The InterPro Database, 2003 brings increased coverage and new features

MULDER, Nicola J, et al.

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).

Reference

DOI : 10.1093/nar/gkg046
PMID : 12520011

Available at:
http://archive-ouverte.unige.ch/unige:30568

Disclaimer: layout of this document may differ from the published version.
The InterPro Database, 2003 brings increased coverage and new features

Nicola J. Mulder1,*, Rolf Apweiler1, Teresa K. Attwood3, Amos Bairoch4, Daniel Barrell1, Alex Bateman2, David Binns1, Margaret Biswas5, Paul Bradley1,3, Peer Bork6, Phillip Bucher7, Richard R. Copley8, Emmanuel Courcelle9, Ujjwal Das1, Richard Durbin2, Laurent Falquet7, Wolfgang Fleischmann1, Sam Griffiths-Jones2, Daniel Haft10, Nicola Harte1, Nicolas Hulo4, Daniel Kahn9, Alexander Kanapin1, Maria Krestyaninova1, Rodrigo Lopez1, Ivica Letunic6, David Lonsdale1, Ville Silventoinen1, Sandra E. Orchard1, Marco Pagni7, David Peyruc9, Chris P. Ponting11, Jeremy D. Selengut10, Florence Servant1, Christian J. A. Sigrist4, Robert Vaughan1 and Evgueni M. Zdobnov6,12

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

<table>
<thead>
<tr>
<th>Release</th>
<th>Date</th>
<th>Entries</th>
<th>SPTR coverage (%)</th>
<th>New feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>March 2000</td>
<td>2990</td>
<td>60.2</td>
<td>PROSITE, PRINTS, Pfam</td>
</tr>
<tr>
<td>2.0</td>
<td>October 2000</td>
<td>3204</td>
<td>63.9</td>
<td>ProDom included</td>
</tr>
<tr>
<td>3.0</td>
<td>March 2001</td>
<td>3875</td>
<td>73.3</td>
<td>SMART included</td>
</tr>
<tr>
<td>4.0</td>
<td>November 2001</td>
<td>4691</td>
<td>72.2</td>
<td>TIGRFAMs included</td>
</tr>
<tr>
<td>5.0</td>
<td>May 2002</td>
<td>5312</td>
<td>74.0</td>
<td>General updates*</td>
</tr>
<tr>
<td>Currently</td>
<td>September 2002</td>
<td>5629</td>
<td>74.0</td>
<td>General updates*</td>
</tr>
</tbody>
</table>

*Updates to include new member database releases.

To view the complete output click [here](#).
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](image_url)

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