InterPro: Quick tour

Alex Mitchell [1]

- Proteins
- Beginner
- 0.5 hour

This quick tour provides a brief introduction to InterPro, the EBI’s database of protein families, domains and functional sites. For a more detailed walkthrough of InterPro have a look at our InterPro tutorial [2].

Learning objectives:

- A basic understanding of InterPro and how it can help you to relate protein architecture to function
- Knowing where to find out more about InterPro

What is InterPro?

Why do we need InterPro?

Classifying proteins into families and identifying important domains and sites is invaluable for helping biologists to identify distantly related proteins and to predict their functions. A daunting array of resources, each with different strengths and weaknesses, is now available for searching genomes and proteomes with ‘protein signatures’ – diagnostic entities that are used to recognise a particular domain [3] or family. InterPro [4] combines protein signatures from multiple, diverse databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool for protein sequence classification.

Applications

InterPro is used in the large-scale analysis of whole proteomes, genomes and metagenomes, as well as for the characterisation of individual proteins. Within the EBI, we use InterPro to help us annotate protein sequences in UniProtKB [5].

InterPro data

InterPro has 13 member databases, each of which uses a different method to classify proteins. InterPro curators manually integrate protein signatures from member databases, merging signatures that represent the same protein family [6], domain or site into single InterPro entries. Where possible, they also trace biological relationships between entries. They check the biological accuracy of the individual signatures and add pertinent information, including consistent entry names, descriptive abstracts, links to the biomedical literature and Gene Ontology [7] terms. Links are also made to various other databases, such as ENZYME [8] and PDB [9]. Figure 1 provides an overview of the data sources used to construct InterPro.
Figure 1 An overview of the data sources used to construct InterPro

**Member databases**

The following databases contribute data to InterPro:

- **CDD** [10] at NCBI, Bethesda, USA
- **PANTHER** [12] at University of Southern California, CA, USA
- **PIRSE** [13] at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, USA
- **Pfam** [14] at the EMBL-EBI, Hinxton, UK
- **PRINTS** [15] at the University of Manchester, UK
- **PROSITE** [16] and **HAMAP** [17] at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland
- **SFLD** [18] at the University of California, San Francisco, USA
- **SMART** [19] at EMBL, Heidelberg, Germany
- **SUPERFAMILY** [20] at the University of Bristol, UK
- **TIGRFAMs** [21] at the J. Craig Venter Institute, Rockville, MD, US

**Searching and visualising data from InterPro**

**Searching InterPro**

You can search InterPro using:
- sequence search, which uses the InterProScan [22] analysis tool to predict the family and domain [3] information for an amino acid [23] sequence;
- text search, using InterPro entry names and identifiers, UniProt [24] accessions, GO [7] terms, PDB [25] identifiers, or free text, to find information in InterPro relating to your query;
- search for a domain architecture;
- browse through InterPro member databases and/or InterPro entry types

Figure 2 highlights the location of the different search entry points on the InterPro web interface.

Visualising InterPro data

InterPro matches to proteins are displayed graphically (illustrated in Figure 3). The protein family [6] membership section (see figure) provides information about the family to which the protein is predicted to belong.

All signatures - and the InterPro entries where they are integrated - are displayed in the protein sequence viewer. The signatures and InterPro entries are grouped by type (family, domain, homologous superfamily, repeat, site). The coloured bars indicate the location of signature matches on the protein sequences. Each matching InterPro entry is displayed on a separate line. They can be coloured by accession [26], member database or domain relationship.

On top of the sequence viewer, different icons allow to collapse the visualisation, disable tooltips information, show InterPro entries only, display the viewer on full screen, take a screenshot or display signatures - and the InterPro entries where they are integrated - full names.

The residues conservation scores can be calculated on demand for sections of the protein matching a Pfam [27] signature.

You can access further information about an InterPro entry or a member database signature by clicking on the accession on the right side, which will take you to the corresponding InterPro page.
Getting data from InterPro

In addition to the web-based sequence and text searches, you can use the InterPro [28] API [29] or download both the InterPro data and the InterProScan [30] analysis tool via anonymous FTP [31] and run InterProScan locally. We would recommend this if you have a large number of searches to perform, or are concerned about confidentiality. Additional files that are available on the FTP site include a list of all InterPro entries of each type, a file of InterPro mappings to GO terms, and a list of InterPro hits in UniProtKB [32]. The database and matches are available in XML [33] format with a corresponding document type definition (DTD) file.

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on InterPro

Documentation

You can find documentation about InterPro [28] (the user manual, release notes, etc) linked from our web page [4]
or you can download the information from our FTP server [34].

Contact
For all support enquiries, please contact the InterPro interhelp [at] ebi.ac.uk (help desk).

References


Contributors

Alex Mitchell [1]
EMBL-EBI
Curation co-ordinator, InterPro and EBI Metagenomics

Alex Mitchell is co-ordinator for the InterPro and EBI Metagenomics databases at EMBL-EBI. He obtained his DPhil in pharmacology from the University of Oxford, and was previously employed as a molecular biologist at the Institute of Psychiatry. He moved to the University of Manchester to work on protein sequence analysis and functional classification, before joining EMBL-EBI in 2011.
Amaia Sangrador [41]

EMBL-EBI
Scientific curator, InterPro

Amaia Sangrador is a curator for InterPro at the European Bioinformatics Institute in Cambridge, UK. She joined EMBL-EBI in 2010. She has a PhD in Molecular Biology and has been involved in several research projects in the areas of immunology and comparative genomics.

Typhaine Paysan-Lafosse [42]

EMBL-EBI
Bioinformatician

Typhaine Paysan-Lafosse is working as a bioinformatician and scientific curator [43] for the InterPro [28] database, previously she worked on the Genome3D project for InterPro and PDBe [44] databases. She joined EMBL-EBI in 2016. Beforehand she worked for the international ImMunoGeneTics information system in Montpellier.

Source URL: https://www.ebi.ac.uk/training/online/course/interpro-quick-tour

Links
[1] https://www.ebi.ac.uk/training/online/trainers/mitchell
[3] https://www.ebi.ac.uk/training/online/glossary/domain
[6] https://www.ebi.ac.uk/training/online/glossary/protein-family