Applications of InterPro in protein annotation and genome analysis

Margaret Biswas, John F. O'Rourke, Evelyn Camon, Gill Fraser, Alexander Kanapin, Youla Karavidopoulou, Paul Kersey, Evgenia Kriventseva, Virginie Mittard, Nicola Mulder, Isabelle Phan, Florence Servant and Rolf Apweiler

Date received (in revised form): 10th June 2002

Margaret Biswas
has a background in biochemistry and molecular structure and has worked in bioinformatics for over 12 years.

John O'Rourke
has an honours degree in Life Sciences Biochemistry (Trinity College Dublin), studied for a PhD in Biochemistry/ Enzymology (Aberdeen University) and carried out post-doctoral research in mammalian gene expression and transcriptional regulation at Oxford University, before taking up a position as Database programmer at the EBI.

Evelyn Camon
is reading for a PhD in Immunology and has three years experience as a curator of the EMBL-Bank database. Currently she is coordinating the annotation of Gene Ontology terms to SWISS-PROT and TrEMBL databases in a project called 'GOA' (Gene Ontology Annotation @ EBI).

Gill Fraser
is a curator for the SWISS-PROT and TrEMBL databases at the EMBL outstation – European Bioinformatics Institute. She is involved in the automated annotation of TrEMBL using RuleBase.

Keywords: *automatic annotation, genome comparison, protein classification, functional annotation, proteome comparison*

Rolf Apweiler,
EMBL Outstation – European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge, UK

Tel: +44 (0)1223 494 435
Fax: +44 (0) 1223 494 468
E-mail: apweiler@ebi.ac.uk

Abstract

The applications of InterPro span a range of biologically important areas that includes automatic annotation of protein sequences and genome analysis. In automatic annotation of protein sequences InterPro has been utilised to provide reliable characterisation of sequences, identifying them as candidates for functional annotation. Rules based on the InterPro characterisation are stored and operated through a database called RuleBase. RuleBase is used as the main tool in the sequence database group at the EBI to apply automatic annotation to unknown sequences. The annotated sequences are stored and distributed in the TrEMBL protein sequence database. InterPro also provides a means to carry out statistical and comparative analyses of whole genomes. In the Proteome Analysis Database, InterPro analyses have been combined with other analyses based on CluSTr, the Gene Ontology (GO) and structural information on the proteins.

INTRODUCTION

As increasing numbers of predicted protein sequences from genome sequencing projects enter the protein sequence databases, one of the challenges is to provide meaningful annotation for these sequences. Protein signature databases provide vital tools for identifying distant relationships in novel sequences and hence are used for the classification of protein sequences and for inference of function. InterPro[1] combines many of the commonly used protein signature databases as an integrated resource. The power of InterPro has been demonstrated at various stages and at different levels, to contribute to the demanding task of automatic and manual annotation of proteins from genome sequencing projects.

With the growing number of organisms for which complete genome sequences are available, tools that facilitate statistical and comparative analyses using functional, structural and other information are of growing importance. InterPro has been demonstrated to provide an extremely useful basis for proteome analysis complementing functional and structural information in the CluSTr database[2,3] and in the Gene Ontology (GO).[4,5]

AUTOMATIC ANNOTATION OF PROTEIN SEQUENCES

The automation of functional annotation is of paramount importance to mine the avalanche of sequence data. To enhance the annotation of uncharacterised protein sequences in TrEMBL, the sequence database group at the EBI developed a novel method for the automatic annotation of protein sequences.[6] This method selects proteins in the SWISS-PROT protein sequence database that belong to the same group of proteins as a given unannotated protein, extracts the annotation shared by all functionally characterised proteins of this group, and assigns this common annotation to the unannotated protein.

Alexander Kanapin
has an MS and PhD in biophysics (1986 and 1997 correspondingly), has been at the EBI since 1998, and his main responsibilities are InterPro production and external services.

Youla Karavidopoulou
works as a curator for the SWISS-PROT and TrEMBL databases.

Paul Kersey
has a PhD in yeast genetics, and currently works as a database programmer for the European Bioinformatics Institute.

Evgenia Kriventseva
is an EMBL PhD student in the Sequence Database Group headed by Dr Rolf Apweiler.

Virginie Mittard
has a PhD and postDoc in Structural biology by NMR, and currently works as a database curator for the European Bioinformatics Institute.

Nicola Mulder
co-ordinates the InterPro database production and maintenance activities.

Dr Isabelle Phan
currently works as a database programmer for SWISS-PROT in Geneva. She participates in the development of a schema xml for SWISS-PROT.

Florence Servant
is a database programmer in the SWISS-PROT group at the EBI, whose research interests include automated clustering methods of proteins, multiple alignment improvements and graphical user interface design.

Rolf Apweiler
co-heads the Sequence Database Group, which encompasses the EMBL, SWISS-PROT, TrEMBL and InterPro database activities at the EMBL Outstation – European Bioinformatics Institute.

Built from SWISS-PROT and InterPro, RuleBase is used to annotate TrEMBL

To implement this methodology for the automated large-scale functional annotation of proteins three major components are required:

- A reference database that serves as the source of annotation. SWISS-PROT is used because of its highly reliable, well-annotated and standardised information.

- The highly diagnostic protein family signature database InterPro supplies the means to assign proteins to groups. InterPro allows the reliable classification of proteins into families and the recognition of the domain structure of multi-domain proteins. Currently, InterPro classifies around 70 per cent of all known protein sequences and this information is incorporated into SWISS-PROT and TrEMBL in the form of database cross-references to InterPro and its member databases.

- A database (RuleBase) that stores and manages the annotation rules, their sources and their usage. Currently, RuleBase supports around 500 rules that are frequently applied on proteins in the TrEMBL database.

RuleBase

The actual flow of information during the automatic annotation can be divided into five steps:

- Use InterPro to extract the information necessary to assign proteins to groups ('conditions') and store the conditions in the RuleBase.

- Group the proteins in SWISS-PROT by the conditions.

- Extract from SWISS-PROT the common annotation shared by all functionally characterised proteins of each group and store this common annotation together with its conditions in RuleBase. Every rule consists of conditions and the annotation common to all proteins of this group characterised by these conditions.

- Group the unannotated TrEMBL entries by the conditions stored in RuleBase.

- Add the common annotation to the unannotated TrEMBL entries. The predicted annotation will be flagged with evidence tags, which will allow users to recognise the predicted nature of the annotation as well as the original source of the inferred annotation.

As the reliability of the conditions is crucial to the reliability of the methodology, a multiple-step procedure is used to minimise false positive automatic annotation:

- The InterPro database used to extract conditions to assign proteins to groups integrates different computational techniques for the recognition of signatures diagnostic for different protein families or domains. All the different approaches integrated in InterPro (hidden Markov models (HMMs), profiles, fingerprints, regular expressions, etc.) have different strengths and weaknesses. The combination of the strengths of the different signature recognition methods, coupled with statistical and biological significance test, allows overcoming the various drawbacks of the individual methods.

- An important condition in every rule is that the taxonomic classification of the unannotated protein sequences must be within the known taxonomic range of the experimentally characterised proteins. For instance, a match of an *a priori* prokaryotic signature against a human protein is regarded as violating the conditions of the rule for this protein family and thus considered as false positive and filtered out.

- In cases where a protein family is

characterised by more than one signature in InterPro, all signatures must be found in the unannotated protein sequence. For instance, bacterial rhodopsins have a signature for a conserved region in helix C (PS00950) and another signature for the retinal binding lysine (PS00327) (Figure 1). If an unannotated protein sequence matches only the helix-C-pattern, but not the retinal-binding pattern, it will not be regarded as a bacterial rhodopsin.

A 'common annotation' approach to overcome the major limitations of automatic annotation

Several methods for automatic functional characterisation of unknown protein sequences use high-level sequence similarity searches against known proteins or collect the results of different programs. The drawbacks associated with these solutions include: (i) lack of detailed, standardised annotation coverage of sequence properties; (ii) assignment of a single function to multifunctional proteins; (iii) propagation of incorrect annotation due to top similarity search hits to unknown or poorly annotated proteins; (iv) lack of coverage of position-specific annotation such as active sites or modified residues; and (v) lack of a means to refresh outdated or incorrect annotation. In the automated 'common annotation' approach described here some limitations of the existing automatic annotation methods have been overcome:

- By using the annotation from SWISS-PROT as the reliable reference database for our predictions, the propagation of wrong annotation, one of the big problems in functional annotation, has been reduced drastically.
- By using the 'common annotation' of multiple entries, the implemented methodology produces significantly fewer over-predictions than methods based on the best hit of a sequence similarity search.
- Using the 'common annotation' from SWISS-PROT with its standardised annotation and nomenclature allows the standardised annotation of uncharacterised proteins by avoiding the use of wrong nomenclature and of different descriptions for the same biological fact.
- Since the method takes both position-independent and position-specific common annotation available in the reference database into account, it is possible to achieve a much higher level of annotation, including position-specific annotation such as active sites.
- The 'common annotation' approach can be used not only with protein families, but also with conditions aiming at a higher level in the protein family hierarchy. Only the annotation common to all members of this, for instance, superfamily will be copied over. The automatic annotation on a superfamily level will obviously lead to more generic and limited annotation than on family level.
- This methodology is independent of the multi-domain organisation of

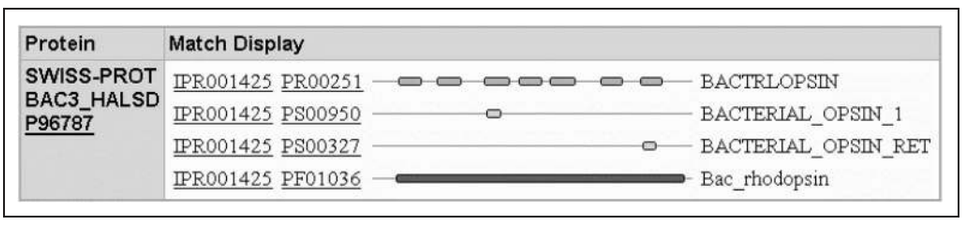


Figure 1: InterPro graphical view of bacterial rhodopsin from *Halorubrum sodomense* showing matches to both the signature for a conserved region in helix C (PS00950) and the signature for the retinal binding lysine (PS00327)

proteins. If a certain condition aims at a single domain that occurs with various other domains, it can be expected that only the annotation referring to this single domain will be found in all relevant characterised proteins. On the other hand, if the single domain always occurs with another domain, the information for the other domain will be picked up as well.

- Evidence tags allow the automatic update of the predicted annotation if the underlying conditions or the 'common annotation' in the RuleBase changes.

RuleBase update mechanism

Proteome Analysis Database provides non-redundant proteome sets

An update and maintenance mechanism is used to ensure that RuleBase adjusts to changes both in InterPro and SWISS-PROT. Changes to the signatures in an InterPro entry will affect the validity of rule conditions. These effects are minimised by using primarily conditions for the individual signatures rather than conditions for the associated InterPro entry. Synchronisation with SWISS-PROT is achieved by tracking of annotation that has been reformatted or withdrawn.

RuleBase currently is responsible for adding automatic annotation to nearly 30 per cent of the entries in TrEMBL that would otherwise remain without annotation (Figure 2). As RuleBase grows the number of automatically annotated entries will grow.

PROTEOME ANALYSIS

The aim of the proteome analysis project[7] is to provide proteome sets for whole genomes with comprehensive statistical and comparative analyses, compiled using InterPro,[1] CluSTr[3] and GO Slim,[8] and also containing structural information derived from the HSSP[9] and PDB[10] databases. There is an accompanying program designed to perform interactive InterPro-based proteome comparisons for any combination of proteomes in the database.[11]

The database provides a variety of ways to query and compare proteome composition, and the combination of both structural information and functional classification provides depth. For example, systematically conserved proteins that are likely to have orthologues across species and be involved in a common core biological process can be identified. Conserved families that are missing in a given genome or proteins unique to a particular species that may well define the species can be investigated.

Producing the proteome sets

Multiple submissions of the same protein sequence produce a problem for protein sequence databases such as SWISS-PROT and TrEMBL, and the problem is exacerbated by the submission of complete genomes. A small number of well-annotated proteins are supplemented by a large and partially overlapping set of newly predicted coding sequences. Redundant data can bias similarity searches and provide misleading data in whole genome analysis.

Manually merging entries for proteins predicted by genome submission with pre-existing entries is a priority task for the curators of SWISS-PROT. Additionally, automatic methods (based on sequence similarity and identifier tracking) are employed to produce wholly non-redundant sets (proteome sets) of SWISS-PROT and TrEMBL for use in comparative analysis.[12] For mouse and human, additional sets incorporating the latest coding sequence predictions from Ensembl[13,14] are also available. These proteome sets provide the underlying data used in all subsequent analyses.

InterPro in Proteome Analysis

InterPro covers between 31 and 67 per cent of the proteins from each of the complete genomes of which there are currently 73 in the database, the most recent addition being the genome of *Schizosaccharomyces pombe*. The Proteome Analysis pages make available an InterPro-

Rule from RuleBase:

```
#RULE   RU000418
#DATE   2001-10-26
#USER   OPS$JOROURKE
#PACK   InterPro
?IPRO   IPR002423
?IPRO-  IPR002194
?PSAC   PS00296
?EMOT   PS00296
?PSTX   A?EP?
?_OR_
?IPRO   IPR002423
?IPRO-  IPR002194
?PRAC   PR00298
?PSTX   A?EP?
!CCSI   BELONGS TO THE CHAPERONIN (HSP60) FAMILY
!SPKW   ATP-binding
!SPKW   Chaperone
```

Figure 2 (a): Rule RU000418 formulated using IPR002423 and IPR002194. The rule contains a conjunctive condition set denoted by lines beginning with '?'. Condition types for matches to InterPro entries (?IPRO), PROSITE signatures (?PSAC) and the eukaryotic taxonomic kingdom (?PSTX) are present. '?EMOT' refers to a condition for a separate statistical verification of PROSITE matches. Several different SWISS-PROT action types are present in lines beginning with '!'. Note: text following '#' symbols is non-parsed information such as headers or comments. Disjunctive condition sub-sets are separated by an explicit '?_OR_' symbol

based statistical and comparative analysis of each of the proteomes.

Comparative analyses

InterPro-based comparative analysis for each of the complete proteomes

Interactive InterPro-based proteome comparisons between any of the organisms in the database can be carried out using a web-based tool[11,15] that allows the user to compare a reference proteome with any other (one or more) proteomes in the database (Figure 3). The InterPro-based statistics are generated from the InterPro tables of protein matches and the lists of non-redundant complete proteome sets for each organism:

- *General statistics* where all InterPro entries with matches to the reference proteome are listed and the number of proteins matched for each InterPro entry are displayed. The number of proteins matched converts to a table that shows the protein accession number and the description taken from the SWISS-PROT/TrEMBL entry. The table has links to the InterPro graphical view and the CluSTr data. In this and other tables, the proteins are ranked according to a computed number of InterPro hits for each protein sequence. This number refers to the total of the number of times each of the signatures in the reference InterPro entry has a hit to the protein sequence and acts as an indicator of how well the protein sequence matches the signatures in the InterPro entry.

- *Top 30 hits* where the top 30 InterPro entries with the highest number of protein matches for the reference proteome are listed. The number of proteins matched for each InterPro entry and the percentage of proteome coverage that this number represents are displayed. The rank is indicated and represents the position of each InterPro entry in the list ordered according to the number of protein matches (eg see Figure 4).

- *15 most common families*, *15 most common domains* and *15 most common repeats* contain lists of the top 15 InterPro entries (of type 'family', 'domain' and 'repeat' respectively) with the largest number of protein matches, PARENTS only, for the reference proteome and displays the number of protein matches and clicking on the number of proteins matched brings up a table that lists the protein sets that match the reference InterPro entry, as described above.

- *Top 30 proteins with the highest number of different InterPro hits* lists the top 30 proteins from the reference proteome containing the largest number of different InterPro hits. The table shows the protein accession number and the description taken from the SWISS-PROT/TrEMBL entry, and has links to the InterPro graphical view.

- An additional feature of the interactive

Annotated TrEMBL entry using rule RU000418:

```
ID   Q9UVK5      PRELIMINARY;   PRT;   271 AA.
AC   Q9UVK5;
DT   01-MAY-2000 (TrEMBLrel. 13, Created)
DT   01-MAY-2000 (TrEMBLrel. 13, Last sequence update)
DT   01-MAR-2002 (TrEMBLrel. 20, Last annotation update)
DE   Heat shock protein 60 (Fragment){EI2}.
GN   hsp60{EI2}.
OS   Trichophyton mentagrophytes.
OC   Eukaryota; Fungi; Ascomycota; Pezizomycotina; Eurotiomycetes;
OC   Onygenales; Arthrodermataceae; mitosporic Arthrodermataceae;
OC   Trichophyton.
OX   NCBI_TaxID=82077{EI2};
RN   [1]{EI2}
RP   SEQUENCE FROM N.A.
RC   STRAIN=TM10;
RA   Raska M., Kopecek P., Zemanova E., Weigl E.;
RT   "Trichophyton mentagrophytes HSP60 cDNA. Isolation, sequencing and
RT   homology estimation.";
RL   Submitted (OCT-1999) to the EMBL/GenBank/DDBJ databases.
CC   -!- SIMILARITY: BELONGS TO THE CHAPERONIN (HSP60) FAMILY{EA1}.
DR   EMBL; AF199024; AAF07213.1; -.{EI2}
DR   HSSP; P06139; 1GRL.{EI3}
DR   InterPro; IPR001844; Chaprnin_Cpn60.
DR   InterPro; IPR002423; Cpn60/TCP-1.
DR   Pfam; PF00118; cpn60_TCP1; 1.
DR   PRINTS; PR00298; CHAPERONIN60.
KW   ATP-binding{EA1}; Chaperone{EA1}.
FT   NON_TER       1      1       {EI2}
FT   NON_TER     271    271       {EI2}
**
**   #################    INTERNAL SECTION    ##################
**EV EA1; Rulebase; -; RU000418; 29-JAN-2002.
**EV EI2; EMBL; -; AAF07213.1; 30-SEP-2001.
**EV EI3; HSSP_ADD; -; P06139; 15-APR-2002.
**PM Pfam; PF00118; cpn60_TCP1; 1; 271; T; 31-JAN-2002;
**PM PRINTS; PR00298; CHAPERONIN60; 159; 183; T; 09-AUG-2000;
**PM PRINTS; PR00298; CHAPERONIN60; 242; 268; T; 09-AUG-2000;
SQ   SEQUENCE   271 AA;  29200 MW;  672A2F1824FD14DC CRC64;
     AGCNPMDLRR GIQAAVQSVV EYLQANKRDI TTTEEIAQVA TISANGDTHV GKLISNAMEK
     VGKEGVITVK DGKTIDDELE VTEGMRFDRG YTSPYFITDP KTQKVEFEKP LILLSEKKIS
     AVQDILPALE ASTTLRRPLV IIAEDIDGVA LAVCILNKLR GQLQVAAVKA PGFGDNRKSI
     LGDIGILTNA TVFTDELDMK LDKATPDMLG STGSITITKE DTIILNGEGS KDAIAQRCEQ
     IRGVIADPAT SDYEKEKLQE RLAKLSGGLA V
//
```

Figure 2 (b): An example of a TrEMBL entry that has had annotation added as a result of implementation of this rule

comparisons tool is the option to compute a list of shared InterPro entries that are common to all the selected proteomes (this is similar in concept to the overlapping region of a Venn diagram).

Pre-computed InterPro-based proteome comparisons cover some of the most obvious proteome comparisons for selected organisms (Figure 5).

SWISS-PROT and InterPro have adopted GO standard vocabulary

Gene Ontology and InterPro

The GO project[4,5] arose from the need to have a universal annotation system for describing and querying genes or gene products. The GO vocabulary consists of three ontologies that describe molecular function, biological process and cellular component. Each vocabulary is structured as directed acyclic graphs (DAGs), wherein any term may have more than one parent as well as zero, one, or more children. This makes attempts to describe biology much richer than would be possible with a hierarchical graph.

To fully exploit the potential of the GO project to unite and transfer knowledge between database resources, the SWISS-PROT group at EBI has joined the GO Consortium and has adopted its standard vocabulary to characterise the activities of proteins in the SWISS-PROT, TrEMBL and

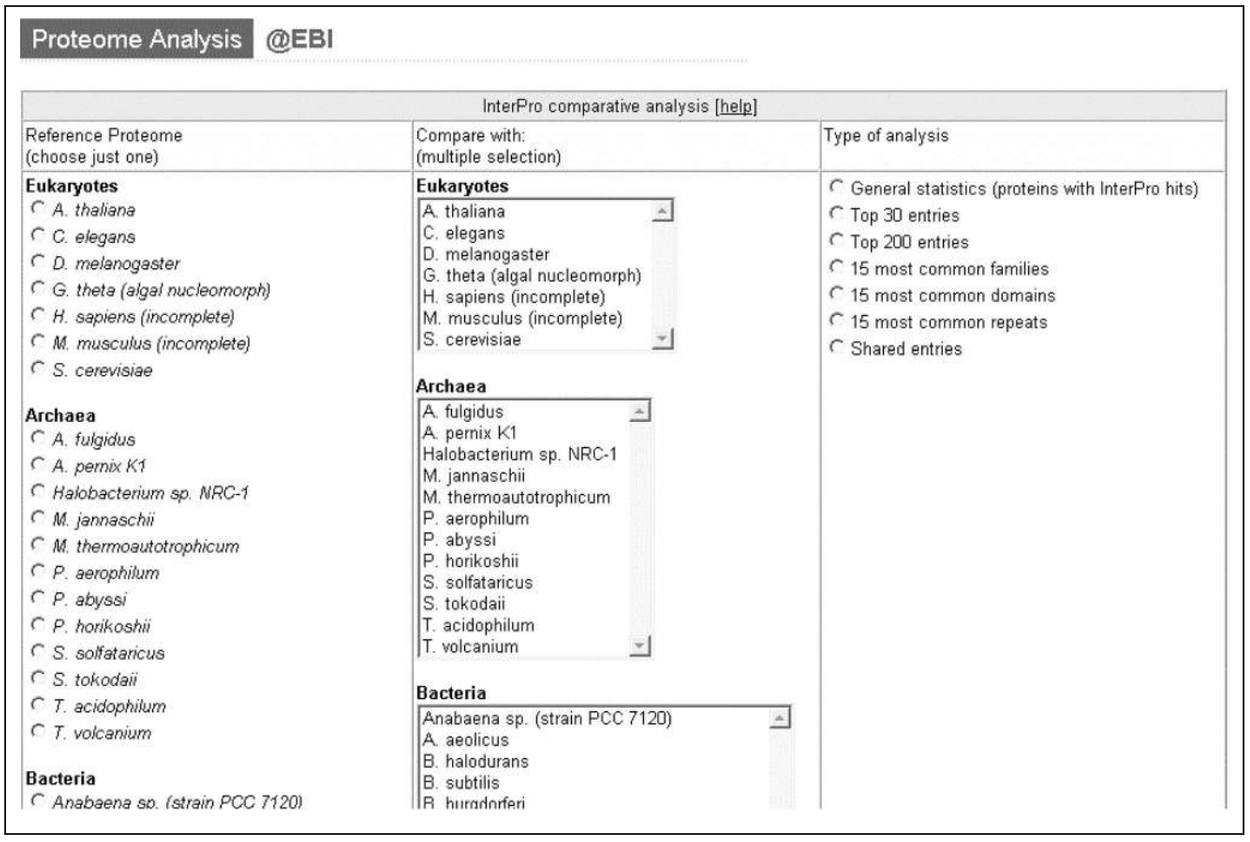
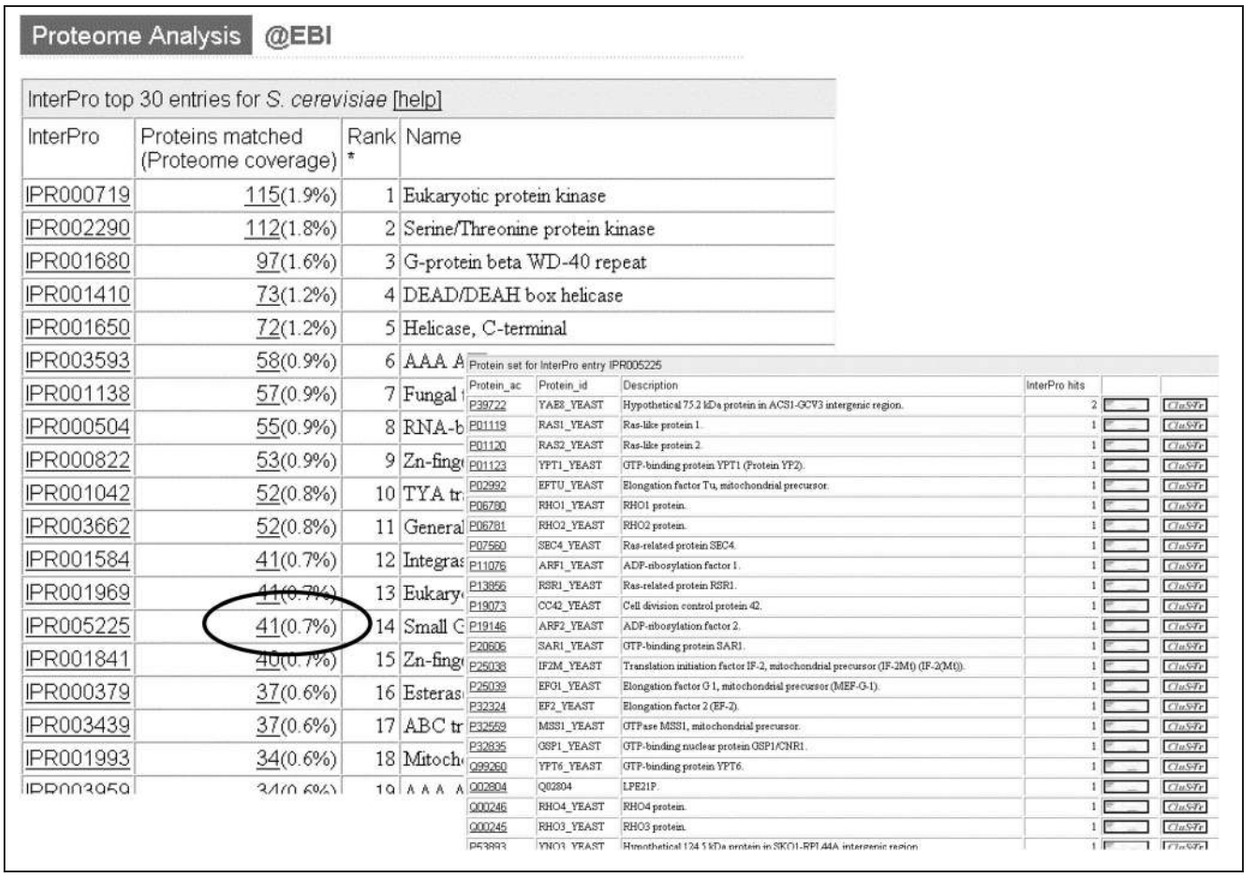


Figure 3: Interactive InterPro-based proteome comparisons can be carried out using this web-based tool

InterPro was manually mapped to GO

InterPro databases. Manual assignment of GO terms to protein sequences involves reading available information on a particular sequence and searching the GO ontologies for the appropriate terms to associate with the sequence. During this process curators frequently extend the scope of the GO ontologies by requesting new terms when necessary. Understandably the manual process is very slow, so automatic methods have been developed that use existing SWISS-PROT flatfile properties (the keywords (spkw2go)[16] and Enzyme Commission (EC) numbers (ec2go)[17]) manually mapped to high-level GO terms. These mappings have been electronically transferred to a table of matching SWISS-PROT and TrEMBL proteins and are available on the GO home page for use by external and collaborating databases.

InterPro plays a central role in the automatic assignments of GO terms to SWISS-PROT and TrEMBL records. InterPro entries provide annotation describing a set of related proteins, some of which may have identical molecular functions, be involved in the same processes, and perform their function in the same cellular locations. The assignment of GO terms to InterPro entries was done by manual inspection of the abstract of the entries and annotation of proteins in the match lists, and mapping of the appropriate GO terms of any level which apply to the whole protein, not necessarily only the domain described. The associated GO terms should also apply to all proteins with true hits to all signatures in the InterPro entry. For each associated term the name of the term and GO accession number is given, and these are visible in InterPro entries, with links to the EBI QuickGO GO



Proteome Analysis @EBI

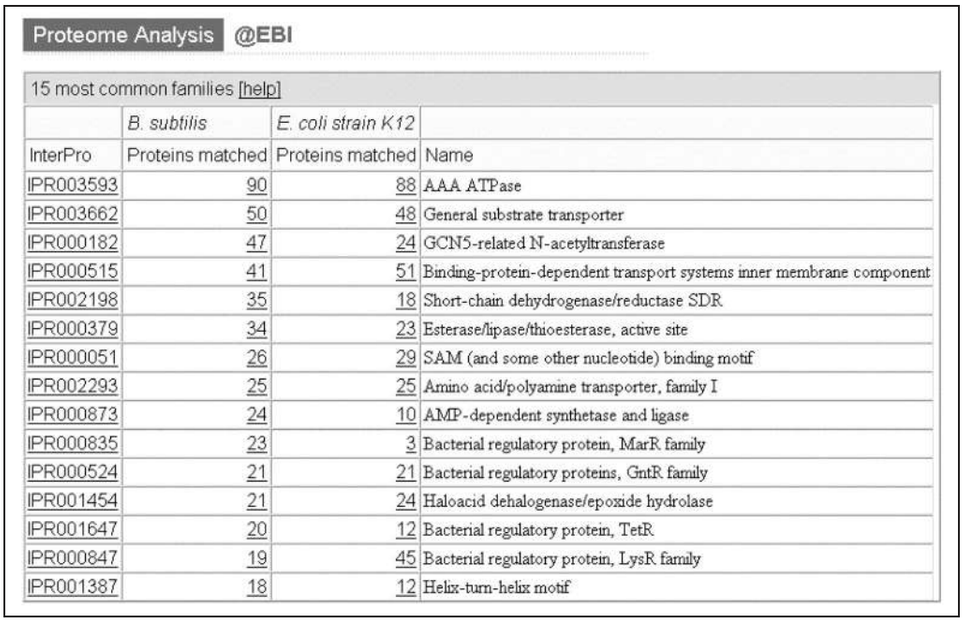
InterPro top 30 entries for *S. cerevisiae* [help]

InterPro	Proteins matched (Proteome coverage)	Rank *	Name
IPR000719	115(1.9%)	1	Eukaryotic protein kinase
IPR002290	112(1.8%)	2	Serine/Threonine protein kinase
IPR001680	97(1.6%)	3	G-protein beta WD-40 repeat
IPR001410	73(1.2%)	4	DEAD/DEAH box helicase
IPR001650	72(1.2%)	5	Helicase, C-terminal
IPR003593	58(0.9%)	6	AAA A
IPR001138	57(0.9%)	7	Fungal
IPR000504	55(0.9%)	8	RNA-b
IPR000822	53(0.9%)	9	Zn-fing
IPR001042	52(0.8%)	10	TYA tr
IPR003662	52(0.8%)	11	General
IPR001584	41(0.7%)	12	Integras
IPR001969	41(0.7%)	13	Eukary
IPR005225	41(0.7%)	14	Small G
IPR001841	40(0.7%)	15	Zn-fing
IPR000379	37(0.6%)	16	Esteras
IPR003439	37(0.6%)	17	ABC tr
IPR001993	34(0.6%)	18	Mitoch
IPR003959	34(0.6%)	19	AAA A

Protein set for InterPro entry IPR005225

Protein_ac	Protein_id	Description	InterPro hits		
P39722	YAE8_YEAST	Hypothetical 75.2 kDa protein in ACS1-GCV3 intergenic region.	2		CluSTr
P01119	RAS1_YEAST	Ras-like protein 1.	1		CluSTr
P01120	RAS2_YEAST	Ras-like protein 2.	1		CluSTr
P01123	YPT1_YEAST	GTP-binding protein YPT1 (Protein YP2).	1		CluSTr
P02992	EFTU_YEAST	Elongation factor Tu, mitochondrial precursor.	1		CluSTr
P06780	RHO1_YEAST	RHO1 protein.	1		CluSTr
P06781	RHO2_YEAST	RHO2 protein.	1		CluSTr
P07560	SEC4_YEAST	Ras-related protein SEC4.	1		CluSTr
P11076	ARF1_YEAST	ADP-ribosylation factor 1.	1		CluSTr
P13856	RSR1_YEAST	Ras-related protein RSR1.	1		CluSTr
P19073	CC42_YEAST	Cell division control protein 42.	1		CluSTr
P19146	ARF2_YEAST	ADP-ribosylation factor 2.	1		CluSTr
P20606	SAR1_YEAST	GTP-binding protein SAR1.	1		CluSTr
P25038	IF2M_YEAST	Translation initiation factor IF-2, mitochondrial precursor (IF-2M) (IF-2(Mt)).	1		CluSTr
P25039	EFG1_YEAST	Elongation factor G 1, mitochondrial precursor (MEF-G-1).	1		CluSTr
P32324	EF2_YEAST	Elongation factor 2 (EF-2).	1		CluSTr
P32559	MSS1_YEAST	GTPase MSS1, mitochondrial precursor.	1		CluSTr
P32835	GSP1_YEAST	GTP-binding nuclear protein GSP1/CNR1.	1		CluSTr
Q99260	YPT6_YEAST	GTP-binding protein YPT6.	1		CluSTr
Q02804	Q02804	LPE21P.	1		CluSTr
Q00246	RHO4_YEAST	RHO4 protein.	1		CluSTr
Q00245	RHO3_YEAST	RHO3 protein.	1		CluSTr
P53893	YNO3_YEAST	Hypothetical 124.5 kDa protein in SKO1-RPL44A intergenic region.	1		CluSTr

Figure 4: Top 30 hits for *organism* showing the number of proteins matched for each InterPro entry and the percentage of proteome coverage that this number represents. The rank indicates the position of each InterPro entry in the list ordered according to the number of protein matches. Clicking on the number of protein matches brings up a table showing the protein accession number and the description taken from the SWISS PROT/TrEMBL entry. The links to the InterPro graphical view and the CluSTr data are in the right hand columns



Proteome Analysis @EBI

15 most common families [help]

	B. subtilis	*E. coli strain K12*	
InterPro	Proteins matched	Proteins matched	Name
IPR003593	90	88	AAA ATPase
IPR003662	50	48	General substrate transporter
IPR000182	47	24	GCN5-related N-acetyltransferase
IPR000515	41	51	Binding-protein-dependent transport systems inner membrane component
IPR002198	35	18	Short-chain dehydrogenase/reductase SDR
IPR000379	34	23	Esterase/lipase/thioesterase, active site
IPR000051	26	29	SAM (and some other nucleotide) binding motif
IPR002293	25	25	Amino acid/polyamine transporter, family I
IPR000873	24	10	AMP-dependent synthetase and ligase
IPR000835	23	3	Bacterial regulatory protein, MarR family
IPR000524	21	21	Bacterial regulatory proteins, GntR family
IPR001454	21	24	Haloacid dehalogenase/epoxide hydrolase
IPR001647	20	12	Bacterial regulatory protein, TetR
IPR000847	19	45	Bacterial regulatory protein, LysR family
IPR001387	18	12	Helix-turn-helix motif

Figure 5: The pre-computed comparison of the '15 most common families' for *Bacillus subtilis* versus *Escherichia coli* K-12

browser.[18] Mapping of InterPro entries to GO terms thus provides an automatic means of assigning GO terms to the protein sequences that form the match table of a particular InterPro entry. An additional feature is that multifunctional proteins can be mapped to multiple GO terms through associations with more than one InterPro entry.

GO Slim and Proteome Analysis

GO Slim to summarise the functional annotation of the complete proteomes

To summarise the attributes of genes or gene products a slimmed down version of the GO vocabulary has been created. Under each ontology, a set of high-level terms has been selected to cover most aspects of each of the three ontologies without overlapping in paths of the GO hierarchy. This set of terms is described here as GO Slim. The Proteome Analysis pages display the GO statistics for each complete proteome set using the GO Slim terms.[19] These statistics are generated using the manual assignments of GO terms to protein sequences along with those generated automatically through the assignments based on SWISS-PROT keywords, EC numbers and InterPro. The mappings to more specific terms are collapsed to the parent GO Slim term and the number of proteins in the proteome set that map to each of the GO Slim terms is calculated. The results are displayed in a table of statistics for each organism (see eg Figure 6). Since proteins may be assigned to more than one GO term some proteins will have been counted more than once.

The functional classification and mapping to GO of InterPro families and domains, as well as SWISS-PROT keywords and EC numbers, provides a simple means of describing the composition of individual proteomes and provides a basis for comparative analysis. It also sets a framework for the automatic mapping to GO of proteins in SWISS-PROT and TrEMBL that have the keywords or EC numbers in the SWISS-PROT or TrEMBL flatfiles or have matches in InterPro. In addition, by virtue of the computed matches to InterPro entries new or previously uncharacterised protein sequences can be automatically mapped to GO.

CluSTr

The CluSTr (*Clu*sters of *S*WISS-PROT and *Tr*EMBL proteins) database[2,3] offers an automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. The clustering is based on the analysis of all pair-wise

GO Classification for *S. cerevisiae*

Term		Proteins	
GO:0003674	**molecular_function**	**2970**	**48.4%**
GO:0003676	nucleic acid binding	816	13.2%
GO:0030528	transcription regulator	118	1.9%
GO:0003754	chaperone	72	1.1%
GO:0003774	motor	17	0.2%
GO:0003824	enzyme	1636	26.6%
GO:0030234	enzyme regulator	48	0.7%
GO:0005194	cell adhesion molecule	1	0.0%
GO:0005198	structural molecule	193	3.1%
GO:0005215	transporter	409	6.6%
GO:0005488	ligand binding or carrier	1537	25.0%
GO:0004871	signal transducer	69	1.1%
GO:0008150	**biological_process**	**2643**	**43.0%**
GO:0008152	metabolism	1961	31.9%
GO:0006810	transport	506	8.2%
GO:0016265	death	3	0.0%
GO:0006928	cell motility	2	0.0%
GO:0006950	stress response	89	1.4%
GO:0007049	cell cycle	242	3.9%
GO:0007154	cell communication	191	3.1%
GO:0007275	developmental processes	31	0.5%
GO:0000004	biological_process unknown	2	0.0%
GO:0005575	**cellular_component**	**2409**	**39.2%**
GO:0005576	extracellular	5	0.0%
GO:0005623	cell	2362	38.4%
GO:0030312	external protective structure	31	0.5%
GO:0005941	unlocalized	25	0.4%
GO:0008372	cellular_component unknown	1	0.0%

Figure 6: The GO Slim statistics page for *Saccharomyces cerevisiae* showing the functional classification

comparisons between proteins using the Smith–Waterman algorithm. The statistical significance of the alignments is estimated using Monte-Carlo simulation resulting in a *Z*-score. Analysis carried out at different levels of protein similarity (based on the *Z*-score) yields a hierarchical organisation of clusters. Working with clusters at different levels of similarity biologically meaningful clusters can be selected for different groups of proteins greatly increasing the flexibility of the database.

CluSTr data is available for six complete eukaryote proteomes (*Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*) and for more then 50 prokaryotic proteomes.

InterPro and CluSTr

Links to the InterPro graphical interface allow users to see whether proteins from a cluster share the same domain architecture. Analysis of a cluster domain composition is even more apparent with the condensed graphical view, which shows a single representative for proteins with exactly the same domain architecture. In addition, the CluSTr database has links to InterPro and from there to the corresponding functional classification codes and GO terms making it is possible to identify protein functions within clusters. CluSTr also has cross-references to the structural databases HSSP and PDB.

CluSTr and Proteome Analysis

CluSTr data for each of the studied organisms is available and includes: (i) *general statistics* where the number of clusters of size greater than one and the number of singletons (clusters with one protein) at different levels of protein similarity as measured by their *Z*-scores are listed; (ii) *list of singletons*; (iii) *30 biggest clusters* and their InterPro-based functional classification; (iv) *clusters without InterPro matches*; and (v) *clusters without high-scoring segment pair (HSSP) links* (predicted secondary structure).

Structural information in Proteome Analysis

The structural information in the Proteome Analysis database is generated from the lists of non-redundant complete proteome sets for each organism. The information includes: *protein length distribution*, *amino acid composition* and links to *secondary and tertiary structure* represented the HSSP[20] and PDB[21] databases.

SUMMARY

Through both RuleBase and Proteome Analysis, InterPro provides a powerful way of classifying the large volumes of genomic data. Since the classification system that InterPro provides can be applied across species, the classification can be utilised to carry out cross-species comparisons that have strong biological meaning.

References

1. URL: http://www.ebi.ac.uk/interpro/
2. Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M. and Apweiler, R. (2001), 'CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins', *Nucleic Acids Res.*, Vol. 29(1), pp. 33–36.
3. URL: http://www.ebi.ac.uk/clustr
4. The Gene Ontology Consortium (2001), 'Creating the gene ontology resource: design and implementation', *Genome Res.*, Vol. 11, pp. 1425–1433.
5. URL: http://www.geneontology.org
6. Fleischmann, W., Moller, S., Gateau, A. and Apweiler, A. (1999), 'A novel method for automated functional annotation of proteins', *Bioinformatics*, Vol. 15, pp. 228–233.
7. URL: http://www.ebi.ac.uk/proteome/
8. URL: http://www.ebi.ac.uk/proteome/goslim_terms.html
9. Holm, L. and Sander, C. (1999), 'Protein folds and families: sequence and structure alignments', *Nucleic Acids Res.*, Vol. 27, pp. 244–247.
10. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.*, Vol. 28, pp. 235–242.
11. Kanapin, A., Apweiler, R., Biswas, M. *et al.* (2002), 'Interactive InterPro-based

comparisons of proteins in whole genomes', *Bioinformatics*, Vol. 18, pp. 374–375.

12. URL: http://www.ebi.ac.uk/proteome/CPhelp.html
13. Hubbard, T., Barker, D., Birney, E. *et al.* (2001), 'The Ensembl genome database project', *Nucleic Acids Res.*, Vol. 30, pp. 38–41.
14. URL: http://www.ensembl.org
15. URL: http://www.ebi.ac.uk/proteome/comparisons.html
16. URL: http://www.geneontology.org/spkw2go
17. URL: http://www.geneontology.org/ec2go
18. URL: http://www.ebi.ac.uk/ego/
19. URL: http://www.ebi.ac.uk/proteome/goslim_terms.html
20. URL: www.sander.ebi.ac.uk/hssp/
21. URL: http://www.ebi.ac.uk/msd/