4, 2011; Revised October 10, 2011; Accepted October 12, 2011

ABSTRACT

InterPro (http://www.ebi.ac.uk/interpro/) is a database that integrates diverse information about protein families, domains and functional sites, and makes it freely available to the public via Web-based interfaces and services. Central to the database are diagnostic models, known as signatures, against which protein sequences can be searched to determine their potential function. InterPro has utility in the large-scale analysis of whole genomes and meta-genomes, as well as in characterizing individual protein sequences. Herein we give an overview of new developments in the database and its associated software since 2009, including updates to

*To whom correspondence should be addressed. Tel: +44 (0) 1223 494 481; Fax: +44 (0) 1223 494 468; Email: hunter@ebi.ac.uk
Correspondence may also be addressed to Alex Mitchell. Email: mitchell@ebi.ac.uk
Correspondence may also be addressed to Philip Jones. Tel: +44 (0) 1223 492610; Fax: +44 (0) 1223 494484; Email:pjones@ebi.ac.uk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

database content, curation processes and Web and programmatic interfaces.

INTRODUCTION

The InterPro database integrates predictive models, or signatures, from multiple, diverse source repositories: Pfam (1), PRINTS (2), PROSITE (3), SMART (4), ProDom (5), PIRSF (6), SUPERFAMILY (7), PANTHER (8), CATH-Gene3D (9), TIGRFAMs (10) and HAMAP (11). Each source has its own distinct biological focus and/or methodology of signature production. The aim of InterPro is to combine their individual strengths to provide a single resource through which scientists can access comprehensive information about protein families, domains and functional sites.

Member database signatures are integrated into InterPro manually. Curators combine signatures representing the same protein family, domain or site into single database entries, and, where possible, trace biological relationships between the constituent signatures. They check the biological accuracy of the individual signatures and add pertinent information, including consistent names, descriptive abstracts (with links to original publications) and Gene Ontology (GO) (12) terms. Semi-automatic procedures create and maintain links to an array of other databases, including the protease resource MEROPS (13), the protein interaction database IntAct (14), the ENZYME database (15) and the 3D structure database PDB (16).

InterPro signature matches to the UniProt Knowledgebase (UniProtKB) (17) and the UniParc protein sequence archive are calculated using the InterProScan software package (18). This information is made available to the public in XML files as well as through Web interfaces, where users can search with either a protein sequence or a protein identifier. These data are also used to aid UniProtKB curators in their annotation of Swiss-Prot proteins and are utilized by the automatic system that adds annotation to UniProtKB/TrEMBL.

InterProScan can also be used to perform automated analysis of protein sequences. The software is available (i) as a browser-based tool for analysing single protein sequences (http://www.ebi.ac.uk/Tools/pfa/iprscan/); (ii) programmatically via Web services (19) that allow up to 25 sequences to be analysed per request (SOAP-based service documented at http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_soap and REST-based service at http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_rest); and (iii) as a downloadable package for local installation from the EBI's FTP server: (ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan).

NEW FEATURES IN INTERPRO

Updates to database content

InterPro curators continue to integrate new entries into the resource. There have been 15 major public releases (where at least one member database has been updated) and 1 minor release (where only the underlying protein sequence database has been updated) of InterPro since 2009; the latest release (version 34.0) contains 31 685 member database signatures integrated into 22 245 InterPro entries, and provides matches to ~80% of the sequences in UniProtKB (Table 1).

Changes to terminology and data structure

InterPro entries are classified according to the type of signature they group together. In order to make it clear to end users what can be inferred from a match to a particular entry in the database, the different entry types have been reviewed and terminology has been standardized. Entry types now comprise families, domains, repeats, post-translational modifications, active sites, binding sites and conserved sites, with formal definitions for each type clearly stated on the InterPro Web site.

The relationships between different entry types have also been revised. Family and domain entries continue to be organized into hierarchies, with top-level entries describing broad families or domains that share higher level structure and/or function, and entries further down the hierarchy describing more specific functional subfamilies or structural/functional subclasses of domains. However, family and domain entries are no longer permitted to occur within the same hierarchy, and are now classified into distinct hierarchies that relate domain architectures to protein families. These data structures are presented on the Web interface in an intuitive tree view, as well as being available on the FTP site in a flat-file (ftp://ftp.ebi.ac.uk/pub/databases/interpro/ParentChildTreeFile.txt).

Integration of the HAMAP database

A new member database, HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes), was added in 2009, InterPro release 22.0. The HAMAP database contains over 1600 signatures, specifically describing archaeal, bacterial and plastid-encoded protein families. It is based upon weighted-matrix signatures, of the type used in the PROSITE profiles database. The integration of HAMAP into InterPro has helped improve diagnostic coverage and specificity for prokaryotic sequences, providing matches to nearly 600 functionally- and/or taxonomically-specific protein families and subfamilies for which no other member database provides corresponding signatures.

Table 1. Coverage of the major sequence databases UniProtKB, UniParc and UniMES by InterPro signatures

Sequence database	Number of proteins in database	Number of proteins with one or more matches to InterPro (%)
UniProtKB/Swiss-Prot	532 146	507 297 (95.3 %)
UniProtKB/TrEMBL	16 886 838	13 365 742 (79.1 %)
UniProtKB (Total)	17 418 984	13 873 039 (79.6 %)
UniParc	28 628 639	20 974 897 (73.3 %)
UniMES	6 028 191	4 442 162 (73.7 %)

Mapping to GO terms

The GO (12) provides a controlled vocabulary that can be used to describe gene products in terms of their molecular functions, biological processes and the subcellular components in which they are found, in a consistent and structured fashion. InterPro entries are manually annotated with these terms, allowing GO terms to be inferred for sequences that match the entries. To date, over 10 000 InterPro entries have been annotated with one or more GO terms, with almost 25 000 GO terms in total mapped to the resource. InterPro GO mappings are currently cross-referenced over 66 million times in UniProtKB, providing GO terms for over 11 million individual proteins.

Recently, improvements have been made to InterPro GO-term-mapping procedures to help bring them into line with the GO's taxonomic restrictions (20). The revisions ensure, for example, that mammalian-specific terms are not assigned to InterPro entries that match non-mammalian proteins. The mapping procedures have also been adapted to take account of InterPro entry types so that entries representing domains will no longer be allocated GO terms based on the general function of the entire protein.

NEW CROSS-REFERENCES

New cross-references have been added, linking InterPro entries to related enzyme and pathway information in the PRIAM (21), Reactome (22), KEGG (23), MetaCyc (24) and UniPathway (25) resources. An automatic procedure checks the type of proteins matched to an InterPro entry and, if a significant proportion (>80%) are found to belong to a particular enzyme family or pathway, a link is made to the appropriate resource. By adding this information, InterPro can now be used for pathway analysis; for example, to examine whether or not a complete genome contains the protein components predicted to be sufficient for a particular reaction or pathway.

XML formats

A new XML schema has been adopted by all InterPro Consortium members to promote data exchange with each other and with third-parties.

The schema defines three data formats: signature annotation, protein matches and nucleotide sequence matches for all six reading frames. Currently the signature annotation XML format is used in the InterPro production process to import annotation from four Consortium members (PRINTS, PROSITE, Pfam and PIRSF), which has led to a reduction in import time and complexity. The intention is to roll this format out to other Consortium partners in the near future. The protein-match XML format is available from the beta version of InterProScan 5 (see below) to facilitate interoperability and integration with third-party pipelines and applications: this facility will be available for nucleotide sequences shortly.

A new user interface

With the aim of improving the InterPro user experience, a new Web-based interface has been developed. The interface has been publicly available at http://wwwdev.ebi.ac.uk/interpro as a beta release since January 2011. Several goals have been addressed in this development, including improvements in usability, the provision of additional functionality, and many improvements to the aesthetics of the interface. These goals have been driven by a user-centred design approach to improve usability and identify important functionality, coupled with a professional graphic design process. Findings gathered from user surveys, formal usability testing, user interviews and reviews of several years of support requests have allowed the InterPro team to focus interface development on real user needs.

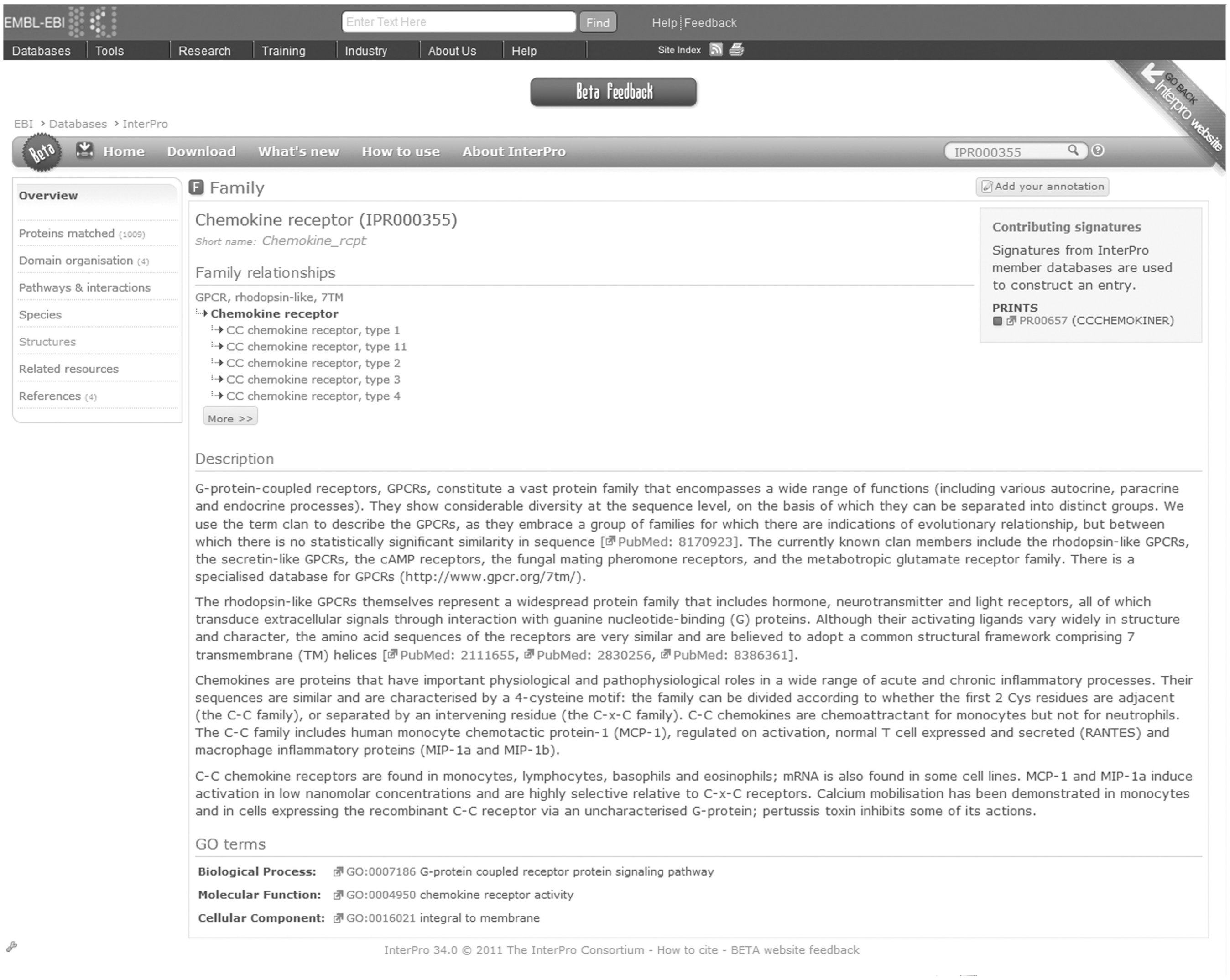
Developing an interface to a conceptually complex system such as InterPro is challenging. The complexity of the underlying data model and the integrated nature of the InterPro resources make it difficult to avoid placing a high cognitive load on the user. A major emphasis of the new design has been to develop individual pages that are as clutter-free and intuitive as possible, freeing the user to focus on the biological problem that they are attempting to address, rather than forcing them to think about how to interact with the interface.

Concrete examples of these improvements include the division of the previously complex and confusing 'Entry page' into eight separate, cross-referenced pages. Each page is clearly named, so users can easily find the content they require, without having to wade through irrelevant detail. Graphical elements have been employed to provide contextual clues, including icons representing proteins, member database signatures and InterPro entries, with the latter having different icons to represent protein families, domains, sites and repeats. This simple change has had a demonstrably positive impact, allowing users to identify the entities presented on the interface with greater ease and speed.The entry 'overview' page is illustrated in Figure 1.

Users can now search InterPro directly with a protein sequence by pasting the sequence into the text area provided on the home page. InterPro then performs a fast look-up of proteins for which matches have already been calculated. If the sequence is available in InterPro, the user is taken to the new protein page directly. If the sequence is not present in InterPro, it is submitted automatically to the InterProScan service, which returns results once the analysis is complete. Tighter integration of these two search services is currently being developed to ensure that users are presented with results in the same way by both InterPro and InterProScan. This improvement will be included in the final released version of the new InterPro Website.

InterProScan 5

Over the last 3 years, InterProScan has been completely re-written using the Java programming language. The new InterProScan is now available as a beta release (version 5beta2) for public evaluation and comment;



Chemokine receptor (IPR000355)

Short name: Chemokine_rcpt

Family relationships

GPCR, rhodopsin-like, 7TM
- **Chemokine receptor**
  - CC chemokine receptor, type 1
  - CC chemokine receptor, type 11
  - CC chemokine receptor, type 2
  - CC chemokine receptor, type 3
  - CC chemokine receptor, type 4

Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

PRINTS
PR00657 (CCCHEMOKINER)

Description

G-protein-coupled receptors, GPCRs, constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the term clan to describe the GPCRs, as they embrace a group of families for which there are indications of evolutionary relationship, but between which there is no statistically significant similarity in sequence [PubMed: 8170923]. The currently known clan members include the rhodopsin-like GPCRs, the secretin-like GPCRs, the cAMP receptors, the fungal mating pheromone receptors, and the metabotropic glutamate receptor family. There is a specialised database for GPCRs (http://www.gpcr.org/7tm/).

The rhodopsin-like GPCRs themselves represent a widespread protein family that includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework comprising 7 transmembrane (TM) helices [PubMed: 2111655, PubMed: 2830256, PubMed: 8386361].

Chemokines are proteins that have important physiological and pathophysiological roles in a wide range of acute and chronic inflammatory processes. Their sequences are similar and are characterised by a 4-cysteine motif: the family can be divided according to whether the first 2 Cys residues are adjacent (the C-C family), or separated by an intervening residue (the C-x-C family). C-C chemokines are chemoattractant for monocytes but not for neutrophils. The C-C family includes human monocyte chemotactic protein-1 (MCP-1), regulated on activation, normal T cell expressed and secreted (RANTES) and macrophage inflammatory proteins (MIP-1a and MIP-1b).

C-C chemokine receptors are found in monocytes, lymphocytes, basophils and eosinophils; mRNA is also found in some cell lines. MCP-1 and MIP-1a induce activation in low nanomolar concentrations and are highly selective relative to C-x-C receptors. Calcium mobilisation has been demonstrated in monocytes and in cells expressing the recombinant C-C receptor via an uncharacterised G-protein; pertussis toxin inhibits some of its actions.

GO terms

Biological Process: GO:0007186 G-protein coupled receptor protein signaling pathway
Molecular Function: GO:0004950 chemokine receptor activity
Cellular Component: GO:0016021 integral to membrane

Figure 1. The ‘Overview’ page on the new set of InterPro entry pages, including the family hierarchy for this entry, an extensive description of the family and cross references to three GO terms that are associated with this family. In this case, the entry comprises a single integrated PRINTS signature. Note the red ’F’ icon that indicates that this entry describes a protein family.

details of how to obtain and install it can be found at http://code.google.com/p/interproscan/wiki/RunningStandaloneInterProScan5. The new version exploits modern, stable Java technologies. A major focus of development has been to improve both the reliability and the scalability of InterProScan to allow it to support large-scale, high-throughput sequence analysis. The final version will be easy to download and install on a variety of platforms.

New functionality has been incorporated into InterProScan version 5, including a fast pre-calculated match lookup Web-service. This has the advantage that users wishing to install InterProScan locally are not obliged to download the complete set of pre-calculated matches; however, it is possible to download and install this service locally, should users wish to make confidential use of InterProScan behind a firewall. The existing cross-references to InterPro entries and GO annotations are also provided, as in the current version of InterProScan. A mechanism to allow matches to be calculated against nucleotide sequence data will be available in the final version, using the EMBOSS getorf program. This new service allows the mapping of predicted features back to coordinates on the submitted nucleic acid sequence.

InterPro BioMart

In July 2009, a BioMart was added to the InterPro suite of services. BioMart provides users with the ability to retrieve large sets of data, based on sophisticated queries that may incorporate multiple filters. Users are able to specify precisely which fields are included in the results returned. The InterPro BioMart has been described previously (26), including a detailed explanation of how to use the BioMart with several example queries.

The most important benefit provided by this feature is the ability to interrogate InterPro for multiple entries, proteins or member database signatures in a single query, which is a feature not available from the main InterPro Web interface. In addition, BioMart provides an easy to use REST Web service for programmatic access to InterPro data. The InterPro BioMart is linked from the InterPro home-page, and is also available directly from the BioMart Central Portal at http://www.biomart.org. The BioMart is exploited extensively throughout the main InterPro Web pages to allow users to download results in 'tab-separated values' (TSV) format. The BioMart user interface is illustrated in Figure 2.

InterPro DAS service

The Distributed Annotation System, DAS (27) is used extensively throughout bioinformatics to allow sharing of annotation on both nucleotide and protein sequences and protein structure. InterPro data were previously available as a single DAS data-source provided and maintained by the Ensembl team at the Wellcome Trust Sanger Institute.

In March 2010 InterPro DAS-service provision moved to the EBI, at the same time being extended to provide three DAS data-sources as described in Table 2.

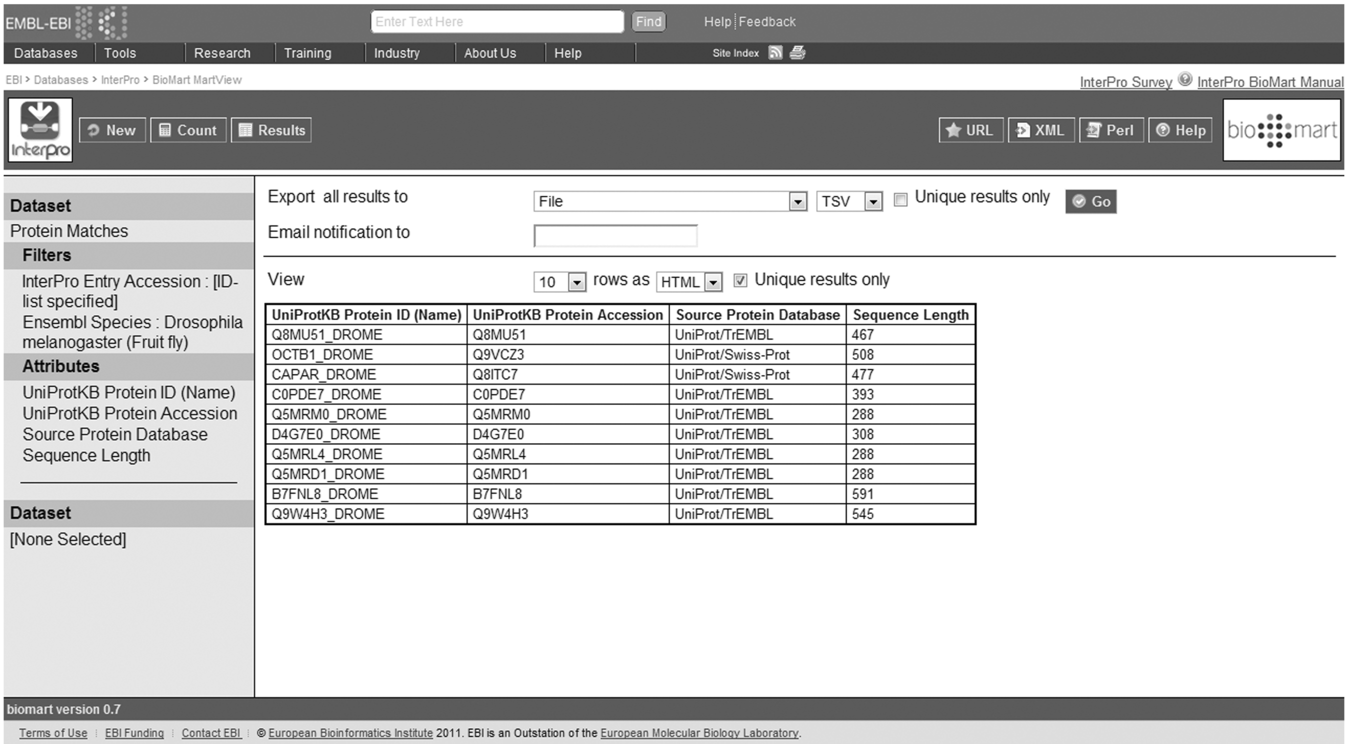
In November 2010 the InterPro DAS service was upgraded to comply with the new DAS 1.6 specification (http://www.biodas.org/documents/spec-1.6.html), implemented using the MyDas Java DAS Server API (http://code.google.com/p/mydas/). All three data sources are registered with the DAS Registry http://www.dasregistry.org/ with IDs as indicated above.

AVAILABILITY

The database and related software are freely available for download and distribution, provided the appropriate Copyright notice is supplied (as described in the accompanying Release Notes). Data can be downloaded in a flat-file format (XML) and via the Web interface and Web services described in the text.

DISCUSSION

InterPro continues to be an important protein structural and functional classification tool that is used directly by high-profile, large-scale sequence databases and genomics projects, and for the characterization of individual protein sequences via the Web. In 2011, the EBI-hosted version of InterProScan averaged more than two million sequence searches per month, which represents a 4-fold increase in monthly searches since 2009. Given its high (and growing) usage statistics, and the clear value of the resource to the scientific community, it is important that InterPro continues to expand and adapt to meet users' changing requirements. InterPro's sequence coverage has kept pace with the rapid growth of UniProt (which has grown from over 6 million to more than 17 million sequences during the last 3 years) thanks to the on-going development of new signatures by its partner databases and continued integration efforts of its curators. Improvements to the methods used in InterPro entry curation (e.g. standardized definitions to help streamline signature integrations) and data processing (e.g. the adoption of the new



Dataset
Protein Matches
Filters
InterPro Entry Accession : [ID-list specified]
Ensembl Species : Drosophila melanogaster (Fruit fly)
Attributes
UniProtKB Protein ID (Name)
UniProtKB Protein Accession
Source Protein Database
Sequence Length

Dataset
[None Selected]

Export all results to File TSV ☐ Unique results only
Email notification to

View 10 rows as HTML ☑ Unique results only

UniProtKB Protein ID (Name)	UniProtKB Protein Accession	Source Protein Database	Sequence Length
Q8MU51_DROME	Q8MU51	UniProt/TrEMBL	467
OCTB1_DROME	Q9VCZ3	UniProt/Swiss-Prot	508
CAPAR_DROME	Q8ITC7	UniProt/Swiss-Prot	477
C0PDE7_DROME	C0PDE7	UniProt/TrEMBL	393
Q5MRM0_DROME	Q5MRM0	UniProt/TrEMBL	288
D4G7E0_DROME	D4G7E0	UniProt/TrEMBL	308
Q5MRL4_DROME	Q5MRL4	UniProt/TrEMBL	288
Q5MRD1_DROME	Q5MRD1	UniProt/TrEMBL	288
B7FNL8_DROME	B7FNL8	UniProt/TrEMBL	591
Q9W4H3_DROME	Q9W4H3	UniProt/TrEMBL	545

Figure 2. The InterPro BioMart. This example illustrates the use of the BioMart to return a large set of data. In this case, a query has been built to return all proteins that are predicted to be members of the rhodopsin-like GPCRs (IPR000276) in Drosophila melanogaster.

Table 2. InterPro DAS data sources

DAS registry ID	Data source name	URL	Provision
DS_327	InterPro	http://www.ebi.ac.uk/das-srv/interpro/das/InterPro	details of InterPro signature matches coordinated on UniProtKB protein sequences
DS_1028	InterPro-matches-overview	http://www.ebi.ac.uk/das-srv/interpro/das/InterPro-matches-overview	summary matches of InterPro entries coordinated on UniProtKB protein sequences and is a default data source on the Dasty3 DAS client [http://www.ebi.ac.uk/dasty, (28)]
DS_1029	InterPro-UniParc-matches	http://www.ebi.ac.uk/das-srv/interpro/das/InterPro-UniParc-matches	details of InterPro member database signature matches coordinated on UniParc (UniProt Archive) protein sequences.

XML format for data exchange that has sped-up the process of loading and checking data) have also helped.

The InterPro software development team have focused on improving the usability of InterPro for both direct human interaction and for programmatic access. An improved primary Web interface and the addition of the BioMart user interface have improved users' experience of InterPro.

The provision of programmatic access to InterPro has been extended through the development of new Web services, including the extended DAS 1.6 services and the InterPro BioMart REST Web service. The development of InterProScan 5 will provide benefits to users installing the service locally, including improved ease of installation.

Future plans for InterPro include improving the text-based searching of the database and development of an experimental semantic representation of InterPro's data to support the drive to develop sophisticated semantic queries across bioinformatics resources.

FUNDING

European Union under the program 'FP7 capacities: Scientific Data Repositories' (grant number 213037). The title for the project is IMproving Protein Annotation and Co-ordination using Technology (IMPACT); BBSRC Bioinformatics and Biological Resources Fund (grant BB/F010508/1); and improvements to InterProScan have been supported in part by a grant from the United States National Institutes of Health (grant GM081084). Funding for open access charge: EMBL-EBI.

Conflict of interest statement. None declared.

REFERENCES

1. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D22.
2. 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.
3. Sigrist,C.J.A., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.
4. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
5. Bru,C., Courcelle,E., Carrère,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.
6. Nikolskaya,A.N., Arighi,C.N., Huang,H., Barker,W.C. and Wu,C.H. (2006) PIRSF family classification system for protein functional and evolutionary analysis. *Evolutionary Bioinformatics Online*, **2**, 197–209.
7. de Lima Morais,D.A., Fang,H., Rackham,O.J.L., Wilson,D., Pethica,R., Chothia,C. and Gough,J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.
8. Mi,H., Dong,Q., Muruganujan,A., Gaudet,P., Lewis,S. and Thomas,P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, D204–D210.
9. Lees,J., Yeats,C., Redfern,O., Clegg,A. and Orengo,C. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**, D296–D300.
10. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
11. Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Bulliard,V., de Castro,E., Lachaize,C., Baratin,D. *et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, **37**, D471–D478.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
13. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.
14. 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.
15. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

16. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
17. The UniProt Consortium. (2010) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
18. 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.
19. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.
20. Deegan née Clark,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.
21. Claudel-Renard,C., Chevalet,C., Faraut,T. and Kahn,D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.
22. D'Eustachio,P. (2011) Reactome knowledgebase of human biological pathways and processes. *Methods Mol. Biol.*, **694**, 49–61.
23. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
24. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
25. Morgat,A., Coissac,E., Coudert,E., Axelsen,K., Keller,G., Bairoch,A., Bridge,A., Bougueleret,L., Xenarios,I. and Viari,A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
26. Jones,P., Binns,D., McMenamin,C., McAnulla,C. and Hunter,S. (2011) The InterPro BioMart: federated query and web service access to the InterPro Resource. *Database*, **2011**, doi:10.1093/database/bar033.
27. Jenkinson,A.M., Albrecht,M., Birney,E., Blankenburg,H., Down,T., Finn,R.D., Hermjakob,H., Hubbard,T.J.P., Jimenez,R.C., Jones,P. *et al.* (2008) Integrating biological data–the Distributed Annotation System. *BMC bioinformatics*, **9(Suppl)8, S3.**
28. Villaveces,J.M., Jimenez,R.C., Garcia,L.J., Salazar,G.A., Gel,B., Mulder,N., Martin,M., Garcia,A. and Hermjakob,H. (2011) Dasty3, a WEB Framework for DAS. *Bioinformatics*, **27**, 2616–2617.