

## InterPro



**Sarah Hunter**

*MSc University of Manchester, 1998. Pharmaceutical and Biotech Industry (Sweden), 1999-2004. At EMBL-EBI since 2005.*

### DESCRIPTION OF SERVICES

Our team coordinates the InterPro and Metagenomics projects at EMBL-EBI.

InterPro is used to classify proteins into families and predict the presence of domains and functionally important sites. The project integrates signatures from 11 major protein signature databases (Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH-Gene3D, PANTHER and HAMAP) into a single resource. During the integration process, InterPro rationalises instances where more than one protein signature describes the same protein family or domain, uniting these into single InterPro entries and noting relationships between them where applicable. Additional biological annotation is included, together with links to external databases such as GO, PDB, SCOP and CATH. InterPro precomputes all matches of its signatures to UniProt Archive (UniParc) proteins using the InterProScan software, and displays the matches to the UniProt KnowledgeBase (UniProtKB) in various formats, including XML files and web-based graphical interfaces. InterPro has a number of important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects. InterPro is used by Ensembl and in the GOA project to provide large-scale mapping of proteins to GO terms.

Metagenomics is the study of the sum genetic material found in an environmental sample or host species, typically using next-generation sequencing (NGS) technology. The Metagenomics Portal, a resource established at EMBL-EBI in 2011, enables metagenomics researchers to submit sequence data and associated descriptive metadata to the public nucleotide archives. Deposited data is subsequently functionally analysed using an InterPro-based pipeline and the results generated are visualised via a web interface.

### SUMMARY OF PROGRESS

- Issued five major releases of the InterPro database: created 2122 new entries and integrated 1176 signatures;
- Issued two new releases of InterProScan, a Perl-based signature-scanning software;
- Migrated InterPro to Hmmer3 for TIGRFAMs and Gene3D. Upgraded Gene3D results processing to use DomainFinder3 algorithm;
- Redesigned the InterPro website and released in beta;
- Launched the Metagenomics Portal in beta;
- Adapted Metagenomics data-submission tools and created pipelines for data analysis;
- Retired the CluSTr resource, which was previously maintained by the InterPro team, in June 2011.

### MAJOR ACHIEVEMENTS

During 2011 development of the InterPro beta website continued, with particular focus on the representation of matches of InterPro's signatures to protein sequences. Previously, the graphical output from InterProScan differed markedly from the InterPro website's view of the same data. The new display now looks identical, regardless of the source, giving users a more consistent experience. Other features on the site include a cleaner aesthetic and design; a new home page; better delineated entry pages that are split into sub-sections; and updated documentation to reflect these changes.

InterPro curators continued to add content to the database, and integrated over 1000 signatures during 2011. The most recent release (v35.0) has seen an improvement in coverage:

95.4% of UniProtKB/Swiss-Prot proteins match at least one InterPro entry, 79.2% of UniProtKB/TrEMBL and 79.7% overall for UniProtKB (Swiss-Prot and TrEMBL). A major focus of curation work during 2011 was improving the process for Gene Ontology term mapping to InterPro. Links to pathway databases such as KEGG and Reactome were added automatically where appropriate.

Two additional InterPro member databases, TIGRFAMs and Gene3D, were updated to use HMMER3, following Pfam's adoption of the algorithm in 2010. As with Pfam, it was necessary to check the validity of existing integrations as HMMER3 can more sensitively detect remote homologs compared to its predecessor. The algorithm was updated in InterProScan for these two resources.

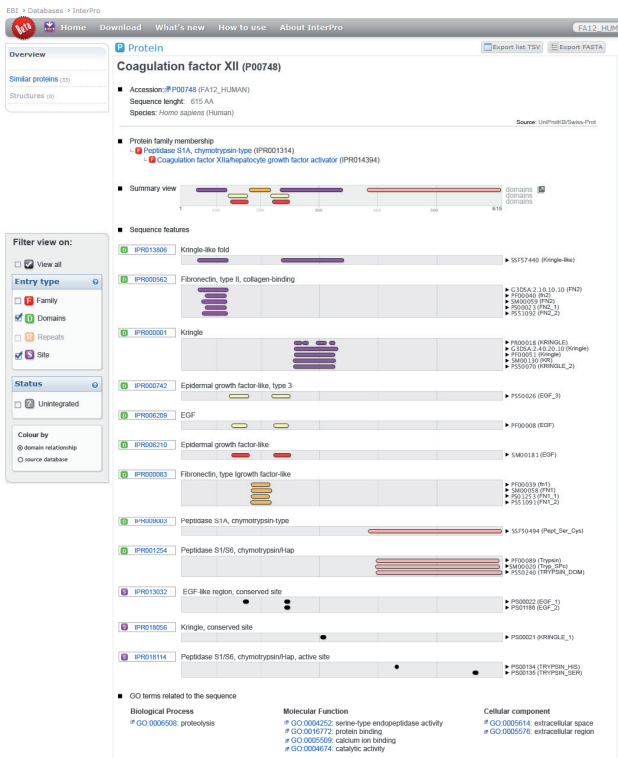


Figure 1. View of the domains and sites in a protein (FA12\_HUMAN) on the new InterPro beta site.

InterProScan5 was developed to utilise 'best-in-breed' Java technologies to improve the robustness of the software compared to the previous version. InterProScan5 outputs a descriptive XML format, which can then be transformed into a graphical, tab-delimited or GFF3 representation of results. If a user searches using a nucleotide sequence, InterProScan will translate the sequence into six open reading frames. The software then maps protein domain and family results back to the original DNA sequence once they have been calculated. Users may also request a look-up of the Gene Ontology terms and pathways associated with the InterPro entries matching their sequence(s). InterProScan5 contains an additional algorithm called Phobius, which is used to predict the presence of signal peptides and transmembrane regions in proteins.

**Selected publications**

Hunter, C., *et al.* (2011) The EBI metagenomics archive, integration and analysis resource. In: Frans J. de Bruijn, Ed, *Handbook of Molecular Microbial Ecology I: Metagenomics and complementary approaches*. Wiley-Blackwell.

Hunter, S., Jones, P., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40 (Database issue), D306-12.

Jones, P., Binns, D., *et al.* (2011) The InterPro BioMart: federated query and web service access to the InterPro Resource. *Database (Oxford)* 2011, bar033.

McDowall, J. and Hunter, S. (2011) InterPro Protein Classification. *Methods Mol. Biol.* 694, 37-47.

The Metagenomics Portal was launched in beta in late 2011, offering a number of public datasets for browsing. The initial version was designed to assist researchers in submitting, organising and analysing their metagenomic datasets. We built an analysis pipeline comprising quality control, clustering and filtering steps, followed by an InterPro-based functional characterisation step (which includes the association of GO terms). We also created an interface that enables users to submit their raw sequence and sample metadata to the European Nucleotide Archive (ENA) and retrieve any subsequent analysis results. Data may be held privately prior to publication, although policy dictates that all data must eventually be made available in the public domain.

**FUTURE PLANS**

The InterPro website (currently in beta) will be launched to the public in early 2012; InterProScan5 will be released concurrently. InterPro is expected to be served from EMBL-EBI's London data centres by the end of 2012.

Further developments are planned for the Metagenomics Portal. Existing data pipelines will be enhanced by the addition of 16S marker gene analysis software so that the taxonomic diversity of metagenomics samples can be estimated. Submission tools will similarly be improved to increase the ease with which data can be uploaded for analysis and archiving. New visualisation tools will be created that allow researchers to compare the functional and taxonomic profiles of different metagenomics samples in an intuitive way. We intend to make these improvements by working closely with target users and by extensive usability testing.

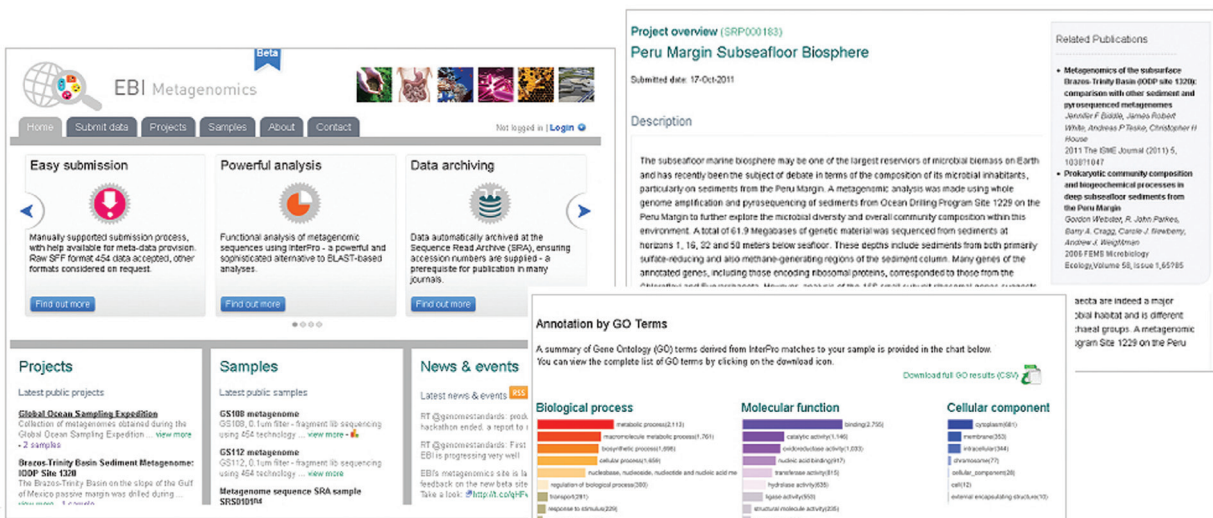


Figure 2. The Metagenomics Portal home page, an example of a page describing a metagenomics project and a summary of the annotation of a sample's sequences using the Gene Ontology.