Website Review

InterPro (The Integrated Resource of Protein Domains and Functional Sites)

<http://www.ebi.ac.uk/interpro/>

Christopher Southan*

*Correspondence to:

C. Southan, Bioinformatics Target
Discovery, SmithKline Beecham
Pharmaceuticals, New Frontiers
Science Park North, Third Avenue,
Harlow, Essex CM19 5AW, UK.
E-mail:
Christopher_D_Southan@sbphrd.com

Received: 15 September 2000
Accepted: 20 September 2000

Abstract

The family and motif databases, PROSITE, PRINTS, Pfam and ProDom, have been integrated into a powerful resource for protein secondary annotation. As of June 2000, InterPro had processed 384 572 proteins in SWISS-PROT and TrEMBL. Because the contributing databases have different clustering principles and scoring sensitivities, the combined assignments compliment each other for grouping protein families and delineating domains. The graphic displays of all matches above the scoring thresholds enables judgements to be made on the concordances or differences between the assignments. The website links can be used to analyse novel sequences and for queries across the proteomes of 32 organisms, including the partial human set, by domain and/or protein family. An analysis of selected HtrA/DegQ proteases demonstrates the utility of this website for detailed comparative genomics. Further information on the project can be found at the European Bioinformatics Institute at <http://www.ebi.ac.uk/interpro/> Copyright © 2000 John Wiley & Sons, Ltd.

Introduction

Secondary databases describing functional sites and domains are key resources for identifying structure–function relationships in protein sequences. PROSITE, initiated in 1988, is the archetype for these databases. Although many more have been developed since, they have used different approaches to the recognition of conserved sequence. These have been variously described as patterns, clusters, motifs, signatures or fingerprints but they all represent structurally conserved features related to protein function. The initial phase of the InterPro project has unified four complementary methods: PROSITE (Hofmann *et al.*, 1999); PRINTS (Attwood *et al.*, 2000); Pfam (Bateman *et al.*, 2000); and ProDom (Corpet *et al.*, 2000), each of which can be applied to an individual protein sequence using a formal score-based recognition cut-off. The first InterPro release (June 2000) contains 3052 entries, representing 574 domains, 2418 families, 46 repeats and 14 post-translational modification sites (Apweiler *et al.*, 2000). In addition to comprehensive protein annotation, the objectives of the project include the incorporation of new protein families and domains. It will also be

used to enhance the automated functional annotations in TrEMBL, the protein database supplementing SWISS-PROT (Bairoch and Apweiler, 2000) and is being applied as a proteome annotation tool for genome projects.

Site outline

The home page (Figure 1) includes a link to the Documentation page. This includes links to the release notes, user manual, a list of deleted InterPro entries, the scheme of the database, a fully annotated sample entry and references for the member databases. Brief descriptions and links to the member databases are given in the Databases page. The Search pages include a form to directly search the underlying ORACLE database for InterPro, Pfam, PRINTS, Prosite, SWISS-PROT, TrEMBL accession numbers and names, database names, entry types, *etc.* There is also a Sequence Retrieval System (SRS) link for more complex and/or multiple database queries. A key feature of the search page is ability to query a sequence directly against all three source databases and get the results back in the InterPro graphical format described

Figure 1. The InterPro home page

below. This would be useful either for a novel sequence or one that had entered TrEMBL since the last InterPro run. An extra feature is the option to include a Smith and Waterman search against the SWISS-PROT + TrEMBL (SPTR or SWALL) non-redundant protein database.

The protein and protein family displays

Before going further into what the site has to offer, the best way to grasp the concepts behind InterPro is by inspecting the graphic display for an individual protein entry and following the domain and

family links. This can be done via the example given on the home page but it is preferable to use a multi-domain protein for which you already have some background knowledge. Those less familiar with the theory and practice of protein sequence analysis will find useful information in the InterPro user manual, as well as in the documentation and publications from the source databases.

Figure 2 illustrates the concordance and discrepancies between the source entries for an *Escherichia coli* HtrA/DegQ gene that contains a serine protease domain and two C-terminal PDZ domains. In this example the PROSITE profiles and the hidden Markov models (HMMs) in Pfam have

Figure 2. An InterPro graphical display of entry TrEMBL 025663 from the IPR001940 HtrA/DegQ protease family <http://www.ebi.ac.uk/servlets/interpro/production/IEntry?ac=IPR001940>. The Prosite matches are orange, Pfam blue and PRINTS green

recognized both the trypsin and one of the PDZ domains, although these would be classified separately in the two source databases. In this case PRINTS provides a more discriminatory six-element fingerprint that is specific to the HtrA/degQ family of serine proteases, a subfamily (or 'child' in InterPro terminology) of the larger 'parent' trypsin family. However, in this case PRINTS recognizes only one of the two PDZ domains because many family members have only one rather than two PDZ domains. It should also be noted that the distal PDZ domain has been recognized by PROSITE but not by Pfam at the cut-off scores used.

From an individual protein entry there can be three types of links. Each of the graphic signatures has a link back to the source database, PS for PROSITE, PF for Pfam and PR for PRINTS. The second type of link is an InterPro domain type, e.g. all PDZ domains, regardless of what protein families they cluster with, will be linked to the domain entry IPR001478 as shown in Figure 3.

The numbers within this domain classification

indicate further subtleties and complementarity of the InterPro approach. While PROSITE and Pfam recognize (by application of their individual scoring cut-offs) 567 and 601 PDZ domains, respectively, the combination has an increased detection sensitivity for 629. The utility of this is shown in Figure 2, where PROSITE scored the second PDZ domain that Pfam and PRINTS 'missed'. This is a non-trivial result in that it immediately suggests the non-equivalence of the two domains, *i.e.* the sequence of the terminal domain is not a simple repeat.

The second type of InterPro entry describes a protein family as defined by the extended alignments in Pfam and/or the extended motif groups in PRINTS. The protease in Figure 2 links to the family-specific IPR001940 entry for the HtrA/DegQ proteases. This family (Figure 4), numbering 79 members in this first release, is defined by conformity to the PRINTS HtrA/DegQ motif group, which includes one proximal PDZ domain, and by 'membership' of the parent trypsin protease family IPR001254.

Database	InterPro
Accession	IPR001478 (matches 629 proteins)
Name	PDZ domain (also known as DHR or GLGF)
Type	Domain
Dates	08-OCT-1999 (created) 10-APR-2000 (last modified)
Signatures	PS50106; PDZ (567 proteins) PF00595; PDZ (601 proteins)
Abstract	PDZ domains are found in diverse signaling proteins in bacteria, yeast, plant, insect and vertebrate [1]. They may function in targeting signalling molecules to sub-membranous sites [2].

Figure 3. The InterPro PDZ domain entry. All the InterPro entries contain extensive descriptions and links but only the header sections are shown in these snapshots

Database	InterPro
Accession	IPR001940 (matches 79 proteins)
Name	HtrA/DegQ protease family
Type	Family
Dates	08-OCT-1999 (created) 01-MAR-2000 (last modified)
Signatures	PR00834; PROTEASES2C (79 proteins)
Parent	IPR001254; Serine proteases, trypsin family (1332 proteins)

Figure 4. The InterPro HtrA/DegQ protease family entry

As indicated in Figure 5, the parent (superfamily) for the HtrA proteases is the trypsin family, IPR001254. This entry also shows an apparent increase in sensitivity achieved by combining the results of the PROSITE and Pfam profile recognition methods for the detection of 1332 trypsin-like sequences. However, the entry also indicates that only 852 proteins have an active-site serine motif that conforms to the 11-element PROSITE regular expression, PS00135. The sensitivity, defined as true hits/(true hits+false negatives), for PS00135 is reported as 94% within SWISS-PROT and almost all the false negatives were attributable to loss of a conserved Gly in the first element. These apparent discrepancies in total family numbers illustrate the power of InterPro to highlight significant differences between the recog-

nition methods revealed by this scale of analysis. In this case we could conclude that 36% of the profile matches are catalytically inactive. However, an alternative interpretation is that the sensitivity of PS00135 may be lower for the expanded phylogenetic range of trypsin proteases now processed by InterPro compared with the 301 sequences matching PS00135 in SWISS-PROT, the majority of which are mammalian proteins. Support for this interpretation is provided by the HtrA/DegQ protease family. None of the 79 members have an exact match to the PS00135 active-site serine motif but many have been experimentally verified as proteases. In addition to the HtrA/DegQ example, the InterPro trypsin family has spawned two additional 'children', IPR00136, the alpha-lytic endopeptidases and IPR001314, the

Database	InterPro
Accession	IPR001254 (matches 1332 proteins)
Name	Serine proteases, trypsin family
Type	Family
Dates	08-OCT-1999 (created) 01-MAR-2000 (last modified)
Signatures	PS00134; TRYPSIN_HIS (919 proteins) PS00135; TRYPSIN_SER (852 proteins) PS50240; TRYPSIN_CATAL (1204 proteins) PF00089; trypsin (1182 proteins)
Children	IPR001316; Alpha-lytic endopeptidase serine protease (S2A) (22 proteins) IPR001940; HtrA/DegQ protease family (79 proteins) IPR001314; Chymotrypsin serine protease family (S1) (892 proteins)

Figure 5. The InterPro trypsin serine protease family entry

chymotrypsins. These are both defined by their scores against PRINTS signatures, although the latter is not classified as a separate family on the basis of accepted protease nomenclature (Barrett *et al.*, 1998)

Proteome analysis

InterPro is one of the tools the SWISS-PROT group at the EBI are using to give a perspective on domain structure and function, gene duplication and protein families in different genomes. An analysis has already been produced for the 31 proteomes from completely sequenced organisms and the incomplete human set (<http://www.ebi.ac.uk/proteome/>). The page of links for the human proteome so far is shown in Figure 6. A total of 18988 proteins were analysed from a combination of all human SWISS-PROT and TrEMBL entries. The number that had an assignment to at least one

domain was 13 275, *i.e.* 69% of all human proteins. Comparative figures for fly, worm and yeast are 54%, 54% and 52%, respectively. This difference could mean that the domain collections (and therefore the cut-off scores) are historically skewed towards mammalian proteins. Alternatively, the more primitive phyla may have taxon-specific domains which have not yet been analysed.

There are a large number of ways to query and compare the data from the Proteomics page, depending on the objectives of the analysis. The example chosen in Figure 7 shows the top 20 human protein domains compared with the three other completely sequenced eukaryotes. The results allow instant comparative genomic observations, so long as caution is exercised. For example, the conclusions that the fly, worm and yeast contain 3%, 2% and 1.8% protein kinases, respectively, but no histocompatibility proteins, seems solid enough. However, a visual inspection of the yeast Ig and

InterPro [help]	CluSTR [help]	Structure	Functional classification - Gene Ontology (GO)	Known Genes	Comparative analysis
General statistics (proteins with InterPro hits)	General statistics List of singletons	Primary Amino acid composition	General statistics (InterPro proteins with GO hits)	By Chromosome location Chr. 1 Chr. 2 Chr. 3 Chr. 4 Chr. 5 Chr. 6 Chr. 7 Chr. 8 Chr. 9 Chr. 10 Chr. 11 Chr. 12 Chr. 13 Chr. 14 Chr. 15 Chr. 16 Chr. 17 Chr. 18 Chr. 19 Chr. 20 Chr. 21 Chr. 22 Chr. X Chr. Y	InterPro-based proteome comparisons vs. <i>C. elegans</i>, <i>D. melanogaster</i> and <i>S. cerevisiae</i> InterPro top 30 hits vs. <i>C. elegans</i>, <i>D. melanogaster</i> and <i>S. cerevisiae</i> InterPro top 200 hits vs. <i>C. elegans</i>, <i>D. melanogaster</i> and <i>S. cerevisiae</i> 10 biggest InterPro protein families vs. <i>C. elegans</i>, <i>D. melanogaster</i> and <i>S. cerevisiae</i> 15 most common InterPro protein domains vs. <i>C. elegans</i>, <i>D. melanogaster</i> and <i>S. cerevisiae</i>
Top 30 hits	30 biggest clusters	Secondary Proteins with HSSP links in SPTR			
Top 200 hits	Clusters without InterPro links	Tertiary Proteins with PDB links in SPTR			
10 biggest protein families	Clusters without HSSP links				
15 most common protein domains					
15 proteins with the highest occurrence of a given domain					
15 proteins with the highest occurrence of Pfam signatures					
15 proteins with the highest occurrence of different InterPro hits					

Figure 6. A preliminary statistical and comparative analysis of the *H. sapiens* proteome is presented with links to alternative ways of sorting the data

	<i>H. sapiens</i> (Incomplete)	<i>D. melanogaster</i>	<i>C. elegans</i>	<i>S. cerevisiae</i>	
InterPro	Proteins matched	Proteins matched	Proteins matched	Proteins matched	Name
IPR003006	800	149	85	1	Immunoglobulin and major histocompatibility complex domain
IPR000719	652	219	363	118	Eukaryotic protein kinase
IPR000822	512	343	172	52	Zinc finger, C2H2 type
IPR000276	423	101	365	0(0)	Rhodopsin-like GPCR superfamily
IPR002290	376	136	184	112	Serine/Threonine protein kinases active-site
IPR000561	348	92	167	1	EGF-like domain
IPR001245	233	76	95	4	Tyrosine kinase catalytic domain
IPR001452	232	59	54	24	Src homology 3 (SH3) domain
IPR000504	224	145	106	57	RNA-binding region RNP-1 (RNA recognition motif)
IPR001777	201	49	39	2	Fibronectin type III domain
IPR001849	201	64	55	26	Pleckstrin homology (PH) domain
IPR002048	199	130	108	17	EF-hand family
IPR001356	192	95	83	7	Homeobox domain
IPR001680	187	160	111	88	G-protein beta WD-40 repeats
IPR001039	178	0(0)	0(0)	0(0)	Major histocompatibility complex protein, Class I
IPR001841	155	101	123	33	RING finger
IPR002126	151	18	15	0(0)	Cadherin domain
IPR002965	149	167	36	0(0)	Proline rich extensin
IPR001478	142	65	57	2	PDZ domain (also known as DHR or GLGF)
IPR001254	141	193	14	1	Serine proteases, trypsin family

Figure 7. The InterPro top 20 human hits vs. *C. elegans*, *D. melanogaster* and *S. cerevisiae*

EGF-like domain matches reveals them as false positives. Such conclusions can be quickly backed up by launching an SRS BLAST search directly from the protein link.

Following on from using the HtrA/DegQ example above, we can see that yeast has one trypsin and two PDZ domains. Inspection of the single yeast trypsin assignment shows these have converged in the entry for P53920, *i.e.* it contains both a trypsin and a PDZ domain but confined to the N-terminal section of the 997-residue protein (Figure 8). The intriguing explanation was provided in a review of the HtrA proteases some time ago (Pallen and Wren, 1997). The authors established by conventional sequence comparison that P53920 contains an ancient tandem duplication of an HtrA/DegQ protease, similar to Figure 2, with two PDZ domains but that the second trypsin domain had lost essential catalytic positions. Although the catalytic triad can be discerned for the N-terminal

section, the BLAST score is only $5e-7$ against the bacterial HtrA/DegQ proteases and the score drops to 0.001 for the C-terminal section. The consequences, seen in Figure 8, are that only the N-terminal trypsin domain and one of the PDZs have scored above the thresholds for the PROSITE and Pfam profiles. The single PROSITE profile match therefore brings P53920 into the IPR001254 trypsin family but the PRINTS match score was below the threshold that would have assigned it to IPR001940, the HtrA/DegQ subfamily. The distal C-terminal section of the ancient duplication has such low residual sequence similarity that no domains were recognized. Although the InterPro family and domain assignments 'missed' for the yeast P53920 could be considered false negatives, it is important to note that neither catalytic activity nor PDZ-mediated complexing has yet been documented for P53920.

According to the table in Figure 7, the yeast

Figure 8. InterPro domain matches for the N-terminal section of the yeast protein SWISS-PROT P53920

contains only one other PDZ domain; in this case the protein is the probable 26S proteasome regulatory subunit P27, SWISS-PROT P40555. Inspection and a BLAST search indicate that this may be a true positive.

The PRINTS alignment set that defines the HtrA/DegQ protease family is derived from a set of 40 exclusively prokaryotic sequences. The finding of one distantly related member in yeast therefore raises the question as to the possible occurrence in other eukaryotes. Inspection of the yeast InterPro top 200 entries for *C. elegans* (<http://www.ebi.ac.uk/proteome/CAEEL/interpro/top200.html>) lists 57 PDZ-containing proteins and 14 trypsin family assignments. A manual inspection of the trypsin assignments reveals two false positives but no overlaps with PDZs, *i.e.* there are no candidate IPR001940 members in the worm. Interrogation of the fly top 200 assignments (<http://www.ebi.ac.uk/proteome/DROME/interpro/top200.html>) gives 65 PDZs and 193 trypsins. These sets are too large to inspect manually but the intercept for the IPR001940 members can easily be found by using an SRS query. This establishes that the fly has only one HtrA/DegQ protease, TrEMBL Q9VFJ3. The human proteome set has 142 PDZs and 141 trypsin matches. The SRS query locates the intercept of these two features in three human HtrA/DegQ proteases, two full-length Q92743, O43464, and one partial sequence Q9UNS5. The graphic display for the first of these is shown in Figure 9.

Comparison between Figures 9 and 2 shows interesting similarities and differences between the bacterial and human representatives of this family across such a large phylogenetic distance. The human has only one PDZ but has gained two N-terminal domains, as defined by Pfam assignments, one insulin-related growth factor binding domain and one kazal protease inhibitor domain. Using SRS to check for the intercepts of IPR001940,

PF00050 and PF00219 finds five hits, the three human sequences already referred to above and two additional sequences from rat. This particular domain–family combination is therefore uniquely mammalian in the current protein databases.

These comparative genomic observations raise many intriguing questions about the biology and evolutionary history of these proteases that are beyond the scope of this review. However, the examples above show that this type of information can be rapidly acquired by using the tools available from the InterPro website. They can also be cross-checked using conventional database searching and sequence analysis tools from the parent site (<http://www.ebi.ac.uk/Tools/index.html>).

Conclusions

InterPro is destined to become the pre-eminent secondary database for protein annotation. This combination of motif and family profile databases into a single package offers a more discriminatory and comprehensive protein analysis resource than the individual databases. The fact that the contributing databases are built on different concepts produces a valuable compensation effect, *i.e.* their tendencies for false positives and false negatives are skewed in different directions. The key feature of the combined graphic display is that it will often allow a clear judgement between false positives, borderline matches and likely true positives, although it should be borne in mind that the latter should ideally be verified by experiment.

The ability to interrogate and compare the entire proteomes of organisms by domain and/or protein family distributions and combinations is another key feature of InterPro. As illustrated by the protease examples, inspection of the differential cut-off results can reveal subtleties of scientific

Figure 9. An InterPro graphical display of the human HtrA/DegQ protease sequence, TrEMBL 025663

significance that would be difficult or tedious to track down by other methods.

Considering the future, the automation capacity of InterPro should be able to keep pace with the genome projects. It is intended that the Ensembl human genome annotation output, already processed through Pfam, will eventually run through InterPro (Apweiler, personal communication). If this output is blended effectively with human TrEMBL by redundancy removal, this will provide an immediate comparative functional overview of the human proteome. The next production release of InterPro will integrate more of the ProDom database that was only partially incorporated in the first release. It is also planned to incorporate the Gene Ontology (GO) classification system (<http://genome-www.stanford.edu/GO/>) and eventually other signature databases will be included, such as Blocks and SMART (Henikoff *et al.*, 2000; Schultz *et al.*, 2000).

Acknowledgements

The InterPro project is based at the European Bioinformatics Institute (EBI) on the Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. All screen views from the website are reproduced with the kind permission of Rolf Apweiler. Other members of the project team are listed on the website at http://www.ebi.ac.uk/interpro/release_notes.html.

Source databases

Pfam protein families database of alignments of HMMs: <http://www.sanger.ac.uk/Software/Pfam/index.shtml>

PROSITE, database of protein families and domains: <http://www.expasy.ch/prosite/>

PRINTS protein fingerprint database: <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS>

The protein domain database, ProDom: <http://www.toulouse.inra.fr/prodom.html>

SWISS-PROT and TrEMBL: <http://www.ebi.ac.uk/swissprot/Information/information.html>

SRS (Sequence Retrieval System): <http://www.ebi.ac.uk/swissprot/Information/information.html>

European Bioinformatics Institute: database searching, browsing and analysis tools: <http://www.ebi.ac.uk/Tools/index.html>

The Ensemble Project (automated genome assembly): <http://www.ensembl.org/>

References

- Apweiler R, Attwood TK, Bairoch A *et al.* 2000. InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* (in press).
- Attwood TK, Croning MDR, Flower DR *et al.* 2000. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* **28**: 225–227.
- Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**: 45–48.
- Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL. 2000. The Pfam Protein Families Database. *Nucleic Acids Res* **28**: 263–266.
- Barrett AJ, Rawlings ND, Woessner JF (eds). 1998. *Handbook of Proteolytic Enzymes*. Academic Press: New York.
- Corpet F, Servant F, Gouzy J, Kahn D. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28**: 267–269.
- Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. 2000. Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res* **28**: 228–230.
- Hofmann K, Bucher P, Falquet L, Bairoch A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res* **27**: 215–219.
- Pallen MJ, Wren BW. 1997. The HtrA family of serine proteases. *Mol Microbiol* **26**(2): 209–221.
- Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. 2000. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**: 231–234.