The InterPro Consortium is a group of scientists responsible for the production and maintenance of InterPro and its member databases: PRINTS, PROSITE, Pfam, ProDom, SMART and TIGRFAMs. The consortium members are situated at different research centres around the world.

InterPro: An integrated documentation resource for protein families, domains and functional sites

The InterPro Consortium: Nicola J. Mulder, Rolf Apweiler, Terri K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Margaret Biswas, Paul Bradley, Peer Bork, Phillip Bucher, Richard Copley, Emmanuel Courcelle, Richard Durbin, Laurent Falquet, Wolfgang Fleischmann, Jerome Gouzy, Sam Griffith-Jones, Daniel Haft, Henning Hermjakob, Nicolas Hulo, Daniel Kahn, Alexander Kanapin, Maria Krestyaninova, Rodrigo Lopez, Ivica Letunic, Sandra Orchard, Marco Pagni, David Peyruc, Chris P. Ponting, Florence Servant and Christian J. A. Sigrist

Date received (in revised form): 14th June 2002

Abstract

The exponential increase in the submission of nucleotide sequences to the nucleotide sequence database by genome sequencing centres has resulted in a need for rapid, automatic methods for classification of the resulting protein sequences. There are several signature and sequence cluster-based methods for protein classification, each resource having distinct areas of optimum application owing to the differences in the underlying analysis methods. In recognition of this, InterPro was developed as an integrated documentation resource for protein families, domains and functional sites, to rationalise the complementary efforts of the individual protein signature database projects. The member databases – PRINTS, PROSITE, Pfam, ProDom, SMART and TIGRFAMs – form the InterPro core. Related signatures from each member database are unified into single InterPro entries. Each InterPro entry includes a unique accession number, functional descriptions and literature references, and links are made back to the relevant member database(s). Release 4.0 of InterPro (November 2001) contains 4,691 entries, representing 3,532 families, 1,068 domains, 74 repeats and 15 sites of post-translational modification (PTMs) encoded by different regular expressions, profiles, fingerprints and hidden Markov models (HMMs). Each InterPro entry lists all the matches against SWISS-PROT and TrEMBL (2,141,621 InterPro hits from 586,124 SWISS-PROT and TrEMBL protein sequences). The database is freely accessible for text- and sequence-based searches.

Keywords: *signature database, domain, protein classification, functional annotation*

INTRODUCTION

In June 2000 the first draft of the human genome sequence was announced, and was accompanied by promises of significant advances in medical science. However, these promises cannot be met simply with the flood of raw data from the genome sequencing projects, the data need to be converted into useful biological information. To live up to the promises the first obstacle is in classifying the genes and assigning a function to their products.[1] With the scale of the influx of raw sequence data from genome sequencing projects, manual annotation of all gene products is no longer possible, and therefore there is a need for reliable automatic methods for protein sequence analysis and classification. Traditional methods of annotation involve searching the query sequence against an existing protein database. Such methods are often confounded by the presence of low-complexity sequence or repetitive elements as well as local regions of sequence similarity around genetically mobile domains. In addition, sequences may be evolutionarily related, although

Nicola J. Mulder,
EMBL Outstation,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, UK

Tel: +44 (0) 1223 494 602
Fax: +44 (0) 1223 494 468
E-mail: mulder@ebi.ac.uk

their sequence divergence may be to such an extent that they are not picked up in a sequence similarity search.

With the increase in population of protein sequence databases the number of related sequences increases, so when a search is performed, it identifies a large set of highly related sequences and the less related sequence hits may be lost. The occurrence of multi-domain proteins also increases the complexity of sequence searches since some domains may be present in many different combinations in a protein sequence. It is for these reasons that protein signature databases evolved and have become increasingly useful tools for protein sequence analysis and identifying domains or classifying proteins into families. In this paper we use the word 'signature' to describe diagnostic entities for a domain, family, etc., which may be produced by several different methods.

Protein signature databases have evolved

Protein signature methods

The most useful tools use various methods for identifying motifs or domains found in previously characterised protein families. The basic information about a protein comes from the sequence. From one sequence it is difficult to infer any information about the protein; however, as the number of related sequences increases, so an alignment can be built to create a consensus for protein families, or identify conserved domains or highly conserved residues which may be important for function, eg an active site. These conserved areas of a protein family, domain or functional site can be used to recreate identifiable features using several different methods. These include building up regular expressions to show patterns of conserved amino acid residues; producing profiles from sequence alignments; and hidden Markov models (HMMs), which are profiles with a more complex probabilistic scoring mechanism. A profile is built from a sequence alignment, and is a table of position-specific amino acid weights and gap costs, in other words matrices describing the probability of finding an amino acid at a given position in the sequence.[2] The numbers in the table (scores) are used to calculate similarity scores between a profile and a sequence for a given alignment. For each set of sequences a threshold score is calculated so that only sequences scoring above this threshold are considered to be related to the original set of sequences in the alignment.

Consensus for protein families or domains

Each method has its own advantages. For example, patterns are relatively simple to build, and are very useful for small regions of conserved amino acids such as active sites or binding sites, but fail to provide information about the rest of the sequence. Because of the constraints on which amino acids may be found in a given area of the sequence, patterns fail to pick up related sequences with a small divergence in that particular area. Profiles and HMMs compensate for these downfalls in that they generally cover larger areas of the sequence, and since all amino acids have a chance of occurring at a given position, albeit with a lower probability or score, more divergent family members may still be included in the hit-list (hit-list in this paper refers to the list of proteins that match or contain a particular pattern or profile above the required score).

Protein signature databases

There are a number of well-known pattern databases in the public domain which use the methods described above to produce diagnostic signatures for protein families, domains, repeats, active sites, binding sites and post-translational modifications. These include PROSITE regular expressions and profiles,[3] PRINTS fingerprints (groups of aligned, unweighted motifs),[4] Pfam,[5] SMART[6] and TIGRFAMs HMMs,[7] and Blocks aligned, weighted motifs or blocks.[8] There are also several databases that identify protein families or domains using sequence clustering and alignment methods, for example ProDom.[9] A list of the major protein signature databases is shown in Table 1 with their URLs.

Table 1: List of the major pattern databases, a description of the database and their URLs

Database	Description	URL
SWISS-PROT/ TrEMBL	Protein sequence databases	http://www.ebi.ac.uk/swissprot/or http://www.expasy.org/sprot/
PROSITE	Database of patterns and profiles describing protein families and domains	http://www.expasy.org/prosite/
PRINTS	Compendium of protein fingerprints	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
Pfam	Collection of multiple sequence alignments and HMMs	http://www.sanger.ac.uk/Software/Pfam/index.shtml
SMART	A Simple Modular Architecture Research Tool – a collection of protein families and domains	http://smart.embl-heidelberg.de/ http://smart.ox.ac.uk
TIGRFAMs	Protein families based on HMMs	http://www.tigr.org/TIGRFAMs/index.shtml
ProDom	Automatic compilation of homologous domains	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
PIR-ALN	Curated database of protein sequence alignments	http://pir.georgetown.edu/pirwww/dbinfo/piraln.html
ProClass	Non-redundant protein database organised by family relationships	http://pir.georgetown.edu/gfserver/proclass.html
Blocks	Database of protein alignment blocks	http://blocks.fhcrc.org
InterPro	Integrated documentation resource for protein families, domains and functional sites	http://www.ebi.ac.uk/interpro/
IProClass	Integrated protein classification database	http://pir.georgetown.edu/iproclass/

Protein signature databases address different sequence analysis problems

While all of the protein signature databases share a common interest in protein sequence classification, they have each evolved to address different sequence analysis problems, resulting in rather different and, for the most part, independent databases. In terms of family coverage, the pattern databases are similar in size but differ in content. The different areas of optimum application for each resource is due to the different strengths and weaknesses of their underlying analysis methods, as well as differences in their focus. For example, regular expressions are likely to be unreliable in the identification of members of highly divergent superfamilies (where profiles and HMMs excel); fingerprints also have a weakness in highly divergent families where there is insufficient ungapped sequence in the multiple sequence alignment from which to derive the motifs; profiles may 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 subfamily diagnoses (where fingerprints excel). Sequence cluster databases, for example ProDom, are also commonly used in sequence analysis, for example to facilitate domain identification. Unlike the signature databases, the clustered data are derived automatically from the protein sequence databases using different clustering algorithms. This allows ProDom to be comprehensive, but since they do not depend on manual crafting and validation of family discriminators, the biological relevance of clusters can be ambiguous.

Clustering algorithms

Comprehensive

Integrated documentation resource

To facilitate the coverage of the protein signature databases and accuracy of the signatures themselves in protein sequence classification, an integrated documentation resource that combines them into a single coherent database was created. The process of integration is non-trivial, given the disparity in database formats, search algorithms and the output that each database generates. The result is that InterPro,[10] an integrated documentation resource for protein

families, domains and functional sites, was developed. InterPro provides an integrated view of a number of commonly used pattern databases, and provides an intuitive interface for text- and sequence-based searches.[11]

Text- and sequence-based searches

INTEGRATION INTO INTERPRO

Domain composition

InterPro currently contains diagnostic protein signatures from PROSITE, PRINTS, Pfam, ProDom, SMART and TIGRFAMs. Signatures from the different databases that describe the same protein family, domain, repeat or post-translational modification (PTM) are integrated into a single InterPro entry with a unique InterPro accession number. The guidelines for integration are that the signatures must overlap, at least in part, in position on the protein sequence; they should have at least 75 per cent overlap in the protein match lists; and they must all describe the same biological entity whether it be a family, domain, etc. New signatures from member databases are manually integrated by curators using a list of protein matches for the new signatures and a list of overlaps between new and existing signatures. New signatures are either integrated into existing InterPro entries or assigned unique InterPro accession numbers, following the guidelines described above.

There are cases where a signature(s) matches a set of proteins that is a subset of a larger group of proteins matched by a different, but overlapping, signature(s). In this case the signatures are assigned unique InterPro accession numbers, which are then related to each other. There are two types of relationships in InterPro: the parent/child and the contains/found in relationship. In parent/child relationships child signatures should match subsets within the parental set of matches. The requirement for signature specificity is paramount both between different families as well as children within the same family (siblings). An example of such a relationship is the tubulin family, which is described in InterPro entry IPR000217 with matches to all types of tubulins. Children of this family include the more specific alpha (IPR002452), beta (IPR002453), gamma (IPR002454), delta (IPR002967), epsilon (IPR004057) and zeta (IPR004058) tubulins. Proteins matching the children entries also have matches to the parent tubulin entry (IPR000217).

Parent/child and contains/found in relationships

The second relationship between InterPro entries is the contains/found in relationship. This is used to indicate domain composition. Some domains can be found in more than one type of protein or family of proteins, but is not a subtype in the family sense. The domain is a separate, mobile entity, which can be found in proteins with different domain organisations. An example is the C2 domain (IPR000008), which is found in several different protein families, including phosphoinositide-specific phospholipase C (IPR001192) and synaptotagmin (IPR001565).

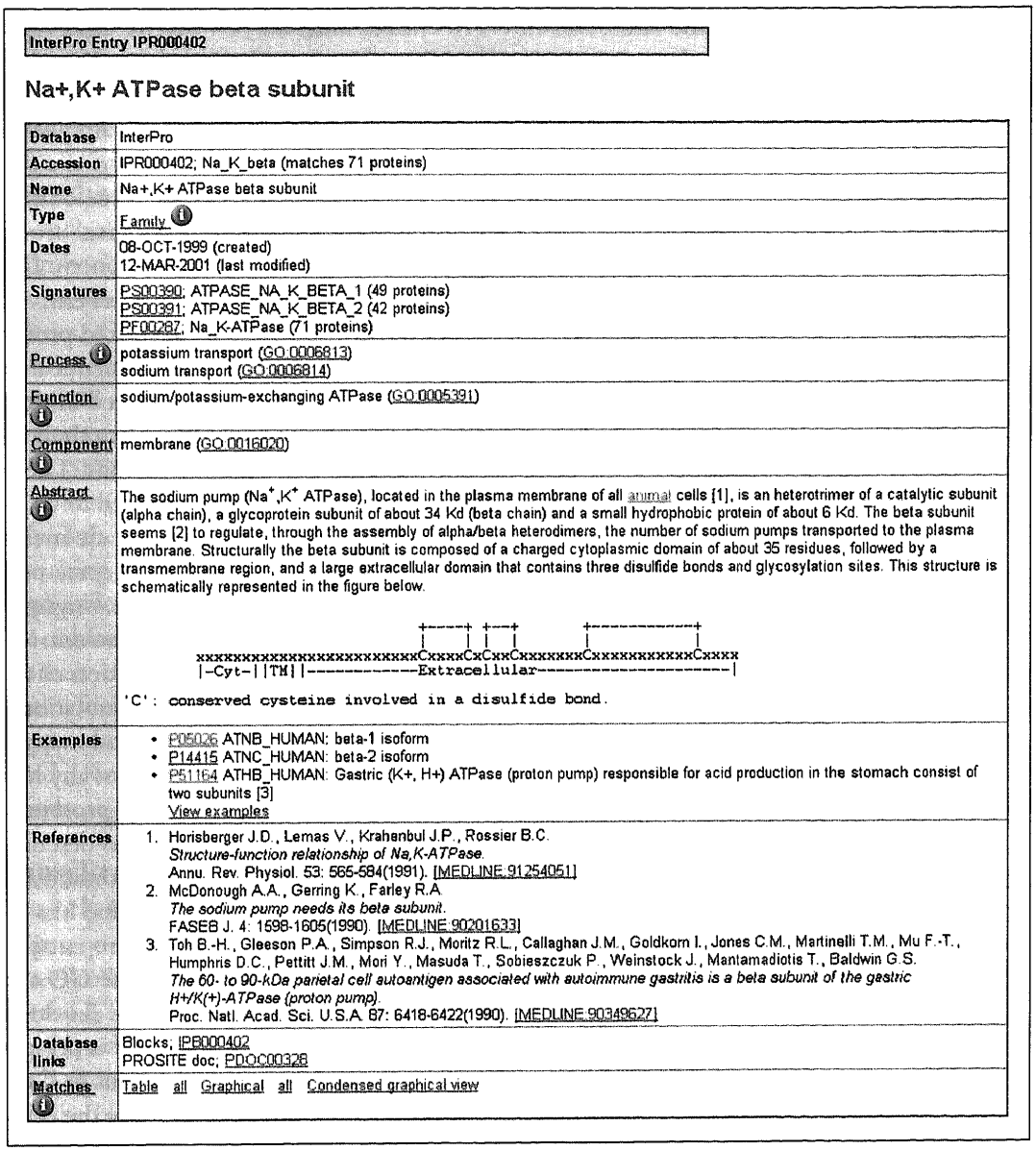
INTERPRO CONTENT

Release 4.0 of InterPro was built from Pfam 6.6, PRINTS 31.0, ProDom 2001.1, PROSITE 16.37, SMART 3.1, TIGRFAMs 1.2, SWISS-PROT 40.1 and TrEMBL 18.1. At the time of the release InterPro contained 4,691 entries, representing 3,532 families, 1,068 domains, 74 repeats and 15 sites of PTMs. Each entry contains one or more signatures from the individual member databases which all describe the same group of proteins. An example entry is shown in Figure 1. All entries contain annotation and a list of the proteins matching the entry.

Protein matches

Protein matches

Probably the most important part of an InterPro entry is the protein match information. Each InterPro entry contains a list of precomputed matches to SWISS-PROT and TrEMBL.[12] The original match lists are provided by the member databases with updates for new or changed protein sequences calculated at the EBI. An exception here concerns



InterPro Entry IPR000402

Na+,K+ ATPase beta subunit

Database	InterPro
Accession	IPR000402; Na_K_beta (matches 71 proteins)
Name	Na+,K+ ATPase beta subunit
Type	Family
Dates	08-OCT-1999 (created) 12-MAR-2001 (last modified)
Signatures	PS00390; ATPASE_NA_K_BETA_1 (49 proteins) PS00391; ATPASE_NA_K_BETA_2 (42 proteins) PF00287; Na_K-ATPase (71 proteins)
Process	potassium transport (GO:0006813) sodium transport (GO:0006814)
Function	sodium/potassium-exchanging ATPase (GO:0005391)
Component	membrane (GO:0016020)
Abstract	The sodium pump (Na^+,K^+ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane. Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below. +----+ +--+ +----------+ | | | | | xxxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxxCxxxxxxxxxxxxCxxxx |-Cyt-||TM||----------Extracellular----------| 'C': conserved cysteine involved in a disulfide bond.
Examples	• P05026 ATNB_HUMAN: beta-1 isoform • P14415 ATNC_HUMAN: beta-2 isoform • P51164 ATHB_HUMAN: Gastric (K+, H+) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3] View examples
References	1. Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. *Structure-function relationship of Na,K-ATPase.* Annu. Rev. Physiol. 53: 565-584(1991). [MEDLINE:91254051] 2. McDonough A.A., Gerring K., Farley R.A. *The sodium pump needs its beta subunit.* FASEB J. 4: 1598-1605(1990) [MEDLINE:90201633] 3. Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T., Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. *The 60- to 90-kDa parietal cell autoantigen associated with autoimmune gastritis is a beta subunit of the gastric H+/K(+)-ATPase (proton pump).* Proc. Natl. Acad. Sci. U.S.A. 87: 6418-6422(1990). [MEDLINE:90349627]
Database links	Blocks; IPB000402 PROSITE doc; PDOC00328
Matches	Table all Graphical all Condensed graphical view

Figure 1: Example InterPro entry depicting the sodium/potassium ATPase beta subunit. This family is described by two signatures from PROSITE and one from Pfam, and contains: an abstract derived from merged annotation of the member databases, mappings to GO terms, a list of representative examples, the literature references cited in the abstract; and links to lists of matches in tabular or graphical form

InterProScan software package

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

Match list viewed graphically

Protein matches are calculated using the InterProScan[14] software package described below. The match lists may be viewed in a tabular form, which lists the protein accession numbers and the positions in the amino acid sequence where each signature from that InterPro entry hits. The match list can also be viewed graphically, in which the sequence is split into several lines, one for each hit by a unique signature. This view includes the hits by all signatures from the

Domain organisation

Consensus domain boundaries

Comprehensive annotation

same and other InterPro entries: thus for each sequence, the domain and/or motif organisation can be seen at a glance. The proteins can also be viewed graphically in a condensed view, which computes the consensus domain boundaries from all signatures within each entry, and splits the protein sequence into different lines for each InterPro entry matched. InterPro entries that are children of other entries are collapsed into one line with the parent entries, while domain entries are shown on separate line, thereby providing a simple view of family and domain composition. This provides the user with an alternative and simpler way of visualising matches to more than one InterPro entry. From this view, all proteins sharing a common domain architecture can be grouped, and the sequences aligned using Jalview[15] or DisplayFam.[16]

Annotation

Each InterPro entry has a unique accession number (which takes the form IPRxxxxxx, where x is a digit), short name and description (name). There is an abstract describing the entry (family, domain, repeat or PTM), derived from merged annotation from the member databases. Literature references used to create the abstract are stored in a reference field in each entry. Some additional uncited references may be present in this field for extra reading, including online cross-references to the Oxford University Press Protein Profiles project.[17] The links point to the information pages, from which there are links to the protein sequence alignments. A list of examples of representative sequences matching the signatures in an entry is provided with a link to the InterPro graphical view of these proteins. Where relationships exist between InterPro entries these are displayed in a 'parent', 'child', 'contains' or 'found in' field. Parent/child relationships can be displayed through a link to a hierarchy tree.

Mapping to Gene Ontology

Additional annotation is available for some entries in the form of mappings to Gene Ontology (GO) terms.[18] The GO project is an effort to provide a universal ontology for describing gene products across all species. The project provides sets of terms in a directed acyclic graph under the three ontologies: molecular function, biological process and cellular component. InterPro entries provide comprehensive 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. Therefore InterPro entries were mapped to GO terms to provide an automatic means of assigning GO terms to the corresponding proteins.

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.

Since GO terms were not and never will be developed to describe domains or sites, it is not possible to map InterPro domains to GO terms directly, but one can use the domain hits in InterPro to group related proteins. The aim of mapping of InterPro entries to GO terms is to provide an efficient automatic means of large-scale GO characterisation of proteins, not of the domains or families themselves. Therefore, while an InterPro domain may not infer a particular function on a protein, that protein may have that function from another domain it contains and therefore has the potential to be mapped to a GO term describing the function. However, mapping of individual domain entries to GO terms does provide a means of assigning multiple GO terms to multifunctional proteins. 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

browser.[19] In this way, all proteins belonging to InterPro entries mapped to GO terms can be automatically mapped to these GO terms.

Database cross-references

Cross-references to other databases

In addition to cross-referencing the member database signatures and GO terms, there is a separate field in InterPro entries, 'Database Links', to provide cross-references to other databases. Included in this field are cross-references to corresponding Blocks accession numbers; PROSITE documentation; and the Enzyme Commission (EC) Database where the EC number(s) for proteins matching the entry are common. Where applicable, there may also be links to specialised websites for example the Carbohydrate-Active EnZymes (CAZy) site, which describes families of related catalytic and carbohydrate-binding modules of enzymes that act on glycosidic bonds.

InterProScan – sequence search tool

The sequence search package InterProScan[14] combines the search methods from each of the databases into a single package and provides an output with all results in a single format, which may be HTML, text or XML (extensible mark-up language). The sequence-based search uses tools provided by the member databases, including ScanRegExp for PROSITE patterns, pfscan for PROSITE profiles,[3] hmmpfam for Pfam,[5] SMART and TIGRFAMs HMMs, fingerPRINTScan for PRINTS fingerprints[20] and BlastProDom for ProDom patterns.[9] InterProScan is more than a simple wrapping of sequence analysis applications since it requires performing considerable data look-ups from some databases and program outputs. The threshold scores for profiles or HMMs are provided by the member databases and are considered to be trustworthy for displaying only true hits. Some post-processing of data is linked to the software package; for example, the Pfam, SMART and TIGRFAMs outputs are filtered through family-specific thresholds for increased accuracy of the results.

Output HTML, text or XML

The results from a sequence search through InterProScan 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 illustrated in Figure 2. Detailed results of matches to the individual database search methods are provided via hyperlinks to each of the parent databases. For all methods except Prosite patterns the results displayed are considered to be true since they have been generated using the hand-curated thresholds of the member databases and post-processing of data. However, it will soon be possible to view the scores of the results so that users can determine for themselves the significance of the hit. Prosite pattern hits are usually reported as unknown (?), but can then be verified as true or false positive by the user. In InterPro there are some protein matches that have been changed to false positive hits by curators; however, not including Prosite patterns, there are only 4,768 false positive out of 2.5 million true hits. A mail server is available for sequence searches.[21] Documentation on using the mail server can be obtained by emailing the address with the word 'help' in the body of the text. In this way, independent researchers may submit their sequences using a Web interface and obtain results of hits in InterPro in both a graphical and tabular view. Groups requiring confidentiality or bulk sequence searches may download a Perl stand-alone InterProScan package that can be run locally. This version also has the option to run the TMHMM (transmembrane prediction)[22] and SignalP (signal peptide prediction)[23] software as plug-ins. It supports all this software but the packages are not free and so are not distributed. It is also possible to link to GO terms which the package retrieves from a file of InterPro to GO mappings.

Submit sequences using web interfaces

Confidentiality or bulk sequence searches

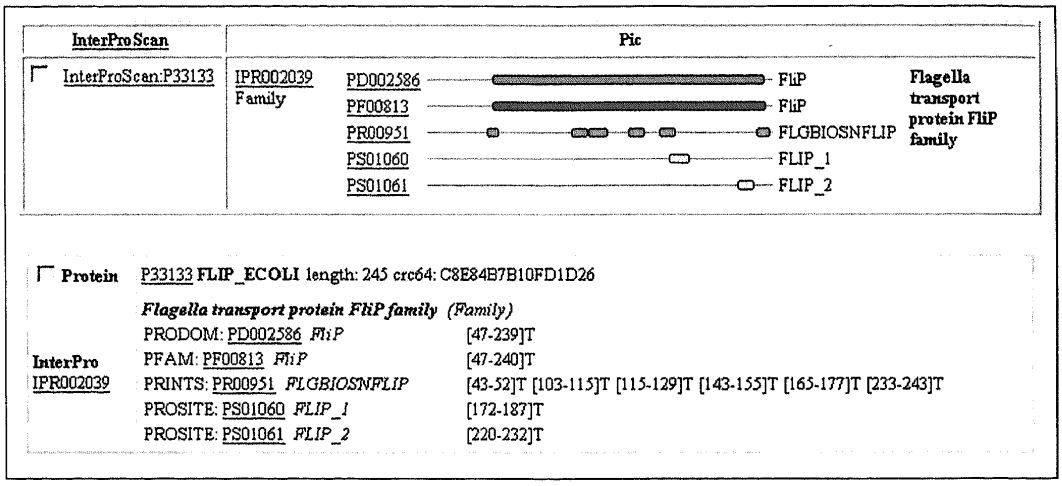


Figure 2: InterProScan output for the *Escherichia coli* flagellar transport protein P33133. The results are provided in a tabular or graphical view. In the latter, the signatures are colour coded, and the widths of the coloured bands represent the boundaries of the signatures. The codes on the left 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. The table provides information on the positions of each match on the sequence

Database access and format

Synchronisation of data is an issue

Interactive use via web server

The InterPro database is implemented in an Oracle relational database, and is accessible for interactive use via the EBI web server,[11] which can also be reached via each of the member databases. All data in InterPro is freely accessible and distributable with the InterPro Copyright agreement. The Web interface allows a simple text-based search accessing the database directly, and text- and and sequence-based searches using SRS.[24]

The InterPro entries are also released in two XML flatfiles, 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. The InterPro data in SRS is based on these XML files. The InterPro flatfiles may be retrieved from the EBI anonymous-ftp server.[25]

Additional files

Additional files available from the Web and ftp servers are: a list of all protein matches; a list of all InterPro entries and their names; and a file of InterPro to GO mappings. There is also documentation available for the database, including release notes and a user manual.

UPDATING INTERPRO

Since InterPro is an integrated resource of up to eight different databases, synchronisation of data is an issue. Good communication and data flow between member databases are required. New member database signatures are integrated into InterPro shortly before member databases produce new releases since the integration process and annotation of new entries takes time. New member database releases are received as flatfiles, all new, changed or deleted signatures are identified and InterPro is updated accordingly. The new and changed signatures are then run over the whole of SWISS-PROT and TrEMBL to compute the match lists. The protein matches in InterPro are updated weekly, triggered by the production of the weekly SPTR (SWISS-PROT and TrEMBL) release. Annotation and documentation updates are ongoing and released at regular intervals. At InterPro release time new XML files are dumped from the database and validated. At the same time all available files from the Web and ftp servers are updated. The future aim is to have more regular releases of the XML files to keep the data in the database,

which serves the Web site, and the XML file, which is indexed in SRS, more in synch.

DISCUSSION

Capitalising on strengths of member databases

The InterPro project has succeeded in capitalising on the strengths of the member databases to produce an integrated resource with benefits not only for individual researchers or genome projects, but also for the member databases themselves. Integration into InterPro reduces duplication of effort in the labour-intensive process of annotation; serves as a quality control mechanism for assessing individual methods; and also highlights the areas where all the member databases are lacking in representation. This is supplemented by the increasing availability of complete genome sequences, which identifies uncharacterised protein families that may be expressed in a single organism or comprise orthologues in a number of different species. Another major use of InterPro is in identifying those families and domains for which the existing discriminators are not optimal and could hence be usefully supplemented with an alternative pattern (eg 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 superfamily it could be beneficial to develop discrete family fingerprints, and so on). The resource acts as a convenient means of deriving a consensus among signature methods particularly when one domain or family is diagnosed by signatures from many of the member databases.

Quality control mechanism

Complete genome sequences

Functional classification of newly determined sequences

A primary application of InterPro's family, domain and functional site definitions is in the computational functional classification of newly determined sequences that lack biochemical characterisation. The EBI is using InterPro for enhancing the automated annotation of TrEMBL.[26] This process is more efficient and reliable than using each of the pattern databases separately, because InterPro provides internal consistency checks and deeper coverage. One limitation is that not all InterPro entries represent signatures diagnostic of known functions. There are approximately 350 out of over 4,500 entries that describe 'proteins of unknown function', and several more that hit a set of hypothetical or uncharacterised proteins. Nevertheless, they still function in grouping related proteins, so that if the function of one of the proteins is elucidated it may shed some light on the function of the related proteins. InterPro has become a major resource for the annotation of newly sequenced genomes. The database and InterProScan software package have been used for: the comparative genome analysis of *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*,[27] comparative analysis of malaria genomes,[28] the study of fish genomes,[29] initial annotation of the human genome[30] and analysis of the mouse cDNAs[31] and the rice (*Oryza sativa*) genome,[32,33] exemplifying the utility of the resource in analysis and comparison of complete genomes.

FUTURE DIRECTIONS

While the initial InterPro release was created around PRINTS, PROSITE and Pfam, since then ProDom, SMART and TIGRFAMs have been included in stages. The next protein family database to be integrated into InterPro will be the Protein Information Resource (PIR) superfamily database.[34] PIR superfamilies facilitate protein family information retrieval, identification of domain and family relationships and classification of multi-domain proteins. However, the major future objectives are to broaden the scope of InterPro into the area of protein secondary and tertiary structure. Protein structure information has become vital in understanding protein function and evolutionary relationships. Integration of such information into InterPro will enhance the capability of the database in the field of protein classification and characterisation and make the database a

Integration of structural information

true integrated resource for complete protein sequence and structure information. A project has been initiated to integrate the data of SCOP (Structural Classification of Proteins),[35] CATH (Class, Architecture, Topology, Homology)[36] and SWISS-MODEL 3D structure homology models[37] into InterPro. The project will include the development of improved visualisation tools for better views of the integrated data. As InterPro continues to grow in size, scope and strength, so the utility of the data will be extended to more and more users from different fields of biological research.

Acknowledgements
The InterPro project is supported by the ProFuSe grant (number QLG2-CT-2000-00517) of the European Commission.

References

1. Ponting, C. P. (2001), 'Issues in predicting protein function from sequence', *Brief. Bioinform.*, Vol. 2(1), pp. 19–29.
2. Gribskov, M., Luthy, R. and Eisenberg, D. (1990), 'Profile analysis', *Methods Enzymol.*, Vol. 183, pp. 146–159.
3. Falquet, L., Pagni, M., Bucher, P. *et al.* (2002), 'The PROSITE database, its status in 2002', *Nucleic Acids Res.*, Vol. 30, pp. 235–238.
4. Attwood, T. K., Blythe, M. J., Flower, D. R. *et al.* (2002), 'PRINTS and PRINTS-S shed light on protein ancestry', *Nucleic Acids Res.*, Vol. 30, pp. 239–241.
5. Bateman, A., Birney, E., Cerruti, L. *et al.* (2002), 'The Pfam Protein Families Database', *Nucleic Acids Res.*, Vol. 30, pp. 276–280.
6. Letunic, I., Goodstadt, L., Dickens, N. J. *et al.* (2002), 'Recent improvements to the SMART domain-based sequence annotation resource', *Nucleic Acids Res.*, Vol. 30, pp. 242–244.
7. Haft, D. H., Loftus, B. J., Richardson, D. L. *et al.* (2001), 'TIGRFAMs: a protein family resource for the functional identification of proteins', *Nucleic Acids Res.*, Vol. 29(1), pp. 41–43.
8. Henikoff, J. G., Pietrokovski, S., McCallum, C. M. and Henikoff, S. (2000), 'Blocks-based methods for detecting protein homology', *Electrophoresis*, Vol. 21(9), pp. 1700–1706.
9. 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.*, Vol. 28, pp. 267–269.
10. Apweiler, R., Attwood, T. K., Bairoch, A. *et al.* (2001), 'The InterPro database, an integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res.*, Vol. 29(1), pp. 37–40.
11. URL: http://www.ebi.ac.uk/interpro/ (queries may be emailed to interhelp@ebi.ac.uk).
12. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 45–48.
13. 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*, Vol. 95, pp. 5865–5871.
14. Zdobnov, E. M. and Apweiler, R. (2001), 'InterProScan – an integration platform for the signature-recognition methods in InterPro', *Bioinformatics*, Vol. 17(9), pp. 847–848.
15. Clamp, M. E., Cuff, J. A. and Barton, G. J. (1998), 'Jalview – a java multiple alignment editor' (URL: http://www.ebi.ac.uk/~michele/jalview/).
16. Corpet, F., Gouzy, J. and Kahn, D. (1999), 'Browsing protein families via the "Rich Family Description" format', *Bioinformatics*, Vol. 15, pp. 1020–1027.
17. URL: http://www.ebi.ac.uk/sp/proteinprofiles/
18. The Gene Ontology Consortium (2001), 'Creating the gene ontology resource: Design and implementation', *Genome Res.*, Vol. 11, pp. 1425–1433.
19. URL: http://www.ebi.ac.uk/ego/
20. Scordis, P., Flower, D. R. and Attwood, T. K. (1999), 'FingerPRINTScan: Intelligent searching of the PRINTS motif database', *Bioinformatics*, Vol. 15(10), pp. 799–806.
21. Mail to: Interproscan@ebi.ac.uk
22. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes', *J. Mol. Biol.*, Vol. 305(3), pp. 567–580.
23. Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997), 'A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites', *Int. J. Neural Syst.*, Vol. 8(5–6), pp. 581–599.
24. Etzold, T., Ulyanov, A. and Argos, P. (1996), 'SRS: information retrieval system for molecular biology data banks', *Methods Enzymol.*, Vol. 266, pp. 114–128.

25. URL: ftp://ftp.ebi.ac.uk/pub/databases/interpro
26. Fleischmann, W., Möller, S., Gateau, A. and Apweiler R. (1999), 'A novel method for automatic functional annotation of proteins', *Bioinformatics*, Vol. 15, pp. 228–233.
27. Rubin, G. M., Yandell, M. D., Wortman, J. R. *et al.* (2000), 'Comparative genomics of the eukaryotes', *Science*, Vol. 287, pp. 2204–2215.
28. Carlton, J. M., Muller, R., Yowell, C. A. *et al.* (2001), 'Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species', *Mol. Biochem. Parasitol.*, Vol. 118(2), pp. 201–220.
29. Biswas, M., Kanapin, A. and Apweiler, R. (2001), 'Application of InterPro for the functional classification of the proteins of fish origin in SWISS-PROT and TrEMBL', *J. Biosci.*, Vol. 26(2), pp. 277–284.
30. The International Human Genome Consortium (2001), 'Initial sequencing and analysis of the human genome', *Nature*, Vol. 409(6822), pp. 860–921.
31. Kawaji, H., Schonbach, C., Matsuo, Y. *et al.* (2002), 'Exploration of novel motifs derived from mouse cDNA sequences', *Genome Res.*, Vol. 12(3), pp. 367–378.
32. Yu, J., Hu, S., Wang, J. *et al.* (2002), 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)', *Science*, Vol. 296(5565), pp. 79–92.
33. Goff, S. A., Ricke, D., Lan, T. H. *et al.* (2002), 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)', *Science*, Vol. 296(5565), pp. 92–100.
34. Wu, C. H., Xiao, C., Hou, Z. *et al.* (2001), 'iProClass: An integrated, comprehensive and annotated protein classification database', *Nucleic Acids Res.*, Vol. 29(1), pp. 52–54.
35. Lo Conte, L., Brenner, S. E., Hubbard, T. J. *et al.* (2002), 'SCOP database in 2002: Refinements accommodate structural genomics', *Nucleic Acids Res.*, Vol. 30(1), pp. 264–267.
36. Pearl, F. M, Lee, D., Bray, J. E. *et al.*(2002), 'The CATH extended protein-family database: Providing structural annotations for genome sequences', *Protein Sci.*, Vol. 11(2), pp. 233–244.
37. Guex, N. and Peitsch, M. C. (1997), 'SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling', *Electrophoresis*, Vol. 18(15), pp. 2714–2723.