BIOINFORMATICS

Vol. 16 no. 12 2000
Pages 1145–1150

InterPro—an integrated documentation resource for protein families, domains and functional sites

The InterPro Consortium

(R. Apweiler[1,*], T. K. Attwood[2], A. Bairoch[3], A. Bateman[4], E. Birney[1], M. Biswas[1], P. Bucher[5], L. Cerutti[4], F. Corpet[6], M. D. R. Croning[1,2], R. Durbin[4], L. Falquet[5], W. Fleischmann[1], J. Gouzy[6], H. Hermjakob[1], N. Hulo[2], I. Jonassen[7], D. Kahn[6], A. Kanapin[1], Y. Karavidopoulou[1], R. Lopez[1], B. Marx[1], N. J. Mulder[1], T. M. Oinn[1], M. Pagni[5], F. Servant[6], C. J. A. Sigrist[3] and E. M. Zdobnov)[1]

[1]EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]School of Biological Sciences, The University of Manchester, Manchester, UK, [3]Swiss Institute for Bioinformatics, Geneva, Switzerland, [4]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [5]Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland, [6]CNRS/INRA, Toulouse, France and [7]Department of Informatics, University of Bergen, HIB, Bergen, Norway

Received on April 14, 2000; revised on July 28, 2000; accepted on July 31, 2000

Abstract

Motivation: *InterPro is a new integrated documentation resource for protein families, domains and functional sites, developed initially as a means of rationalising the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database projects.*

Results: *Merged annotations from PRINTS, PROSITE and Pfam form the InterPro core. Each combined InterPro entry includes functional descriptions and literature references, and links are made back to the relevant parent database(s), allowing users to see at a glance whether a particular family or domain has associated patterns, profiles, fingerprints, etc. Merged and individual entries (i.e. those that have no counterpart in the companion resources) are assigned unique accession numbers. Release 1.2 of InterPro (June 2000) contains over 3000 entries, representing families, domains, repeats and sites of post-translational modification (PTMs) encoded by 6581 different regular expressions, profiles, fingerprints and Hidden Markov Models (HMMs). Each InterPro entry lists all the matches against SWISS-PROT and TrEMBL (more than 1 000 000 hits from 264 333 different proteins out of 384 572 in SWISS-PROT and TrEMBL).*

Availability: *The database is accessible for text- and sequence-based searches at http://www.ebi.ac.uk/interpro/.*

*To whom correspondence should be addressed.

Contact: *Interhelp@ebi.ac.uk*

Introduction

Pattern databases have become vital tools for identifying distant relationships in novel sequences and hence for inferring protein function. During the last decade, several pattern-recognition methods have evolved to address different sequence analysis problems, resulting in rather different and, for the most part, independent databases. To perform a comprehensive analysis, a user therefore has to know several important things. For example, what are the resources and where can they be found? What is the difference between them in terms of diagnostic performance and family coverage? What do the different search outputs mean? Is it sufficient to use just one of the databases, and if so, which one? Or, given the seeming complexity, won't PSI-BLAST (Altschul *et al.*, 1997) do just as well?

Currently, the most commonly-used pattern databases include: PROSITE, regular expressions and profiles (Hofmann *et al.*, 1999); Pfam, hidden Markov models (HMMs) (Bateman *et al.*, 2000); PRINTS, fingerprints (groups of aligned, un-weighted motifs) (Attwood *et al.*, 2000); and Blocks, aligned, weighted motifs or blocks (Henikoff *et al.*, 2000). Diagnostically, these resources have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. For example, regular expressions are

likely to be unreliable in the identification of members of highly divergent super-families (where profiles and HMMs excel); fingerprints perform relatively poorly in the diagnosis of very short motifs (where regular expressions do well); and profiles and HMMs are less likely to give specific sub-family diagnoses (where fingerprints excel).

In terms of family coverage, the pattern databases are similar in size but differ in content—each contains between 1000 and 2000 entries, spanning a range of globular and membrane proteins, modules and mosaics, repeats, and so on. While all of the resources share a common interest in protein sequence classification, some focus on divergent domains (e.g. Pfam), some focus on functional sites (e.g. PROSITE), and others focus on families, specialising in hierarchical definitions from super-family down to sub-family levels in order to pinpoint specific functions (e.g. PRINTS).

A number of sequence cluster databases are also commonly used in sequence analysis, for example to facilitate domain identification (e.g. ProDom, Corpet *et al.*, 2000). Unlike pattern databases, the clustered resources are derived automatically from sequence databases, using different clustering algorithms. This allows them to be relatively comprehensive, because they do not depend on manual crafting and validation of family discriminators; but the biological relevance of clusters can be ambiguous and may just be artefacts of particular thresholds.

Given these complexities, analysis strategies should endeavour to combine a range of databases, as none alone is sufficient. In concert, however, they can complement routine sequence database searches by providing more specific diagnoses than are possible with tools such as PSI-BLAST. PSI-BLAST highlights generic similarities by gathering sequences into families using an iterative profiling technique. However, there are problems with this approach. For example, if a multi-domain protein is matched, it may not be clear whether the region matched is the functional part of the protein, and hence whether functional annotations can be reliably transferred to the query; similarly, if a large super-family has been matched, it may be difficult to make the correct family or sub-family diagnosis.

In the task of sequence characterisation, we need more reliable, concerted methods for identifying protein family traits and for inheriting functional annotation. This is especially important given our dependence on automatic methods for assigning functions to the raw sequence data issuing from genome projects. Rationalising this process by creating a single coherent resource for diagnosis and documentation of protein families is difficult, given entirely different database formats, different search tools and different search outputs. Nevertheless, in an attempt to address some of these issues, we have developed InterPro. This new resource provides an integrated view of a number of commonly used pattern databases, and provides an intuitive interface for text- and sequence-based searches.

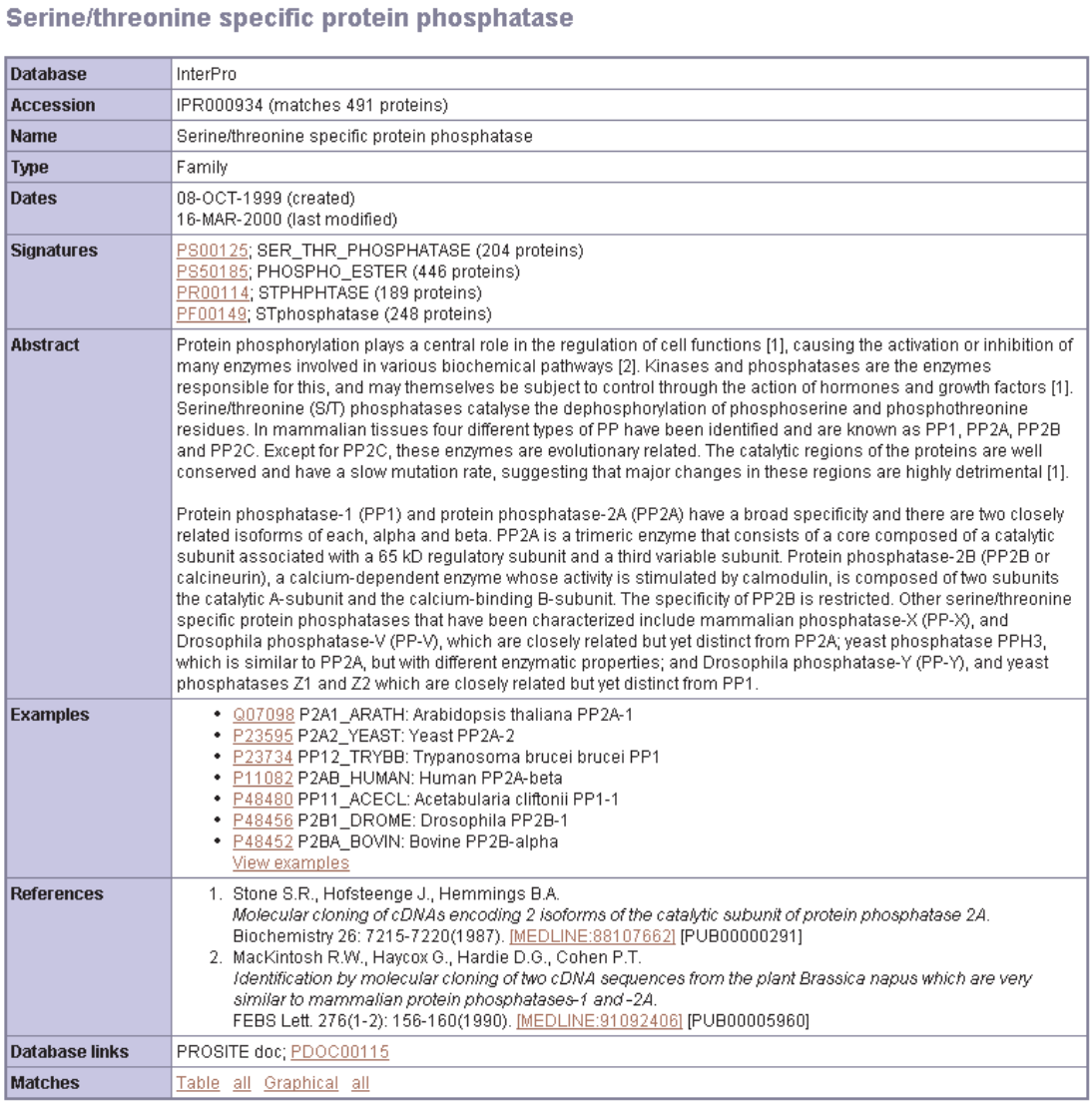
Source databases and methods

Release 1.2 of InterPro was built from Pfam 5.3 (2128 domains), PRINTS 26.1 (1310 fingerprints), ProDom 2000.1 (540 domains), PROSITE 16.0 (1370 families) and 241 preliminary profiles. The design of the member databases is such that they describe different protein patterns, some describe domains, while others describe families or motifs.

Flat-files submitted by each of the groups were systematically merged and dismantled. Where relevant, family annotations were amalgamated, and all method-specific annotation separated out. This process was complicated by the relationships that can exist, both between entries in the same database, and between entries in different databases. Different types of parent–child relationship were evident, leading us to recognise 'sub-types' and 'sub-strings'. A sub-string means that a motif or motifs are contained within a region of sequence encoded by a wider pattern (e.g. a PROSITE pattern is typically contained within a PRINTS fingerprint; or a fingerprint might be contained within a Pfam domain). A sub-type means that one or more motifs are specific for a sub-set of sequences captured by another more general pattern (e.g. a super-family fingerprint may contain several family- and sub-family-specific fingerprints; or a generic Pfam domain may include several family fingerprints).

Having classified the parent–child relationships of overlapping PROSITE, PRINTS and Pfam entries, all recognisably distinct entities were assigned unique accession numbers (which take the form IPR$xxxxxx$, where x is a digit). In doing this, we adopted the general principle that parent and children signatures with sub-string relationships usually have the same IPR numbers, while sub-type parent–child relationships warrant their own IPRs.

Each InterPro entry contains a list of matches to SWISS-PROT and TrEMBL (Bairoch and Apweiler, 2000); the match lists are provided by the member databases. An exception here concerns PROSITE pattern hits against TrEMBL, which undergo a different procedure—these are not provided by PROSITE and must therefore be derived by the TrEMBL group. All TrEMBL entries are scanned for PROSITE patterns. If a match is found, its significance is checked by means of a set of secondary patterns computed with the eMotif algorithm (Nevill-Manning *et al.*, 1998). For each family in PROSITE, the true members are aligned and fed into eMotif, which calculates a near optimal set of regular expressions, based on statistical rather than biological evidence. A stringency of 10^{-9} is used, so that each eMotif pattern is expected to produce a



Serine/threonine specific protein phosphatase

Database	InterPro
Accession	IPR000934 (matches 491 proteins)
Name	Serine/threonine specific protein phosphatase
Type	Family
Dates	08-OCT-1999 (created) 16-MAR-2000 (last modified)
Signatures	PS00125; SER_THR_PHOSPHATASE (204 proteins) PS50185; PHOSPHO_ESTER (446 proteins) PR00114; STPHPHTASE (189 proteins) PF00149; STphosphatase (248 proteins)
Abstract	Protein phosphorylation plays a central role in the regulation of cell functions [1], causing the activation or inhibition of many enzymes involved in various biochemical pathways [2]. Kinases and phosphatases are the enzymes responsible for this, and may themselves be subject to control through the action of hormones and growth factors [1]. Serine/threonine (S/T) phosphatases catalyse the dephosphorylation of phosphoserine and phosphothreonine residues. In mammalian tissues four different types of PP have been identified and are known as PP1, PP2A, PP2B and PP2C. Except for PP2C, these enzymes are evolutionary related. The catalytic regions of the proteins are well conserved and have a slow mutation rate, suggesting that major changes in these regions are highly detrimental [1]. Protein phosphatase-1 (PP1) and protein phosphatase-2A (PP2A) have a broad specificity and there are two closely related isoforms of each, alpha and beta. PP2A is a trimeric enzyme that consists of a core composed of a catalytic subunit associated with a 65 kD regulatory subunit and a third variable subunit. Protein phosphatase-2B (PP2B or calcineurin), a calcium-dependent enzyme whose activity is stimulated by calmodulin, is composed of two subunits the catalytic A-subunit and the calcium-binding B-subunit. The specificity of PP2B is restricted. Other serine/threonine specific protein phosphatases that have been characterized include mammalian phosphatase-X (PP-X), and Drosophila phosphatase-V (PP-V), which are closely related but yet distinct from PP2A; yeast phosphatase PPH3, which is similar to PP2A, but with different enzymatic properties; and Drosophila phosphatase-Y (PP-Y), and yeast phosphatases Z1 and Z2 which are closely related but yet distinct from PP1.
Examples	• Q07098 P2A1_ARATH: Arabidopsis thaliana PP2A-1 • P23595 P2A2_YEAST: Yeast PP2A-2 • P23734 PP12_TRYBB: Trypanosoma brucei brucei PP1 • P11082 P2AB_HUMAN: Human PP2A-beta • P48480 PP11_ACECL: Acetabularia cliftonii PP1-1 • P48456 P2B1_DROME: Drosophila PP2B-1 • P48452 P2BA_BOVIN: Bovine PP2B-alpha View examples
References	1. Stone S.R., Hofsteenge J., Hemmings B.A. *Molecular cloning of cDNAs encoding 2 isoforms of the catalytic subunit of protein phosphatase 2A.* Biochemistry 26: 7215-7220(1987). [MEDLINE:88107662] [PUB00000291] 2. MacKintosh R.W., Haycox G., Hardie D.G., Cohen P.T. *Identification by molecular cloning of two cDNA sequences from the plant Brassica napus which are very similar to mammalian protein phosphatases-1 and -2A.* FEBS Lett. 276(1-2): 156-160(1990). [MEDLINE:91092406] [PUB00005960]
Database links	PROSITE doc; PDOC00115
Matches	Table all Graphical all

Fig. 1. Example InterPro entry depicting the serine/threonine protein phosphatase family. This includes a signature from each of the member databases, an abstract derived from merged annotation of the member databases, a list of representative examples, the literature references cited in the abstract; and links to lists of matches in tabular or graphical form.

random or false positive hit in 10^{-9} matches. All pattern hits confirmed by eMotif are considered true; all others are flagged as unknown.

Database format

To facilitate in-house maintenance, InterPro is managed within a relational database system. However, the InterPro entries are also released in two ASCII (text) flatfiles in XML (extended markup language) format, one containing the core InterPro entries, and the other containing the protein matches. The files come together with a corresponding DTD (document type definition) file, to allow users to keep local InterPro copies on their machines.

Content of the current release

Release 1.2 (June 2000) contains over 3000 entries, representing families, domains, repeats and PTMs (post-translational modifications) encoded by 6581 different regular expressions, profiles, fingerprints and HMMs. Overall, InterPro methods have 1 099 807 hits from 264 333 protein sequences in SWISS-PROT and TrEMBL. Of these, 1 032 290 are true hits, 2522 are

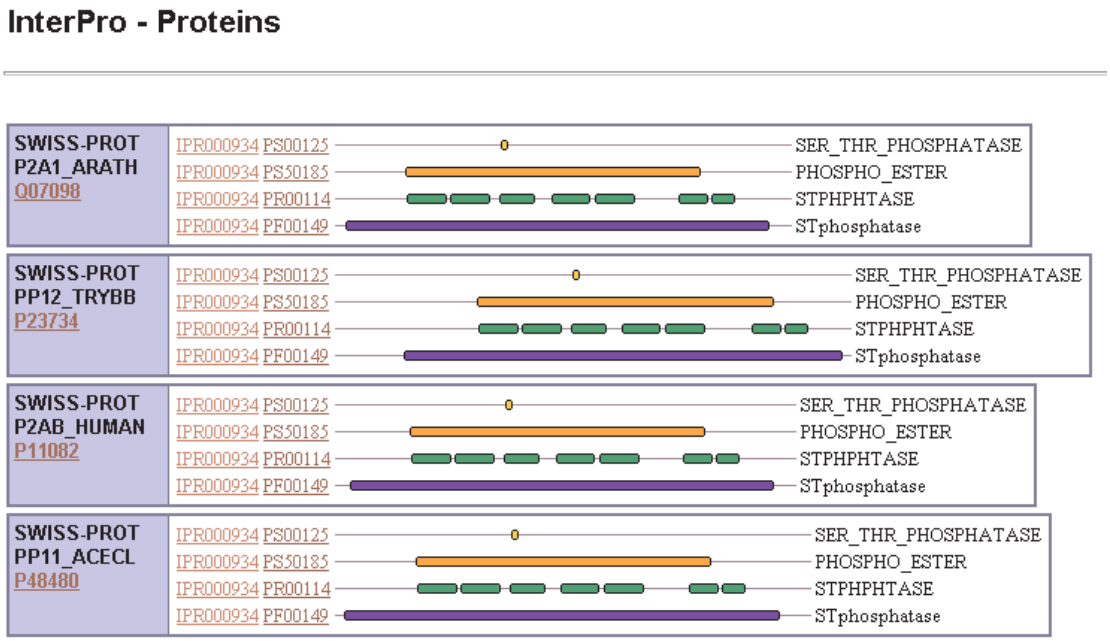


Fig. 2. InterPro graphical output for the serine/threonine specific protein phosphatase family. The signatures are colour coded, where blue depicts a Pfam HMM, green a PRINTS fingerprint, yellow a PROSITE signature, and orange a PROSITE profile. The widths of the coloured bands represent the boundaries of the signatures. The codes on the left-hand side of the figure are the accession numbers of the source databases and their corresponding IPRs, while those on the right-hand side are ID codes from the source databases.

partial hits, 3996 are false negative hits, 7523 are false positive hits (some PROSITE patterns with low specificity give false positive hits), and 53 476 have the status unknown. A complete content list is available from the Web site.

Individual InterPro entries contain: a list of member database signatures, HMMs, profiles or fingerprints associated with the entry; an abstract describing the family, domain, repeat or PTM, derived from merged annotation from the member databases; examples of representative sequences; literature references used to create the abstract; and links to tabular or graphical views of the matches to SWISS-PROT and TrEMBL. An example entry is shown in Figure 1. Interpretation of output is facilitated by means of a graphical user interface. Thus, for each sequence, the domain and/or motif organisation can be seen at a glance, as illustrated in Figure 2.

Database access and distribution

InterPro is accessible for interactive use via the EBI Web server (http://www.ebi.ac.uk/interpro), which can also be reached via each of the member databases. The Web interface allows text-based and sequence-based searches using SRS (Etzold *et al.*, 1996). The sequence-based search uses tools provided by the member databases, namely ppsearch for PROSITE patterns, pfscan for PROSITE profiles (Hofmann *et al.*, 1999), hmmpfam for Pfam HMMs (Bateman *et al.*, 2000), fpscan for PRINTS fingerprints (Attwood *et al.*, 2000), and BlastProDom for ProDom patterns (Corpet *et al.*, 2000). The results display matches to the parent databases and the corresponding InterPro entries, providing the positions of the signatures within the sequence, and a graphical view of the matches, as shown in Figure 3. Detailed results of matches to the individual database search methods are provided via hyperlinks to each of the parent databases. A mail server is available for sequence searches at Interproscan@ebi.ac.uk. Documentation on using the mail server can be obtained by emailing the address with the word ‘help’ in the body of the text.

The InterPro flatfile may be retrieved from the EBI anonymous-ftp server (ftp://ftp.ebi.ac.uk/pub/databases/interpro).

Future directions

While the initial InterPro release was created around PRINTS, PROSITE and Pfam, ProDom is being included in stages. Various factors rendered a step-wise approach to the development of InterPro desirable. First, the scale of the task of amalgamating just the first three databases was immense. The rational merging of apparently equivalent database entries that in fact simultaneously define a specific family, domains within that family, or even

InterPro search Results.

1 Query Sequence submitted Length 210 aa.

InterPro	Results of PPsearch against PROSITE	Results of PFScan against PROSITE	Results of FingerPRINTScan against PRINTS	Results of HMMDecypher against PFAM-A
IPR000485 Bacterial regulatory proteins, asnC family			PR00033 [167-179] [178-198]	
IPR000595 Cyclic nucleotide-binding domain	PS00888 [30-46] PS00889 [71-89]	PS50042 [24-124]		PF00027 [18-112]
IPR001808 Bacterial regulatory proteins, Crp family	PS00042 [168-191]		PR00034 [166-183] [182-198]	PF00325 [166-197]
IPR002373 cAMP-dependent protein kinase			PR00103 [24-39] [39-54] [69-79]	

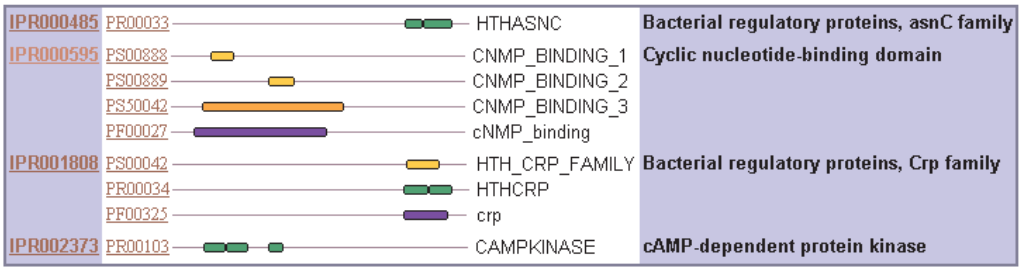


Fig. 3. Example results from a sequence-based search using SRS. The sequence query was the *Escherichia coli* catabolite gene activator, CRP (P03020). The output provides a summary of the hits in tabular and graphical form, with hyperlinks to the detailed results from the individual databases.

repeats within those domains, presented an enormous challenge. Thus, the immediate goal for InterPro was to limit the problem only to databases that offered annotation. A second important consideration was that while Pfam, PRINTS and PROSITE are true pattern databases, ProDom is based solely on automatic clustering of sequences by similarity (i.e. discriminators are not derived). Resulting clusters need not have precise biological correlations and some family designations have changed between database versions. It was therefore necessary that ProDom should adopt stable accession numbers before its entries could be meaningfully considered for inclusion in InterPro. This was recently achieved, and the first 540 ProDom entries were integrated for release 1.2. We expect the integration of a further 700 ProDom families into InterPro for release 2.

The Blocks database also began amalgamating a number of protein motif databases in their new Blocks+ database, which contains ungapped multiple alignments for families in PROSITE, PRINTS, Pfam-A, ProDom and Domo. The Blocks+ database, however, contains no annotation, and, unlike InterPro, systematic merging and dismantling of member database files were not achieved manually. Full Blocks releases are now based on families already in InterPro (J.Henikoff, personal communication), starting with release 12.0. Cross-referencing between Blocks and InterPro was therefore relatively straightforward and has been done for InterPro release 1.2. Once the founder members of the InterPro consortium have been assimilated into the unified resource, other pattern databases will also be included. First, scheduled for release 3, will be the SMART resource (Schultz *et al.*, 1998). Ultimately, we hope to include many other protein family databases to give a more comprehensive view of the resources available.

Discussion

A primary application of InterPro's family, domain and functional site definitions will be in the computational functional classification of newly determined sequences

that lack biochemical characterisation. For instance, the EBI is using InterPro for enhancing the automated annotation of TrEMBL (Fleischmann *et al.*, 1999). This process is more efficient and reliable than using each of the pattern databases separately, because InterPro provides internal consistency checks and deeper coverage. Furthermore, InterPro has been used for the comparative genome analysis of *Drosophila melanogaster, Caenorhabditis elegans*, and *Saccharomyces cerevisiae,* and has there proven its usefulness for whole proteome analysis (Rubin *et al.*, 2000).

Another major use of InterPro will be in identifying those families and domains for which the existing discriminators are not optimal and could hence be usefully supplemented with an alternative pattern (e.g. where a regular expression identifies large numbers of false matches it could be useful to develop an HMM, or where a Pfam entry covers a vast super-family it could be beneficial to develop discrete family fingerprints, and so on). Alternatively, InterPro is likely to highlight key areas where none of the databases has yet made a contribution and hence where the development of some sort of pattern might be useful. For example, sequence groups from ProDom are being analysed using the Pratt pattern discovery tool (Jonassen *et al.*, 1995; Jonassen, 1997) to reveal clusters that can form InterPro families and to create regular expression discriminators. This united approach should thus help us to improve both the utility and the coverage of pattern databases, pinpointing weaknesses and allowing us to remedy them efficiently.

Conclusions

InterPro is an international initiative that was conceived in an attempt to streamline the efforts of the pattern database providers. The project aims to reduce duplication of effort in the labour-intensive, rate-limiting process of annotation, and will facilitate communication between the disparate resources. By uniting these databases, we capitalise on their individual strengths, producing a single entity that is far greater than the sum of its parts. As it evolves, InterPro will streamline the analysis of newly determined sequences for the individual user, and will make a significant contribution in the demanding task of automatic annotation of predicted proteins from genome sequencing projects.

Acknowledgements

The InterPro project is supported by grant number BIO4-CT98-0052 of the European Commission. TKA is a Royal Society University Research Fellow.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Attwood,T.K., Croning,M.D.R., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bairoch,A. and 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,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.

Fleischmann,W., Möller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.

Henikoff,J.G., Greene,E.A., Pietrokovski,S. and 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. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Jonassen,I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, **13**, 509–522.

Jonassen,I., Collins,J.F. and Higgins,D. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.

Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. USA*, **95**, 5865–5871.

Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Li,P.W., Apweiler,R., Fleischmann,W., Cherry,J.M., Henikoff,S., Skupski,M.P., Misra,S., Ashburner,M., Birney,E., Boguski,M.S., Brody,T., Brokstein,P., Celniker,S.E., Chervitz,S.A., Coates,D., Cravchik,A., Gabrielian,A., Galle,R.F., Gelbart,W.M., George,R.A., Goldstein,L.S., Gong,F., Guan,P., Harris,N.L., Hay,B.A., Hoskins,R.A., Li,J., Li,Z., Hynes,R.O., Jones,S.J., Kuehl,P.M., Lemaitre,B., Littleton,J.T., Morrison,D.K., Mungall,C., O'Farrell,P.H., Pickeral,O.K., Shue,C., Vosshall,L.B., Zhang,J., Zhao,Q., Zheng,X.H., Zhong,F., Zhong,W., Gibbs,R., Venter,J.C., Adams,M.D. and Lewis,S. (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA*, **95**, 5857–5864.