PROTEIN FAMILY DATABASES FOR AUTOMATED PROTEIN DOMAIN IDENTIFICATION

ERIK L.L. SONNHAMMER

Computational Biology Branch, National Center for Biotechnology Information,
National Library of Medicine, Building 38A, Room 8N805
National Institutes of Health, Bethesda, Maryland 20894, USA

ABSTRACT

Automatic identification and annotation of protein domains is a major challenge for genome sequencing projects. Simple transfer of the annotation from the overall most similar protein with a known function is relatively reliable for prokaryotic proteins, but often produces misleading and incomplete results for multi-domain proteins, which are common in higher organisms. An alternative approach is to classify protein domains based on matches to a precompiled database of protein domain families. A number of such databases are reviewed here, including an update on the Pfam database. The differences a user can expect to experience when using different databases for domain identification are illustrated by examples of known multi-domain proteins. The advantages and drawbacks of single-sequence versus multiple-alignment methods are also discussed. The degree of protein modularity was surveyed in the genomes of *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Haemophilus influenzae* by matching them to Pfam. While prokaryotic genomes typically have a small fraction of multi-domain proteins, that rarely contain more than three domains, at least 10% of higher eukaryotic proteins have multiple domains, many times with dozens of domains per protein chain.

KEYWORDS

Protein Family Database, Protein Domain Identification, Protein Sequence Clustering, Modular Proteins, Protein Database Searching

INTRODUCTION

A number of new sequence analysis challenges have emerged in the genome era. Predicting the function of each newly found protein has been a main focus of genome analysis (Scharf *et al.*, 1994; Bork *et al.*, 1995; Casari *et al.*, 1996; Koonin *et al.*, 1996; Tatusov *et al.*, 1996). The analyses of the first complete genomes have concentrated on improving the percentage of proteins for which any functional inference can be made, no matter how small a portion of the sequence contains that information. Few attempts have been made to investigate the modularity, or the existence of multiple domains, in these proteins. One reason for this is that all genomes completed to date are either prokaryotic or single-cell eukaryotic, which only contain a small

fraction of multi-domain proteins. As the genome of a higher eukaryote, the nematode *Caenorhabditis elegans*, is nearing completion and the human genome project is accelerating, ignoring the issue of multiple domains on the same protein chain is becoming a more serious problem. Proper domain annotation is vital for biological inferences based on sequence similarity. The simple approach of only carrying over functional annotation from the most similar sequence may conceal important sequence features and lead to incorrect functional interpretations.

To meet the challenge of domain parsing on a large scale, there is a need for automated approaches. One could imagine that this might be achieved by condensing the results from a 'traditional' database search against a single-sequence database automatically, emulating the analytical process a human expert would perform to infer domain boundaries. However, this process involves a substantial amount of intuition and interaction, and is hard to generalize for different types of protein families. For known domains, it is possible to exploit prior knowledge about the family in order to predict new members. A database of such domain families can thus be used for parsing the domain architecture of newly found proteins. Although positive identification is limited to domains represented in the database, new domains in flanking segments also become easier to parse and analyse. To be useful, such a database needs to be as comprehensive as possible. Aside from the better definition of the domain boundaries, the approach of using a database of aligned protein domains has the advantage of being potentially more sensitive to weak similarities, since well-conserved features can be given a higher weight.

This paper addresses the question of how useful protein family databases are for automated protein domain identification, in the sense of what a sequence analyst can expect to gain from searching them, relative to traditional single-sequence database searching. Five protein family that are available for searching, Prosite, Blocks, Prints, Pfam and ProDom, are compared to traditional Blast searching. Without prior knowledge, both multiple alignment and pairwise approaches should be used, since the database of single sequences will always be more comprehensive. However, when no pairwise matches are found, or when they produce a partial or complex picture of homologous domains, a significant improvement in the analysis may result from using a multiple alignment-based database.

To assess how common it is that domain analysis is required, proteins from Swissprot and three genomes were analyzed for the presence of multiple domains by matching them to Pfam.

REVIEW OF PROTEIN FAMILY DATABASES

To construct a database of protein families, a number of basic steps are required. First, clusters corresponding to families must be generated. Each cluster member should have a defined starting and ending position in the full-length sequences to avoid including unrelated domains. For each cluster a multiple alignment is generated, which may be tested for quality and for specificity and sensitivity to find the known members, and documentation may be added. After the initial creation, methods to keep the database up-to-date should be developed. Each of these

steps represents hurdles that have to be resolved by a compromise between quality, efficiency and comprehensiveness. The main issues for each step are outlined below.

Clustering. In general, the quality of clusters created with manual inspection of sequence similarity and known functions far exceeds the quality of clusters generated fully automatically. On the other hand, they cannot compete in terms of comprehensiveness or speed of construction. Also, since many hand-built families embody a particular perspective of that family, it is virtually impossible to incorporate available hand-built families into a self-consistent database. For instance, one family clustering may have been constructed in order to find as many distant members as possible to a superfamily, while another clustering may be meant to distinguish a subfamily. Automated similarity-based clustering approaches face two main concerns. If based on an $\boldsymbol{O}(N^2)$ algorithm, computation time is likely to become a problem. A more severe problem is that multi-domain proteins may incorrectly join unrelated clusters. To solve this, a method to infer domain boundaries is needed, so that only the segment belonging to a particular family is clustered with it. For both manual and automatic approaches, the generation of non-overlapping clusters is, although not strictly necessary, a feature that will make the eventual analysis less ambiguous.

Generation of Multiple Alignments. Given a family clustering, multiple alignments can be generated with a wide range of available methods. This is usually relatively straightforward for the 'core domains' (*i.e.* the highly conserved, usually central parts of the sequences) but is far from a solved problem when only a segment of some sequences can be aligned with each other. Particularly if some sequences contain extra domains, most, if not all, alignment programs will try to incorrectly incorporate such domains instead of rejecting them, which generally leads to overall distorted alignments. This makes full automation of whole-domain alignment construction unreliable. One way of circumventing this problem is to only look for highly conserved short motifs, and leave out all unconserved regions.

Documenting and Maintaining the Clusters. To be useful for domain annotation, each cluster should be documented and linked to the literature. This is nearly a completely manual task, and may be more of a bottleneck than the computational aspects. Since the sequences in the primary databases are continuously updated (for instance when previous fragments are extended, sequencing errors are detected, or gene predictions refined), it is important to keep a multiple alignment database in synchrony with its member sequences. A database generated by a fully automatic clustering method faces two maintenance problems if it requires a complete re-run of the clustering for each release. First, computation time may be a bottleneck. Second, and more serious, if the clusters change in content, appearance or accession number at each release, they cannot be used as a stable reference from other databases.

Searching. To search a protein family database, the query has to be compared to a representation of each multiple alignment. This can be done in several different ways. The simplest method is to only extract the most conserved columns from the alignment, and describe those as a regular expression pattern in which only certain residues are allowed at certain positions. Searching by regular expression matching is very efficient. More information can however be extracted if each column is converted to a vector of 20 rows, with a score for each amino acid. This score is either

a probability, estimated from the observed frequencies in the column (Staden, 1989; Henikoff & Henikoff, 1991), or the average score of the amino acid in a substitution matrix against the observed amino acids in the column (Gribskov *et al.*, 1987). A common technique is to add 'pseudocounts' from substitution matrix scores to the observed frequencies to estimate probabilities of amino acids that may not have been seen due to small sample size. Gaps in the alignment are either avoided entirely, in which case the model is normally called a weight matrix, or they can become properties of the resulting model, which is then often called a profile (Gribskov *et al.*, 1987). A more formal way to describe position-specific probabilities for insertions, deletions, and amino acids is used in 'hidden Markov models' (HMMs) (Krogh *et al.*, 1994; Eddy, 1996). Such models are often called HMM profiles. Depending on the model, different search algorithms may be used. For weight matrices, simple scanning to find the best match is rapid and robust. For (HMM) profiles, dynamic programming methods are needed to find the optimal alignment. This is computationally more complex and hence slower, but can include gaps in the alignment.

Presented below are a selection of available multiple alignment databases. They were chosen for this study based on free availability of the multiple alignments or models, and a method to search a query protein sequence against them. In principle, any multiple alignment search method can be adapted to any multiple alignment database, but in order to avoid re-calibration only the methods provided with the database were used.

Prosite

The protein families in Prosite (Bairoch *et al.*, 1997) have largely been manually compiled from the literature. The current release, 13.0, contains 889 families. The emphasis of Prosite is on *functional motifs*, such as binding sites or other short amino acid patterns that share a common function, and high quality documentation of each family. Nearly all families are characterized by short regular expression-like patterns of allowed residues at conserved positions, often interleaved by non-conserved spacer columns. On average, the patterns contain ten conserved positions. In many cases, this is enough to separate the true members from the non-members, but for families with less conservation, the pattern approach apparently proved insufficient and whole-domain profiles were added. 16 such profiles are included in Prosite 13.0, and another 8 are available from the Prosite WWW site. Prosite is available at http://expasy.hcuge.ch/sprot/-prosite.html. Because of its comprehensiveness and excellent family annotation, Prosite is widely used as a reference for multiple alignment family databases.

Blocks

Prosite does not provide any multiple alignments, but they can be constructed from the list of members in Swissprot (Bairoch & Apweiler, 1997) that is attached to each Prosite family. This has been done automatically to generate short ungapped alignments of conserved regions, which are released in the Blocks database (Henikoff *et al.*, 1997). A given family may give rise to any number of blocks, which are assigned the original Prosite accession number plus a letter A, B, …

for each block in the group, ordered from the N-terminus to the C-terminus. The blocks in release 9.3 vary from 4 to 55 residues in width, and have an average of 3-4 blocks per family, and about 20 member sequences per block. Blocks is available at http://www.blocks.fhcrc.org/.

Prints

The Prints database (Attwood *et al.*, 1997) is similar to Blocks in that each family contains a number of short ungapped alignments of conserved regions. Like Prosite, Prints contains large amounts of documentation for each family, but unlike Blocks, the families are constructed by iteratively searching the database and manually validating the clusters. Regarding the alignments the main difference compared to Blocks is that Prints generally uses more blocks for each family, and that Prints blocks are shorter and have more members listed. Prints families contain more members partly because it uses OWL as primary database, which contains more sequences than Swissprot. Prints 16.0 contains 750 families, of which some 200 are not represented in Blocks (Henikoff *et al.*, 1997). The Prints blocks are 5-33 residues wide, with an average of 5-6 blocks per family, and about 35 members per block. Prints is available at http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html.

ProDom

The Prodom database (Sonnhammer & Kahn, 1994), is constructed by fully automatic clustering and multiple alignment generation. Local pairwise sequence similarities generated by Blastp (Altschul *et al.*, 1991) are processed by the program Domainer, which uses heuristics to infer domain boundaries in order to separate out clusters that supposedly correspond to domains. In earlier versions of ProDom, Domainer used the original pairwise similarity locations to quickly generate multiple alignments, but in the latest release, 34.1, the member sequence segments assigned to each cluster were re-aligned using the program Multalin (Corpet, 1988). In most cases, the ProDom alignments are shorter than true domains, and the alignment quality is significantly lower than what can be achieved by more manual approaches. Unfortunately, the more members in a family, the lower the quality tends to be. The reason for this is that Domainer is inherently sensitive to incorrect data from unmarked fragmentary proteins and incorrect matches reported by Blast. Families with many members thus run a greater risk of incorporating such deleterious information. Searching ProDom is normally done by Blast against consensus sequences from each cluster, or against all member sequences. Since the clusters typically only contain relatively similar sequences, and because the alignment quality is often poor, great leaps in sensitivity cannot be expected from searching against the multiple alignments. The alignments with more than one member are on average 122 residues wide, including gaps, and on average there are about 7 members per family. ProDom is available at http://protein.toulouse.inra.fr/-prodom.html.

Pfam

The main idea of Pfam (Sonnhammer *et al.*, 1997) was to construct a self-consistent, comprehensive collection of permanent, documented protein families with whole-domain alignments. By 'whole domain' is meant the smallest sequence segment that is able to fold and function independently of other segments of the same protein chain. Operationally the Pfam domains may sometimes be slightly smaller due to poor conservation outside the core of the domain. A lesson from ProDom is that having poorly defined and volatile clusters not only reduces the usefulness for cross-referencing from other databases, but also prevents gradual refinement of the clusters and alignments as more data is gathered. Central to Pfam's methodology is that each family is described by two alignments: a 'seed' alignment, which contains a number of representative full-length sequences, and a 'full' alignment containing all known members. The reason for keeping a separate seed alignment is that it is small enough to be checked, manipulated, and updated with ease. Fragmentary or incorrect sequences can thus be avoided in the data used to represent the family. From the seed alignment an HMM is generated which is used to gather all member sequences, and later on to probe Pfam with query sequences. Both alignments have to pass quality tests for correctness and consistency with other Pfam families. If not all members are found, or if either alignment is incorrect, parts of the process are re-iterated.

Most Pfam families have corresponding entries in Prosite. Of the 527 families in Pfam-A 2.0, 79 families do not have a reference to Prosite. However, many of the families that do, do not exactly correspond to the Prosite cluster, since often the level of clustering is different between the two databases. This is a consequence of the two different methods used. A Prosite family may contain a short conserved motif, which in Pfam is described by several whole-domain subfamilies, e.g. the P-loop containing families. Conversely, separate Prosite clusters may be joined in Pfam, because HMMs can represent weakly conserved domains better than patterns.

For Pfam-A release 2.0, 353 new families were added relative to Pfam-A 1.0. 55 of these were taken from Pfam-B, which is a supplementary database automatically generated by running Domainer on sequences not in Pfam-A at each release. The Pfam-B clusters can serve as initial seeds for Pfam-A families, and our goal is to try to incorporate all large Pfam-B families into Pfam-A.

The documentation in Pfam is generally brief, and largely relies on links to Prosite and Prints entries. To further improve the documentation, 69 links to WWW sites with protein family information were added in Pfam-A 2.0. Pfam can thus also serve as a central repository of pointers to electronically published protein family material, like ProWeb (Henikoff *et al.*, 1996). The alignments are on average 275 residues wide, including gaps. There are on average about 75 members per family in the full alignments, and about 22 in the seed alignments. Pfam is available at http://www.sanger.ac.uk/Pfam and http://genome.wustl.edu/Pfam.

CASE STUDIES

A protein containing nine well-studied domains, phosphatidylinositol-specific phospholipase C-γ, was chosen as a test case to illustrate how the results differ between databases. For easy comparison, the output from each database search was parsed into a standard format and drawn as a graphical schematic diagram, shown in Fig. 1. For reference, the first schematic shows the domains in the Swissprot feature table, which might be considered the 'true' domain architecture. It contains one C2 domain, one EF-hand calcium-binding domain, two PH domains, of which the second is split in segments, with two SH2 and one SH3 domain inserted in between, and two phospholipase C-specific domains called X and Y. The halves of the split PH domain are short and not easy to detect. In fact, to detect both of them in a Pfam search, the score cutoff needed to be lowered from 25 to 7 bits. In this case, no spurious matches were reported in the Pfam search at this cutoff level, but in general noise can be expected up to a score of 20-25 bits. A bit is a $\log_2$ information content measure, meaning that a score of 25 bits is expected by chance once in 2^{25} (3×10^7) alignments.

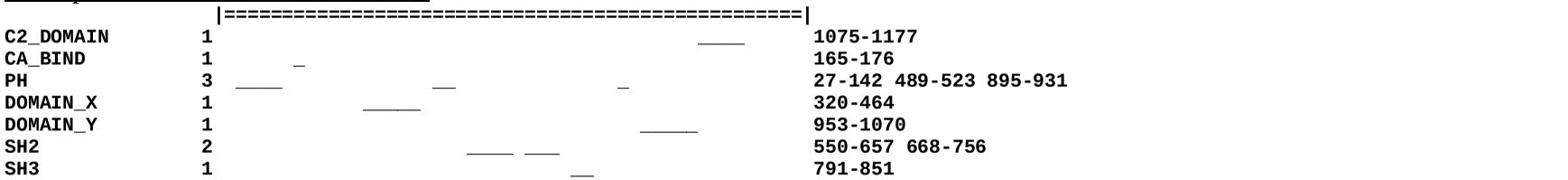
The Prosite patterns only detected the EF hand domain. However, all other domains in this case are represented in the 24 recently added Prosite profiles and were detected in a profile search. Both the Pfam and Prosite searches produce domain matches closely corresponding to the correct architecture.

Both the Blocks and Prints results identify the C2 domain, the X/Y domains, the SH2 and the SH3 domains, albeit in a less clear fashion due to the short motifs. For instance, there are 4 or 5 motif-matches to both the SH2 and SH3 domains, giving no obvious indication that there are two SH2 and only one SH3 domain. Although the PH and EF hand domains were not reported by Blocks or Prints searches, both reported some spurious matches as significant.

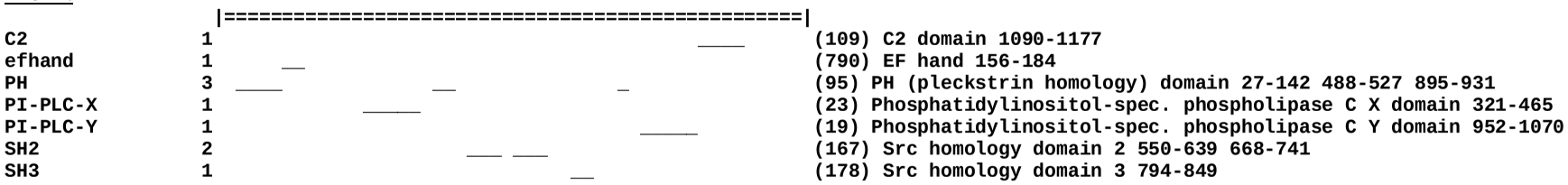
The ProDom output may look confusing at first glance because most matching domain families have very similar descriptions. However, at closer inspection the result corresponds reasonably well to certain domains. For instance, the top two matches, families 1317 and 1316 correspond to domains X and Y. Further down the list, the two matches to family 40 are the two SH2 domains and the match to family 10 is the SH3 domain. It can be seen that these families are 'superfamilies' because they have many more members than the other families (147 and 307, respectively). It is however not easy for a user to trace the fact that these families correspond to the well-known domains SH2 and SH3 as neither of them have the keywords SH2 or SH3 in the description (which is generated by ranking keywords in the sequence annotation of the members). A further complication in this case is that a large proportion of the members in family 10 do not contain SH3 domains, but were nonetheless assigned to this cluster, apparently in many cases due to low complexity matches. The main drawback of the ProDom analysis is thus that it lacks family-level annotation, and the user must make a considerable effort to establish which domains were actually found.

From the output of Blastp, only representative matches are shown. While it does not give a detailed and easily interpretable picture of the domain organization, it does give clues to the fact that domain shuffling is present. It is clear that several other phospholipase C sequences

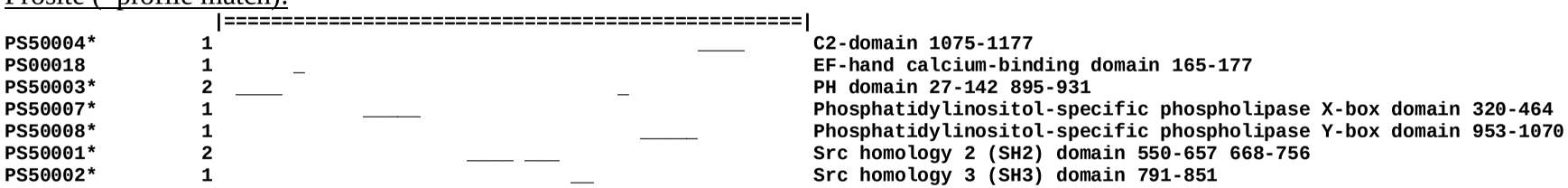
Swissprot Feature Table Annotation:



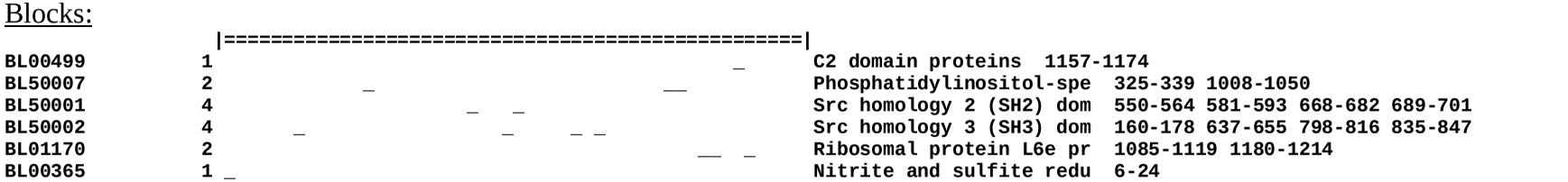
Pfam:



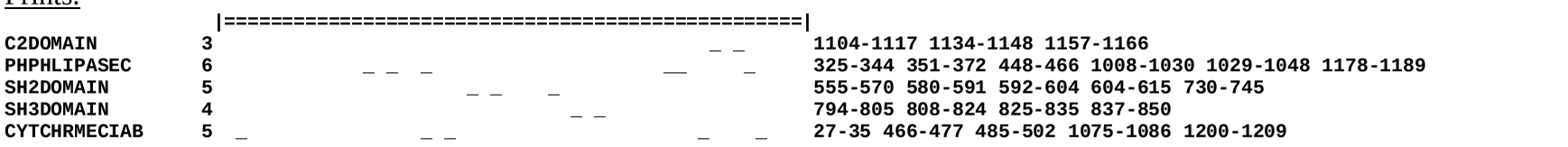
Prosite (*profile match):



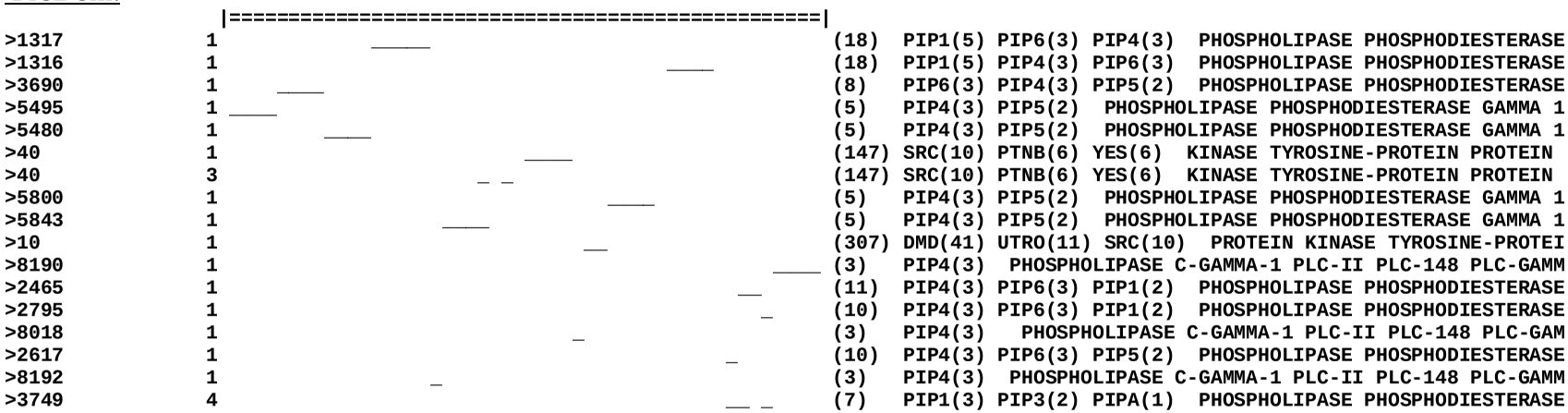
Blocks:



Prints:



ProDom:



Blastp search against Swissprot (pruned):

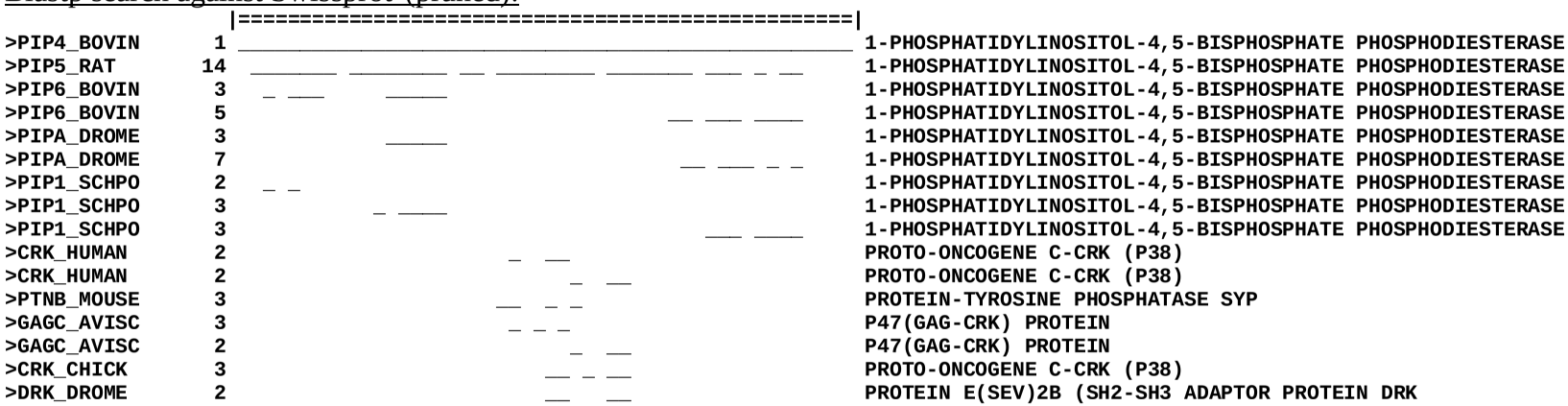


Fig. 1. Example of analysis of a multi-domain protein, phosphatidylinositol-specific phospholipase C-γ (Swissprot PIP4_BOVIN/P08487) by six different methods. The top diagram shows the location of the nine known domains, taken from Swissprot. The next five diagrams show results from search methods that exploit protein family databases, and as comparison the result of a Blastp search is shown last. The query PIP4_BOVIN is shown as a double line, while database matches are single lines. Each match line contains, from left to right: entry name, number of matches, schematic location of match(es), number of members in family (if reported), description (if reported), and start-end positions of each individual match. Multiple matches to the same database entry are merged on one line, except in the Blast output, where they are only merged if considered consistently ordered.

lack the SH2 and SH3 domains, but contain the flanking domains, while other sequences only match at the SH2 and/or SH3 domains. As in the ProDom search, the more subtle domain members, like EF hand and PH, are not easily distinguishable however.

The conclusion for this example is that the search results appear to fall into three classes. The profile-based approaches give a clear, virtually complete picture, at least in this example. The motif databases give a relatively concise picture, but it is less clear where precisely the domains are located; some domain matches are completely absent, and some spurious matches are reported. The fully automatically clustered and the unclustered databases give hints to where domains might be located, but the picture is blurred and incomplete, and the results require substantial manual analysis to produce the complete domain architecture.

Another example, protein kinase C from yeast, (Swissprot KPC1_YEAST/P24583), produces a similar picture (not shown). It contains a C2 domain, a Phorbol esters / diacylglycerol binding domain and a protein kinase domain. Pfam and Prosite profile searches produced matches that correspond well to the known domains, but the Prosite profile search failed to detect the C2 domain here. Pattern and motif searches give matches to all but the C2 domain, and one spurious match each, while Blastp against ProDom or single-sequence databases again produce less easily interpretable results.

HOW COMMON ARE MODULAR PROTEINS?

Modular protein domains, which can be shuffled during evolution, are found in essentially all organisms. They are often used in regulatory signaling systems (Bourret *et al.*, 1989; Pao & Saier, 1997) and transport across membranes (Reizer *et al.*, 1996), and in higher eukaryotes also to a great extent for extracellular structural proteins (Bork, 1991). Apparently the evolution of these systems has been promoted by the ability to quickly generate new combinations of already functional building blocks (Doolittle & Bork, 1993). It has been observed that prokaryotes only contain a small fraction of multi-domain proteins, whereas in higher eukaryotes (which have greater needs for signal transduction, and for large structural proteins) they are more common-place. To give an illustrative quantification of how common modular proteins might be, protein domains were extracted by searching Pfam against Swissprot 34 and three genomes. The results are shown in Table 1 and in Fig. 2. The bacterium *H. influenzae* has very few multi-domain proteins, no more than three domains per chain, while yeast has a larger fraction, up to about 10 domains per protein. The proteins in the nematode *C. elegans* and in Swissprot contain up to about 50 domains. However, most proteins with a large number of domains contain arrays of the same domain, and the maximum number of *different* domains on one chain is not significantly higher in *C. elegans* than for yeast. Presently, the maximum number of domains in one protein chain is about 245. These are immunoglobulin and fibronectin type III domains, and one protein kinase domain, in the 26926 amino acids long human muscle protein titin (Labeit & Kolmerer, 1995). Titin and other extremely long protein sequences are not present in Swissprot. About 8% of the proteins in the nematode *C. elegans* and in Swissprot contain multiple domains. Since Pfam does not yet contain all modular domains, and may have failed to detect a number of them, it does not seem unreasonable to suggest that at least 10% of all higher eukaryotic proteins may consist of multiple domains.

Table 1. The fraction of proteins with multiple domains in Swissprot and the genomes of the nematode *C. elegans* , the yeast *S. cerevisiae* and the bacterium *H. influenzae*. See Fig. 2 for the distribution of proteins with a certain number of domains. *The domain counts are approximate (and probably underestimated), since they were estimated from matches to Pfam.

	Nr. of proteins	Proteins matching Pfam	Nr. of multi-domain* proteins	Max domains* per protein	Max different domains* per protein
Swissprot 34	59021	28169 (48%)	4838 (8%)	60	6
60% of *C. elegans*	7263	1720 (24%)	558 (8%)	44	5
S. cerevisiae	6719	1644 (24%)	360 (5%)	10	4
H. influenzae	1680	358 (21%)	30 (2%)	2	2

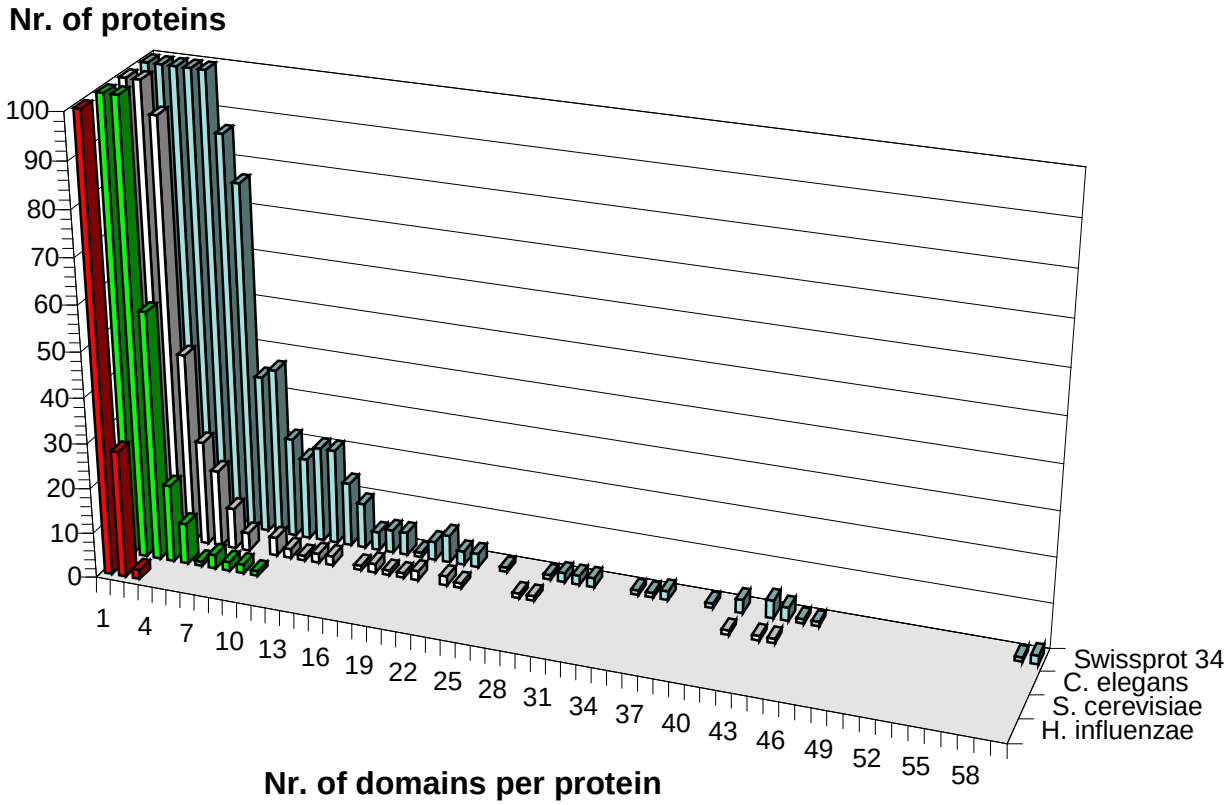


Fig. 2. The modularity of proteins in Swissprot 34 and three genomes, illustrated by a histogram of matches to Pfam domain families. See Table 1 for what percentage in each genome consists of multi-domain proteins.

METHODS AND MATERIALS

The following servers were used for the database searches. Prosite: http://expasy.hcuge.ch/-sprot/scnpsite.html for patterns and http://ulrec3.unil.ch/software/profilescan.html for profiles, Blocks: http://www.blocks.fhcrc.org/blocks_search.html, Prints: http://www.biochem.ucl.ac.uk/-cgi-bin/scordis/fingerPRINTScan/bin/FPSCAN_FORM2.cgi, Prodom: http://-protein.toulouse.inra.fr/prodom/blast_form.html, Pfam: http://www.sanger.ac.uk/Software/-Pfam/HMM_search.shtml.

MSPcrunch 2.1 (Sonnhammer & Durbin, 1994) was used to parse the output of Blastp 1.4, which was used to search the ProDom and Swissprot databases. The schematic diagrams were generated using the 'Big Picture' output function in MSPcrunch. All matches shown in Fig. 1 were reported as significant, except the two C-terminal PH domains in the Pfam search.

The *C. elegans* protein sequences were compiled in the database Wormpep, release 11, which contains 7263 unique genes, approximately corresponding to 60% of all *C. elegans* proteins. Wormpep is available at ftp://ftp.sanger.ac.uk/pub/databases/wormpep. The *S. cerevisiae* protein sequences were extracted from Swissprot and TREMBL (Bairoch & Apweiler, 1997), and the *H. influenzae* protein sequences were provided by TIGR at http://www.tigr.org/.

DISCUSSION

The aim of this paper was to (1) examine the critical issues surrounding automatic protein domain parsing, and (2) to review existing approaches based on protein domain family databases. Since these databases were created in different ways for somewhat different purposes, it is inherently difficult to make a fair comparison. Instead of attempting a large-scale comparison, a few examples were looked into in detail, in order to provide some insight, however anecdotal, into what sort of results a potential user can expect when using these databases for analyzing a query sequence. The main example was chosen because it was a challenging case with many well-characterized domains, not because it was known to favor a particular method. It should be stressed that for queries that only contain subtle, short similarities to known families, the motif databases may be more sensitive than the whole-domain databases. If no match is found to whole-domain databases, or if only a partial match is found, it is therefore wise to search a motif database, such as Blocks or Prints, as the next step. For relatively clear domain homologies, however, the overall result is that protein family databases can be very useful for assisting the domain identification, and that whole-domain approaches generally give a clearer picture than motif-based methods. All pre-clustering approaches offer some advantages over single-sequence searching.

The fact that up to 10% of all proteins appear to contain multiple domains indicates that this issue should be considered an important aspect of genome analysis, especially as genome projects of higher eukaryotic organisms get underway.

ACKNOWLEDGEMENTS

Robert Finn is gratefully acknowledged for preparing most of the new additions to Pfam 2.0. I thank Hugues Sicotte, Lisa Turni and Sean Eddy for critical reading of the manuscript.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E.W. Myers and D. J. Lipman. (1990). Basic local alignment search tool. *J. Mol. Biol.*. 215. 403-410.

Attwood, T. K., M. E. Beck, A. J. Bleasby, K. Degtyarenko, A. D. Michie and D. J. Parry-Smith. (1997). Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res.* 25.. 212-217.

Bairoch, A. and R. Apweiler (1997). The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucleic Acids Res.*. 25. 31-36.

Bairoch, A., P. Bucher and K. Hofmann (1997). The PROSITE database, its status in 1997. *Nucleic Acids Res.*. 25. 217-221.

Bork, P., C. Ouzounis, G. Casari, R. Schneider, C. Sander, M. Dolan, W. Gilbert and P. M. Gillevet. (1995). Exploring the Mycoplasma capricolum genome: a minimal cell reveals its physiology. *Mol. Microbiol.*. 16. 955-967.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett.*. 286. 47-54.

Bourret, R. B., J. F. Hess, K. A. Borkovich, A. A. Pakula and M. I. Simon (1989). Protein phosphorylation in chemotaxis and two-component regulatory systems of bacteria. *J. Biol. Chem.*. 264. 7085-7088.

Casari, G., A. De Daruvar, C. Sander and R. Schneider (1996). Bioinformatics and the discovery of gene function. *Trends Genet.*. 12. 244-245.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*. 16. 10881-10890.

Doolittle, R. F. and P. Bork (1993). Evolutionarily mobile modules in proteins. *Sci. Am.*. 269. 50-56.

Eddy S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.*. 6. 361-365.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*. 84. 4355-4358.

Henikoff, J. G., S. Pietrokovski and S. Henikoff (1997). Recent enhancements to the Blocks Database servers. *Nucleic Acids Res.*. 25. 222-226.

Henikoff, S., S. A. Endow and E. A. Greene (1996). Connecting protein family resources using the proWeb network. *Trends Biochem.. Sci.* 21. 444-445.

Henikoff, S. and J. G. Henikoff (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*. 19. 6565-6572.

Koonin, E. V., E. V. Tatusov and K. E. Rudd (1996). Protein Sequence Comparison at Genome Scale. *Meth. Enz.*. 266. 295-322.

Krogh, A., M. Brown, I. S. Mian, K. Sjoelander and D. Haussler (1994). Hidden Markov model in computational biology. Applications to protein modelling. *J. Mol. Biol.*. 235. 1501-1531.

Labeit, S. and B. Kolmerer (1995). Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science.* 270. 293-296.

Pao, G. M. and M. H. Jr. Saier (1997). Nonplastid eukaryotic response regulators have a monophyletic origin and evolved from their bacterial precursors in parallel with their cognate sensor kinases. *J. Mol. Evol.*. 44. 605-613.

Reizer, J., A. Reizer and M. H. Saier Jr. (1996). Novel PTS proteins revealed by bacterial genome sequencing: a unique fructose-specific phosphoryl transfer protein with two HPr-like domains in *Haemophilus influenzae*. *Res. Microbiol.*. 147. 209-215.

Scharf, M., R. Schneider, G. Casari, P. Bork, A. Valencia, C. Ouzounis and C. Sander (1994). GeneQuiz: a workbench for sequence analysis. In *ISMB-94; Proceedings Second International Conference on Intelligent Systems for Molecular Biology,* pp. 348-353, AAAI Press, Menlo Park.

Sonnhammer, E. L. L. and R. Durbin (1994). A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.*. 10. 301-307.

Sonnhammer, E. L. L., S. R. Eddy and R. Durbin (1997). Pfam: a Comprehensive Database of Protein Domain Families Based on Seed Alignments. *Proteins.* 28. 405-420.

Sonnhammer, E. L. L. and D. Kahn (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*. 3. 482-492.

Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* 5. 89-96.

Tatusov, R. L., A. R. Mushegian, P. Bork, N. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd and E. V. Koonin (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6.. 279-291.