Tutorial

Tools and resources for identifying protein families, domains and motifs

Nicola J Mulder and Rolf Apweiler

Address: The EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Correspondence: Rolf Apweiler. E-mail: apweiler@ebi.ac.uk

Published: 19 December 2001

Genome ***Biology*** 2001, **3(1)**:reviews2001.1–2001.8

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/3/1/reviews/2001

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

With the large influx of raw sequence data from genome sequencing projects, there is a need for reliable automatic methods for protein sequence analysis and classification. The most useful tools use various methods for identifying motifs or domains found in previously characterized protein families. This article reviews the tools and resources available on the web for identifying signatures within proteins and discusses how they may be used in the analysis of new or unknown protein sequences.

What is the problem to be solved?

In June 2000 the first draft of the human sequence was announced, and was considered to be an achievement equal to that of putting the first man on the moon. The announcement brought promises of breakthroughs in treating human diseases, but in fact all it meant was a flood of data to be converted into useful biological information. To live up to the promise of the sequence, the first obstacles are to classify the genes it contains and to assign functions to the gene products [1]. Protein sequences can be classified by identifying the protein type, but they then need to be characterized further, to assign biological function. The challenge is in this application of useful biological knowledge to particular protein sequences.

There are several reasons to choose to characterize proteins rather than DNA sequences. These include: the larger alphabet (21 amino acids versus 4 bases); the lower signal-to-noise ratio in protein sequence searches; the closeness between protein sequence and function; and the availability of good, well annotated databases of protein sequences and protein sequence signatures. Proteins can be characterized at different levels: they perform a function in a cell, but this function is also performed within a particular context, for example as part of a complex pathway, as well as at a defined cellular location. At the functional level, this may come down to analysis of the protein sequence along its entire length, at the level of single domains or motifs, or at the finest level, single important amino-acid residues. With the increased availability of completely sequenced genomes, and using the correct tools and resources, there is scope for protein characterization on all these levels.

The first step in the analysis of new or uncharacterized protein sequences is traditionally to search the protein databases for similar sequences. The main protein sequence databases available are SWISS-PROT and TrEMBL [2,3], the Protein Information Resource (PIR) [4,5], and GenPept, which is a translation of GenBank [6,7]. If the similarity to proteins in a database is significant, information from the proteins in the search results can be inferred to apply also to the query sequence. This relies on the quality of the annotation in the protein sequence databases, and, more generally, on the availability of experimental results in the scientific literature. But problems arise during sequence-similarity searches when more than one domain is present in a protein [1]. A large number of matches to one domain in a sequence may mask 'hits' that match a second domain in the sequence, and thus useful information is lost. It is also possible for sequences to be evolutionarily related but for their sequences to have diverged to such an extent that they are not picked up in a sequence-similarity search. And, with the increase in population of protein sequence databases, the number of related sequences rises, so when a search is performed it identifies a large set of highly related sequences

and the less related sequence hits may be lost. It is for these reasons that protein signature databases evolved and have become increasingly useful tools for protein sequence analysis; they aim to identify domains, or classify proteins into families, and thereby infer function. A signature refers to the diagnostic entity used to recognize a domain or family; it may be derived using a number of methods, including patterns and profiles (discussed below). This article presents the main signature databases available for protein sequence analysis, their methods, and their individual uses.

How is it done?

The basic information about a protein comes from its sequence. From a single sequence it is difficult to infer anything about the protein, but as the number of related sequences increases, so an alignment can be built to create a consensus for a protein family, or to identify conserved domains or highly conserved residues that may be important for function, for example in an active site. These conserved areas of a protein family, domain or functional site can be used to define identifiable features using several different methods. These include building up regular expressions to show patterns of conserved amino-acid residues ('pattern' is used here to mean a precise, contiguous stretch of sequence); producing detailed profiles from sequence alignments; and hidden Markov models (HMMs), which are profiles derived using a more complex probabilistic scoring mechanism. A profile is built from a sequence alignment, and describes the probability of finding an amino acid at a given position in the sequence; the profile constitutes a table, or matrix, of position-specific amino-acid weights and gap costs [8]. The numbers in the table (scores) are used to calculate similarity scores between a profile and a sequence within a given alignment. A threshold score is calculated for each set of sequences, so that only sequences scoring above this threshold are considered to be related to the original set of sequences in the alignment. 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 they fail to provide information about the rest of the sequence, and because of the constraints on which amino acids may be found within a given area of the sequence, patterns fail to pick up related sequences that have even a small divergence in that particular area. Profiles and HMMs compensate for these problems in that they generally cover larger areas of the sequence, and because 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 (the term 'hit list' in this article refers to the list of proteins that match or contain a particular signature above the required score).

There are a number of well known signature databases in the public domain that use these methods to produce diagnostic signatures for protein families, domains, repeats, active sites, binding sites and post-translational modifications. These include PROSITE [9,10], PRINTS-S [11,12], Pfam [13,14], SMART [15,16], TIGRFAMs [17,18] and Blocks [19,20]. There are also several databases that identify protein families or domains using sequence clustering and alignment methods; these include ProDom [21,22], DOMO [23-25], PIR-ALN [26,27], ProClass [28,29], ProtoMap [30,31], SYSTERS [32,33] and CluSTr [34,35]. All these databases are useful tools for protein sequence analysis, and some are discussed in more detail here. A list of useful protein, pattern and integration databases and their URLs is shown in Table 1.

What is available?

PROSITE patterns and profiles

PROSITE [9,10] is a database of both patterns and profiles. PROSITE patterns are built from alignments of related sequences, which are taken from a variety of sources: from a well-characterized protein family; derived from the literature; from the results of sequence searches against SWISS-PROT and TrEMBL; or from sequence clustering. The alignments are checked for conserved regions, which, particularly for the characterized protein families, may have been experimentally shown to be involved in the catalytic activity or to bind a substrate. A core pattern is created in the form of a regular expression that specifies which amino acid(s) may or may not occur at each position. Regular expressions are text strings that describe patterns, used to represent a set of strings. They can be seen as similar to wildcard pattern-matching tools used traditionally under Unix and Unix-like operating system utilities. Regular expressions are much more elaborate and powerful than standard wildcard expressions, but they are also much more complex. Once the core pattern is made, it is tested against the sequences in SWISS-PROT. If the correct set of proteins matches this pattern then it is kept; if it fails to pick up some family members or picks up too many unrelated proteins, the pattern is refined and re-tested until it is optimized.

Patterns have many advantages, but they also have their limitations across whole sequences, which is why PROSITE also creates profiles [36], to complement the patterns. For these, the process also starts with multiple sequence alignments; it then uses a symbol comparison table to convert residue frequency distributions into weights, resulting in a table of position-specific weights [8]. A symbol comparison table comprises values describing the comparison between pairs of amino acids. The table has a value for the match quality of every possible pair of amino acids, and is used to provide scores for the probability of one amino acid being replaced by another at a particular position within the sequence alignment. These numbers are used to calculate a similarity score for the alignment between the profile and sequences in SWISS-PROT; an alignment with a similarity score equal to

Table 1

Useful tools and resources for protein family, domain and motif analysis

Database	Description	URL	Published reference
Blocks	Database of protein alignment blocks	http://blocks.fhcrc.org	[19]
CDD	Conserved domain database	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	[47]
CluSTr	Clusters of SWISS-PROT and TrEMBL proteins	http://www.ebi.ac.uk/clustr/	[34]
DOMO	Protein-domain database based on sequence alignments	http://www.infobiogen.fr/services/domo/	[24]
InterPro	Integrated documentation resource for protein families, domains and functional sites	http://www.ebi.ac.uk/interpro/	[49]
IProClass	Integrated protein classification database	http://pir.georgetown.edu/iproclass/	[45]
MetaFam	Database of protein family information	http://metafam.ahc.umn.edu/	[43]
Pfam	Collection of multiple sequence alignments and hidden Markov models	http://www.sanger.ac.uk/Software/Pfam/	[13]
PIR	Protein Information Resource	http://pir.georgetown.edu/	[4]
PIR-ALN	Curated database of protein sequence alignments	http://pir.georgetown.edu/pirwww/dbinfo/piraln.html	[26]
PRINTS-S	Compendium of protein fingerprints	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	[11]
ProClass	Non-redundant protein database organized by family relationships	http://pir.georgetown.edu/gfserver/proclass.html	[28]
ProDom	Automatic compilation of homologous domains	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html	[21]
PROSITE	Database of patterns and profiles describing protein families and domains	http://www.expasy.ch/prosite/	[9]
ProtoMap	Automatic hierarchical classification of SWISS-PROT proteins	http://www.protomap.cs.huji.ac.il/	[30]
SBASE	Curated protein domain library based on sequence clustering	http://www3.icgeb.trieste.it/~sbasesrv/	[51]
SMART	Simple Modular Architecture Research Tool – a collection of protein families and domains	http://smart.embl-heidelberg.de/	[15]
SWISS-PROT and TrEMBL	Protein sequence databases	http://www.ebi.ac.uk/swissprot/ or http://www.expasy.org/sprot/	[2]
SYSTERS	Systematic re-searching method for sequence searching and clustering	http://systers.molgen.mpg.de/	[32]
TIGRFAMs	Protein families based on hidden Markov models	http://www.tigr.org/TIGRFAMs/	[17]

or greater than a given cut-off value constitutes a true hit. The profile is then refined until only the intended set of protein sequences scores above the threshold for the profile.

Pfam, SMART and TIGRFAMs HMMs

Many databases, such as Pfam, SMART and TIGRFAMs, use HMMs as a way of creating diagnostic signatures for protein families, domains and repeats. The HMMs are built [37] from manually curated sequence alignments using the HMMER2 package [38], which is based on Bayesian statistical models. Pfam [13,14] is a collection of multiple protein-sequence alignments and HMMs, and provides a good repository of models for identifying protein families, domains and repeats. There are two parts to the Pfam database: PfamA, a set of manually curated and annotated models; and PfamB, which has higher coverage but is fully automated (with no manual curation). PfamB HMMs are created from alignments generated by ProDom [21,22] in their automatic clustering of the protein sequences in SWISS-PROT and TrEMBL.

The SMART database ('simple modular architecture research tool') [15,16] produces HMMs that facilitate the identification and annotation of genetically mobile domains and the analysis of domain architectures. The database is highly populated with models for domains found in signaling, extracellular and chromatin-associated proteins. The models rely on hand-curated multiple sequence alignments of representative family members, based on tertiary structures where possible but otherwise found by PSI-BLAST [39]. Once the models are created, they are used to search the database for additional members to be included in the sequence alignment. This iterative process is repeated until no further homologs are detected. TIGRFAMs [17,18] creates HMMs that group homologous proteins that are conserved

with respect to function. The models are produced in a similar way to those in Pfam and SMART, but should only hit equivalogs, proteins that have been shown to have the same function.

FingerPRINTS

The PRINTS-S database [11,12] uses 'fingerprints' as diagnostic signatures, in a variation on the methods described above. A fingerprint is a group of conserved motifs used to characterize a protein family. Rather than focusing solely on small conserved areas, the occurrence of these conserved areas across the whole sequence is taken into account. Once again the starting point is a curated multiple sequence alignment. Profiles are built for small conserved regions in the sequence, and together these make up a fingerprint. The 'fingers', or motifs, are required to be present in the sequence in the correct order for the fingerprint to be counted as a match in a target sequence. During the creation of fingerprints each motif is used to scan the protein sequence database and the resulting hit lists are correlated, to add sequences to the original alignment. New motifs are then generated and the process is repeated until convergence. Recognition of individual elements in the fingerprint is mutually conditional, and true members match all elements in the correct order, while members of a subfamily may match only part of the fingerprint. Many fingerprints have been created to identify proteins at the superfamily as well as the family and subfamily levels; for this reason, many of the fingerprints are related to each other in an ordered hierarchical structure.

Clustering and alignment

An example of a database that solely uses sequence clustering and alignment methods is ProDom [21,22]. This database takes all proteins in the SWISS-PROT and TrEMBL protein databases, removes fragments, identifies the smallest remaining sequence and uses this as a query sequence to search the SWISS-PROT/TrEMBL protein database using PSI-BLAST [39]. The hit-list sequences are made into a new ProDom domain family and removed from the protein database. The remaining sequences are once again sorted by size, and the smallest sequence is again used as a query sequence. This process is repeated until there are no more sequences in the protein database [40]. In this way, ProDom groups all the non-fragment sequences in SWISS-PROT and TrEMBL into more than 150,000 families. Other major alignment databases are: PIR-ALN [26,27], which is a database of annotated protein sequence alignments derived automatically from the PIR sequence database and has alignments at the superfamily and domain levels; and ProtoMap [30,31], an automatic classification of all SWISS-PROT and TrEMBL proteins into groups of related proteins based on pair-wise similarities.

Another protein signature database worth a mention is Blocks [19,20], a collection of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. These alignments are represented as profiles, built up using a tool called PROTOMAT [41]. The profiles are calibrated against the SWISS-PROT database, and the LAMA software tool [42] is used to search new blocks against existing blocks. Blocks+ is a new, extended version of the original Blocks database [20].

Which one(s) should you use?

Many of the databases mentioned here have interactions and exchange data. PfamB, for example, uses ProDom families as a starting point for its automatically built HMMs. Blocks previously used alignments from PROSITE, PRINTS, PfamA and ProDom as a starting point for the creation of the Blocks database, and now uses InterPro, an integration of these four databases (described below). All of the diagnostic protein signature databases have their strengths and weaknesses and are useful for individual researchers as well as large-scale genome sequencing projects. But which ones should be used, and how can the results from the different databases, which all have their own formats and outputs, be integrated? A solution has been provided by several groups who have made an effort to integrate their databases into a single, coherent protein-signature resource. These integrated database resources include MetaFam [43,44], iProClass [45,46], CDD [47,48] and InterPro [49,50]. MetaFam automatically creates supersets of overlapping families from the Blocks+, DOMO, Pfam, PIR-ALN, PRINTS, PROSITE, ProDom and SBASE [51,52] databases, using set theory to compare the databases with one another (Figure 1). MetaFam provides reference domains covering all of the domains represented in all the protein databases. The iProClass database links ProClass, PIR-ALN, PROSITE, Pfam and Blocks into a single database; and CDD is a database of domains derived from SMART, Pfam and contributions from NCBI LOAD (library of ancient domains). These integrated resources all use automatic methods, and in some cases simply allow the user to search the individual databases from one page, whereas InterPro is produced by the curators of the individual databases, and the component databases (at present PROSITE, PRINTS, Pfam, ProDom, SMART and TIGRFAMs) are manually curated and integrated into a single comprehensive format.

InterPro [49,50] consists of nearly 5,000 entries with unique accession numbers and names, each describing a different protein family, domain, repeat or post-translational modification. Each entry (see, for example, Figure 2a) includes one or more signatures from the individual member databases that describe the relevant group of proteins. For example, all PRINTS fingerprints, PROSITE patterns and profiles, ProDom domains, and Pfam, SMART and TIGRFAMs HMMs that provide diagnostic methods for identifying the same domain within protein sequences are grouped together in a single InterPro entry. Each entry also includes an abstract, which provides annotation about the proteins matching the

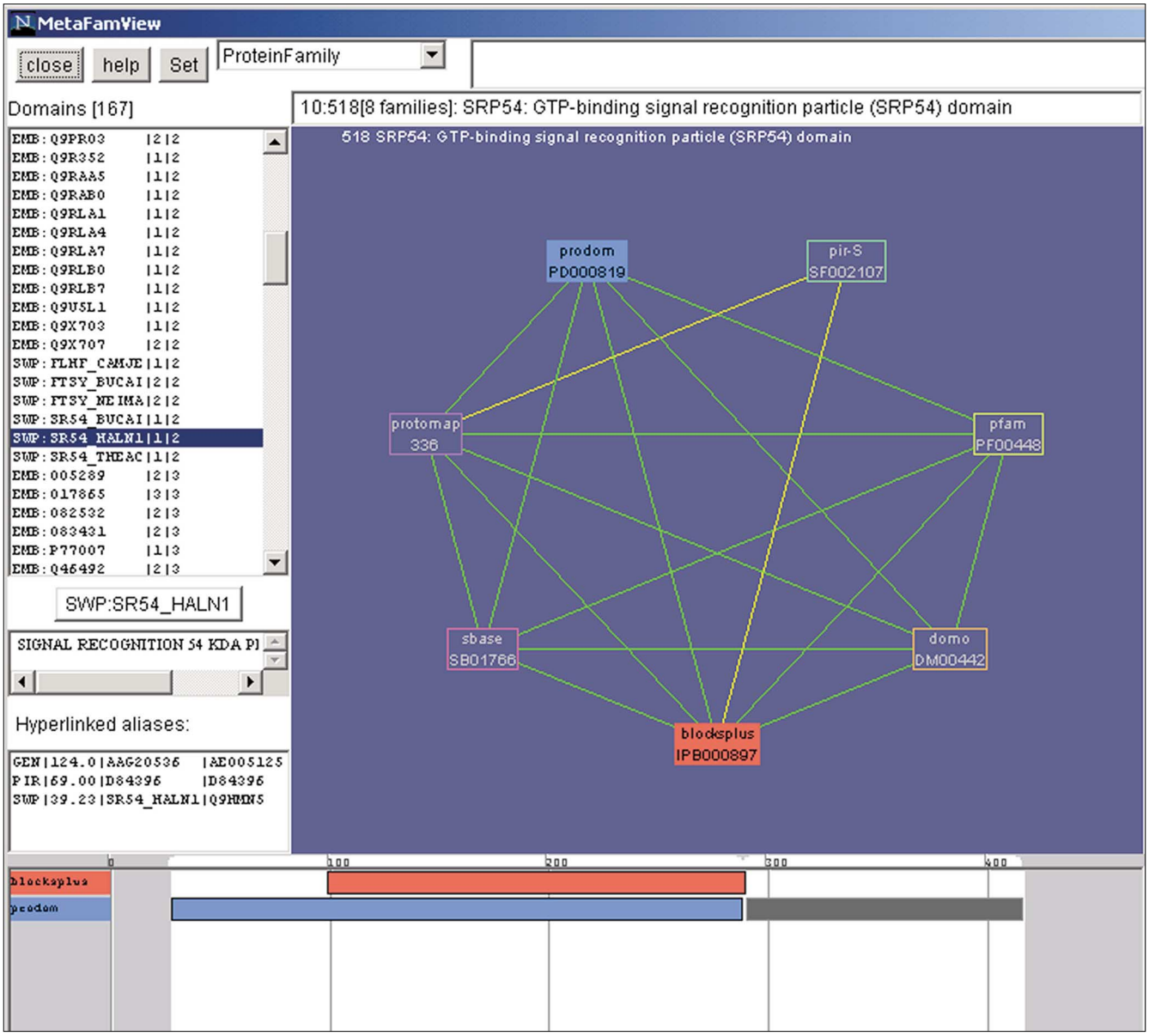

Figure 1
An example of a MetaFam family entry [43,44]. This is the entry for SRP54, the GTP-binding signal recognition particle domain, and shows the links between related entries in ProDom, PIR superfamilies, Pfam, DOMO, Blocks+, SBASE and ProtoMap. The domain structure for the selected SWISS-PROT protein SR54_HALN1 is shown at the bottom of the entry.

entry, and a list of pre-computed matches against the whole of the SWISS-PROT and TrEMBL databases; the match lists may be viewed in a tabular form, with lists of the protein accession numbers and the positions within the amino-acid sequences that each signature from that InterPro entry hits. The match list can also be viewed graphically, with the sequence split into several lines, one for each hit by a unique signature; this view includes the hits by all signatures from the same and other InterPro entries. The proteins can also be viewed graphically in a condensed view that computes the consensus domain boundaries from all signatures within each entry, and splits the protein sequence into different lines for each InterPro entry matched. From this view, all proteins sharing a common domain architecture can be grouped and the sequences aligned using Jalview [53] or DisplayFam [54]. InterPro is implemented in an Oracle relational database, and is accessible using text or sequence searches. The sequence-search package, InterProScan [55], 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 simple text or web-friendly HTML (see Figure 2b) or structured XML. In this way, independent researchers can

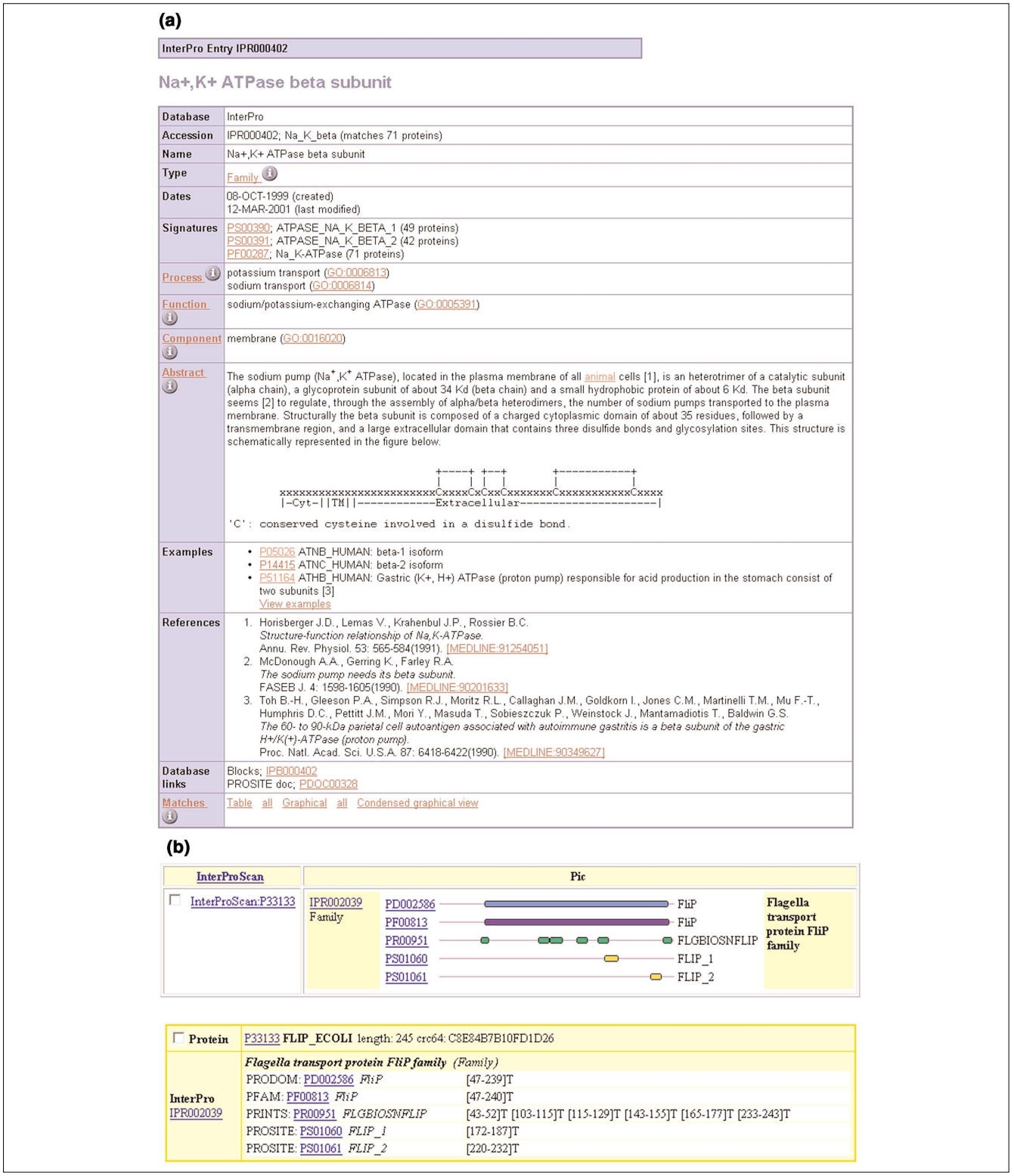

Figure 2

(a) An example of an InterPro entry [49,50]. This is entry IPR000402, which describes the Na^+, K^+ ATPase β subunit. The entry integrates two PROSITE patterns and a Pfam HMM, which are diagnostic for the same family. It includes an abstract describing this family, and match lists of all SWISS-PROT and TrEMBL proteins that belong to the family. **(b)** The output from the sequence search result of InterProScan. The *Escherichia coli* flagellar biosynthetic protein FliP was scanned, and shown to hit InterPro entry IPR002039, which describes the FliP protein family. The results are shown in both a graphical and a tabular view, and include amino-acid positions for the signature matches.

submit their sequences using a web interface and obtain results of hits in InterPro in both a graphical and a tabular view. Groups requiring confidentiality or bulk sequence searches can download a Perl stand-alone InterProScan package that can be run locally. The results give an indication of the family an unknown protein belongs to, and its domain composition; for more information the user can read about related proteins or the protein family in the annotation of the InterPro entries it hits.

The benefits of integration

The integrated resources for protein family and domain signature databases, such as InterPro, MetaFam and CDD, have several uses, not only for the scientific community, for whom they build on the individual strengths of the different methods, but also for the member databases themselves. The integration reduces duplication of effort for the member databases in the labor-intensive, rate-limiting process of annotation, and also facilitates communication between the disparate resources. The integrated resources provide quality control mechanisms for assessing individual methods, and also highlight the areas where all the member databases are lacking in representation. This situation is improved by the increasing availability of complete genome sequences, which help to identify uncharacterized protein families that may be unique to single or groups of related organisms.

It is evident that there are currently a large number of high-quality protein signature databases and integrated databases available for automatic and large-scale protein classification. The challenge remains, however, in the transfer of useful biological knowledge to protein sequences. Automatic methods may provide some useful suggestions of protein architecture or function, but only a biologist can truly assign function to a protein, using these results, and the ultimate confirmation of these assignments must remain experimental evidence.

References

1. Ponting CP: **Issues in predicting protein function from sequence.** *Brief Bioinform* 2001, **2:**19-29.
2. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
3. **SWISS-PROT and TrEMBL** [http://www.expasy.ch/sprot/sprot-top.html]
4. Barker WC, Garavelli JS, Hou Z, Huang H, Ledley RS, McGarvey PB, Mewes HW, Orcutt BC, Pfeiffer F, Tsugita A, *et al.*: **Protein Information Resource: a community resource for expert annotation of protein data.** *Nucleic Acids Res* 2001, **29:**29-32.
5. **Protein Information Resource** [http://pir.georgetown.edu/]
6. Burks C, Cassidy M, Cinkosky MJ, Cumella KE, Gilna P, Hayden JE, Keen GM, Kelley TA, Kelly M, Kristofferson D, *et al.*: **GenBank.** *Nucleic Acids Res* 1991, **Suppl 19:**2221-2225.
7. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
8. Gribskov M, Luthy R, Eisenberg D: **Profile analysis.** *Methods Enzymol* 1990, **183:**146-159.
9. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27:**215-219.
10. **PROSITE** [http://www.expasy.ch/prosite/]
11. Attwood TK, Croning MDR, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Res* 2000, **28:**225-227.
12. **PRINTS-S** [http://bioinf.man.ac.uk/dbbrowser/sprint/printss_lis.html]
13. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL: **The Pfam Protein Families Database.** *Nucleic Acids Res* 2000, **28:**263-266.
14. **Pfam** [http://www.sanger.ac.uk/Software/Pfam/]
15. Ponting CP, Schultz J, Milpetz F, Bork P: **SMART: identification and annotation of domains from signalling and extracellular protein sequences.** *Nucleic Acids Res* 1999, **27:**229-232.
16. **SMART** [http://smart.embl-heidelberg.de/]
17. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Res* 2001, **29:**41-43.
18. **TIGRFAMs** [http://www.tigr.org/TIGRFAMs/]
19. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S: **Increased coverage of protein families with the blocks database servers.** *Nucleic Acids Res* 2000, **28:**228-230.
20. **Blocks** [http://blocks.fhcrc.org/]
21. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28:**267-269.
22. **ProDom** [http://prodes.toulouse.inra.fr/prodom/doc/prodom.html]
23. Gracy J, Argos P: **Automated protein database classification: I. Integration of compositional similarity search, local similarity search and multiple sequence alignment.** *Bioinformatics* 1998, **14:**164-173.
24. Gracy J, Argos P: **Automated protein database classification: II. Delineation of domain boundaries from sequence similarities.** *Bioinformatics* 1998, **14:**174-187.
25. **DOMO** [http://www.infobiogen.fr/services/domo/]
26. Srinivasarao GY, Yeh LS, Marzec CR, Orcutt BC, Barker WC: **PIR-ALN: a database of protein sequence alignments.** *Bioinformatics* 1999, **15:**382-390.
27. **PIR-ALN** [http://www-nbrf.georgetown.edu/pirwww/dbinfo/piraln.html]
28. Huang H, Xiao C, Wu CH: **ProClass protein family database.** *Nucleic Acids Res* 2000, **28:**273-276.
29. **ProClass** [http://pir.georgetown.edu/gfserver/proclass.html]
30. Yona G, Linial N, Linial M**: ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucleic Acids Res* 2000, **28:**49-55.
31. **ProtoMap** [http://www.protomap.cs.huji.ac.il/]
32. Krause A, Stoye J, Vingron M: **The SYSTERS protein sequence cluster set.** *Nucleic Acids Res* 2000, **28:**270-272.
33. **SYSTERS** [http://systers.molgen.mpg.de/]
34. Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R: **CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins.** *Nucleic Acids Res* 2001, **29:**33-36.
35. **CluSTr** [http://www.ebi.ac.uk/clustr/]
36. Bucher P, Karplus K, Moeri N, Hofmann K: **A flexible motif search technique based on generalized profiles.** *Comput Chem* 1996, **20:**3-23.
37. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK: Cambridge University Press, 1998.
38. **HMMER2: Profile hidden Markov models for biological sequence analysis** [http://hmmer.wustl.edu/]
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
40. Gouzy J, Corpet F, Kahn D: **Whole genome protein domain analysis using a new method for domain clustering.** *Comput Chem* 1999, **23:**333-340.
41. Henikoff S, Henikoff JG: **Automated assembly of protein blocks for database searching.** *Nucleic Acids Res* 1991, **19:**6565-6572.
42. Pietrokovski S: **Searching databases of conserved sequence regions by aligning protein multiple-alignments.** *Nucleic Acids Res* 1996, **24:**3836-3845.
43. Silverstein KA, Shoop E, Johnson JE, Retzel EF: **MetaFam: a unified classification of protein families. I. Overview and statistics.** *Bioinformatics* 2001, **17:**249-261.
44. **MetaFam** [http://metafam.ahc.umn.edu/]

45. Wu CH, Xiao C, Hou Z, Huang H, Barker WC: **iProClass: an integrated, comprehensive and annotated protein classification database.** *Nucleic Acids Res* 2001, **29:**52-54.
46. **iProClass** [http://pir.georgetown.edu/iproclass/]
47. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29:**11-16.
48. **CDD: A Conserved Domain Database and Search Service** [http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml]
49. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29:**37-40.
50. **InterPro** [http://www.ebi.ac.uk/interpro/]
51. Murvai J, Vlahovicek K, Barta E, Pongor S: **The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments.** *Nucleic Acids Res* 2001, **29:**58-60.
52. **SBASE** [http://www3.icgeb.trieste.it/~sbasesrv/]
53. **Jalview - a java multiple alignment editor** [http://www.ebi.ac.uk/~michele/jalview/]
54. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16:**10881-10890.
55. Zdobnov EM, Apweiler R: **InterProScan - an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848.