

Domains, motifs and clusters in the protein universe

Jinfeng Liu^{*†} and Burkhard Rost^{*‡}

The rapid growth of bio-sequence information has resulted in an increasing demand for reliable methods that group proteins. A few databases with curated alignments of protein families have demonstrated that expert-driven repositories can keep up with the data deluge in the genome era. These original resources implicitly identify domain-like modules in proteins. An increasing number of automatic methods have sprouted over the past few years that cluster the protein universe. Many of these implicitly dissect proteins into structural domain-like fragments. In a very coarse-grained evaluation, some of the automatic methods appear to be on par with expert-driven approaches. However, neither automatic nor manual methods are currently entirely up to the challenges of tasks such as target selection in structural genomics. Thus, we urgently need refined and sustained automatic clustering tools.

Addresses

^{*}CUBIC and North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

[†]Department of Pharmacology, Columbia University, 630 West 168th Street, New York, NY 10032, USA

e-mail: liu@cubic.bioc.columbia.edu

[‡]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Correspondence: rost@columbia.edu; <http://cubic.bioc.columbia.edu>

Current Opinion in Chemical Biology 2003, 7:5–11

This review comes from a themed issue on
Proteomics and genomics

Edited by Matt Bogyo and James Hurley

1367-5931/03/\$ – see front matter

© 2003 Elsevier Science Ltd. All rights reserved.

DOI 10.1016/S1367-5931(02)00003-0

Introduction

Gordon Moore correctly predicted that the potency of computers should double every 18–24 months (Moore's 'law') [1]. The only example of computer-independent information growing faster may be the unravelling of bio-sequences [2]. And while the growth of computer potency is beginning to slow down [3], the growth-rate for bio-sequences continues to grow. This reality is one of the technical reasons why clustering and classifying proteins is becoming increasingly important. We challenge that there is no reasonable way of clustering and classifying proteins without dissecting proteins into structural domain-like fragments [4^{*},5]. In fact, such domain-like fragments also appear crucial for inferring structure and

function. Here, we review some of the recent manual and automatic methods that attempt to classify proteins (URLs in Table 1). A glossary of terms is provided in Box 1.

Expert-curated databases of protein families

Motifs and domains

Two types of expert-curated resources complement one another: motif-based and domain-based databases. It is extremely difficult to infer similarities in structure or function from short alignments [6]. Particular short sequence motifs such as nuclear localization signals [7] are related to protein function. These motifs often span evolutionarily divergent families. In fact, short motifs may constitute candidates for the 'atoms of evolution' [5]. Even more powerful are motifs defined by proximity in three-dimensional (3D) structures that constitute skeletons of 'functional units' [8]. However, protein families often cannot be characterised by single motifs. In contrast, structural domains constitute regions that share a common fold, have some functional similarity, and may be evolutionarily related. Thus, domain-based families capture biologically crucial features beyond short motifs.

Motif-based classifications

PROSITE motifs are extracted from the literature [9–11]; annotations can then be cross-linked to and updates synchronised with the SWISS-PROT database of protein sequences [12]. Not all motifs are equally informative; this reality is reflected by statistics on how often a certain motif matches in SWISS-PROT. Families are usually defined as 'all proteins that share a certain motif' and the motif can be described as a regular expression (e.g. [KH]DE[LF] abbreviates the following four peptides: KDEL, HDEL, KDEF, HDEF). Profiles have been added to enable detection of diverged families; these profile-extended patterns currently cover 15% of all entries. The Blocks database [13] builds un-gapped, weighted local alignments (blocks) through dynamic programming [14] for proteins grouped by PROSITE, PRINTS [15], Pfam-A [16^{*}], ProDom [17] and Domo [18]. Blocks alignments extend over 5 to 55 residues (Figure 1b). The PRINTS [15] database also contains groups of aligned, un-weighted motifs referred to as 'fingerprints' that are derived through iterative database searches, followed by semi-manual alignments, and by a final manual validation/annotation.

Structure-based domain classification

A particular example of 'human with machine vs. machines' is the SCOP classifications for proteins of

Table 1

Availability of databases and methods

DB/Method	Version	Latest update	Entries	Update	URL (all begin with http://)
Short sequence motifs					
PROSITE	17.23	10/2002	1573	Manual	www.expasy.ch/prosite/
Blocks+		8/2001	8656	Manual	blocks.fhcrc.org/blocks/
PRINTS	35.0	7/2002	1750	Manual	www.bioinf.man.ac.uk/dbbrowser/PRINTS/
Structural domain-like regions					
Pfam-A	7.6	9/2002	4463	Manual	pfam.wustl.edu
TIGRFAM	2.1	9/2002	1622	Manual	www.tigr.org/TIGRFAMs/
SMART	3.4	10/2002	654	Manual	smart.embl-heidelberg.de
SBASE	9.0	10/2002	483	Semi-manual	hydra.icgeb.trieste.it/~kristian/SBASE/
DOMO	2.0	4/1998		Automatic	www.infobiogen.fr/services/domo/
ProDom	2001.3	12/2001		Automatic	prodes.toulouse.inra.fr/prodom/doc/prodom.htm
GeneRAGE				Automatic	www.ebi.ac.uk/research/cgg/services/rage/
TribeMCL				Automatic	www.ebi.ac.uk/research/cgg/tribe/
CHOP		10/2002		Automatic	cubic.bioc.columbia.edu/db/chop/
Integration					
InterPro	5.2	9/2002	5875	N/A	www.ebi.ac.uk/interpro/
MetaFam	4.1	9/2002		N/A	metafam.ahc.umn.edu
Clusters of proteins					
CluStr				Automatic	www.ebi.ac.uk/clustr/
SYSTEMS	3.0			Automatic	systems.molgen.mpg.de
PICASSO	0	3/1998		Automatic	systems.molgen.mpg.de
ProtoNet	1.4	9/2002		Automatic	www.protonet.cs.huji.ac.il/protonet/
ProClust	1.0			Automatic	promoter.mi.uni-koeln.de/~proclust/

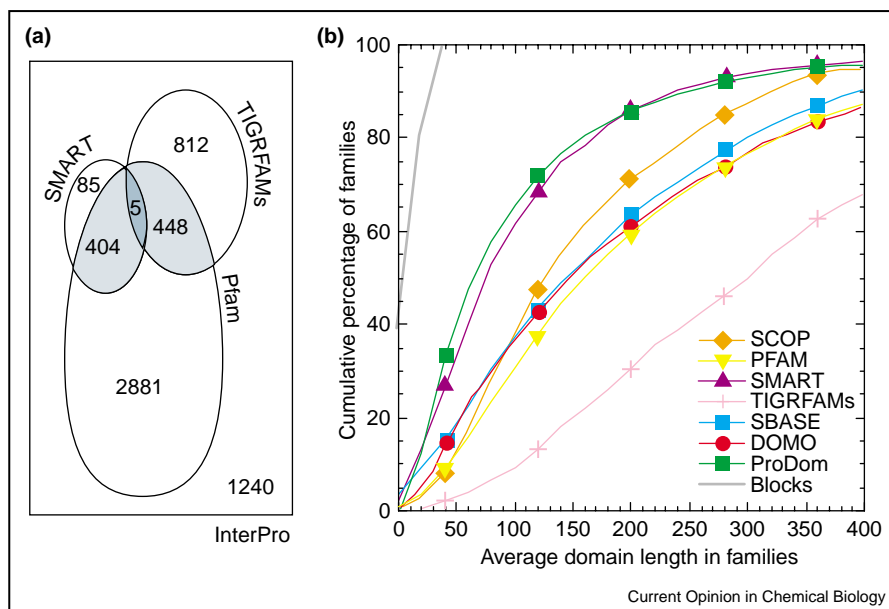
known structure [19]. When structures are added to the PDB [20], Alexei Murzin (MRC, Cambridge University, UK) visually classifies them into 'known fold' and 'new fold'. Folds are further grouped into families and super-families, and structural domains are assigned. The CATH

protein structure family database also classifies structures and defines domains [21]; it has moved steadily from expert-driven to automatic classifications. Fully automated structure-based domain classifications are available through DALI [22], VAST [23] and PrISM [24].

Box 1 Glossary of terms

3D structure	Three-dimensional co-ordinates of protein structure
Blocks	Database of protein alignment blocks derived from multiple compilations [13]
CATH	Class, Architecture, Topology and Homologous superfamily, a database classifying protein domain structures hierarchically [21]
COGs	Clusters of orthologous groups of proteins [30]
DOMO	A database of aligned protein domains [18]
HMM	Hidden-Markov model (i.e. particular alignment method)
InterPro	Database cross-linking SWISS-PROT, TrEMBL, PROSITE, ProDom, PFAM, PRINTS, SMART, and TIGRFAMs [31**,59]
MetaFam	A database of unified classification of protein families [33*]
OMIM	Online Mendelian Inheritance in Man, a database of human genes and genetic disorders [28]
PDB	Protein Data Bank of experimentally determined 3D structures of proteins [20]
Pfam	Database of expert-curated alignments of protein families (strictly called 'Pfam-A') [16*]
PRINTS	Database of probable protein signatures [15]
ProDom	Database of putative protein domains [17]
PROSITE	Database with expert-annotated functional sequence motifs [9–11]
PSSM	Position specific scoring matrix
SMART	Simple Modular Architecture Research Tool, a database of expert-curated protein modules [26]
Smith–Waterman algorithm	A dynamic programming approach to find local sequence similarities
SWISS-PROT	Database of protein sequences [12]
TIGRFAMs	Expert-curated database of protein families based on HMMs developed by The Institute of Genome Research [25]
TrEMBL	Computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT [12]
UniProt	Database merger between PIR, SWISS-PROT and TrEMBL (future)

Figure 1



Manual and automatic domain-based databases. **(a)** Venn-diagram of overlap between Pfam, TIGRFAMs and SMART. For each InterProt entry (release 5.2, 5875 entries), we tracked whether or not Pfam, TIGRFAMs, or SMART were in the 'member list'. The numbers shown in circles are mutually exclusive; for example, 2881 of the Pfam entries were only in Pfam, 404 were in Pfam and SMART, 448 were in Pfam and TIGRFAMs, and 5 were in Pfam, TIGRFAMs and SMART; 1240 entries were not found in any of the three databases. **(b)** Length distribution of fragments. We plotted the average lengths of family entries against the cumulative percentage of families. For SCOP [19] (version 1.59, 1824 families) and SBASE [29], the numbers refer to the average lengths of all sequences in each family/domain, for PFAM [16*], TIGRFAMs [25] and SMART [26] to the lengths of the HMMs, and for Blocks+ [13], DOMO [18] and ProDom [17] to the lengths of the family alignments. The closer the curves to the central line defined by SCOP, the more the entries in that database resemble structural domains. All Blocks+ alignments are shorter than 55 residues. Because Blocks+ is not designed to capture structure-like domains, the Blocks+ distribution constitutes the lower end of the distribution (too fragmented). The corresponding upper end (too long) is given by TIGRFAMs for which the distribution is similar to that of full-length proteins [5]. ProDom and SMART are biased towards short fragments, with almost half of the families shorter than 60 residues. In comparison to the expert-curated SMART modules, the automatic ProDom domain dissection appears surprisingly accurate, on average. The observation that all other data sets fall below SCOP indicates that too many proteins are not dissected into domains.

Classifying structural domain-like families

Another comprehensive expert-curated resource is Pfam [16*] (more precisely Pfam-A). Pfam pioneered two concepts: the building of seed alignments for domain-like regions, and the extension of seed alignments into larger families. Domain seeds from the literature are extended by expert-controlled searches with HMMer. New domain-like entries are added according to their popularity in the recent literature. TIGRFAMs [25] and SMART [26] implement a similar strategy as Pfam; thus all three overlap (Figure 1a). One difference is that SMART seeds are identified by PSI-BLAST [27]. SMART modules are much shorter than structural domains (Figure 1b). SMART estimates the likelihood of a given domain to be secreted, cytoplasmic or nuclear and annotates transmembrane helices, coiled-coils, signal peptides, internal repeats and cross-links to OMIM, a database of human genes and genetic disorders [28]. SBASE [29] groups families by recursively applying k-means clustering to proteins with similar biological names. Families are defined as groups of domain-like regions with significant BLAST similarities; a new query

sequence can be assigned to the family either by nearest-neighbour approach, a probabilistic score or by neural network. Another grouping is realised by the COGs database that attempts to classify proteins according to phylogeny [30].

Database integration

All expert-curated family databases have their strengths and weaknesses. InterPro [31**] provides a unified documentation resource for protein families, domains and functional sites by merging annotations from PROSITE, PRINTS, Pfam, TIGRFAMs, SMART and ProDom. The next-generation extension UniProt will merge SWISS-PROT, TrEMBL, InterPro and PIR resources [32]. MetaFam [33*] combines Blocks, DOMO [18], Pfam, PIR-ALN, PRINTS, PROSITE, ProDom [17], ProtoNet [34], SBASE, and SYSTEMS [35]. MetaFam first converts all proteins in the family databases into a common set of non-redundant proteins, then common families are identified, and supersets are created. Domain boundaries are identified through finding consensus regions among the databases.

Automatic clustering methods

Different objectives yield different clusters

One problem for automatic clustering methods is the definition of similarity thresholds that yield biologically relevant classifications. Another problem is sketched by the following alternative: first, group all proteins that share feature X into one cluster, Y, and second, ascertain that no protein outside cluster Y shares feature X with any protein in Y. Both objectives first must translate ‘similarity in sequence’ into ‘similarity in feature X’. For the feature ‘similarity in structure’, the criteria are well defined in the following way: if the sequence similarity (S_{AB}) between proteins A and B exceeds threshold T, we can reliably infer that A and B are structurally similar [24,36]. However, if $S_{AB} < T$, we have no clue. In other words, we cannot systematically identify all proteins that have common folds [4^{*}]. The threshold problem becomes more difficult when we want to infer similarity in function: different aspects of function such as sub-cellular localisation [37], enzymatic activity [6,38^{*}], or cellular function [39^{*}] require different thresholds. We may seek a way out of this problem by restricting clusters to close homologues, such as COGs [30]. However, the dilemma between the *Skylla* of ‘restrictive thresholds yielding many small clusters’ and the *Charibdis* of ‘permissive thresholds yielding few large clusters’ is a principle one. Neither objective automatic, nor subjective expert-driven classifications can ship around this problem.

Evaluating clustering methods is problematic

Since there is no single ‘correct’ solution to the clustering problem, there is also no unambiguous way to evaluate methods. Methods classifying proteins into structural domains could be compared to large sets of structural domains annotated by SCOP, CATH, DALI [22], VAST [23] and PrISM [24]. However, even structure-based domain assignments agree only to some extent. A simple, coarse-grained feature is the agreement between the distributions of domain lengths suggested by structure-based and by sequence-based domain assignments (Figure 1b).

Clusters establish similarity not distance

Calculating pairwise sequence similarity is usually the first step toward clustering. Expectation values (E-values) from BLAST/PSI-BLAST [40] are adopted to save computer time [17,18,35,41,42]. One problem is that E-values change when adding sequences to the cluster. Another problem is that BLAST E-values are not symmetric (i.e. they differ between aligning A against B and aligning B against A). Most methods account for the asymmetry by *ad-hoc* hacks: for example, use Smith–Waterman alignments when only one BLAST E-value is above the threshold [35,41], or replace asymmetric E-values by averages over both [43^{*}]. Other methods establish similarity through Smith–Waterman alignments [14]. For example, ProClust uses normalised Smith–Waterman

scores [44^{*}]; CluSTr uses Z-scores resulting from Monte-Carlo simulations of Smith–Waterman alignments [45]. ProtoMap [42] ProtoNet [34] and BioSphere [46^{*}] combine measures from Smith–Waterman, BLAST, and FASTA alignments. One important reality of sequence comparisons is that alignment methods optimise the similarity between two sequences. ‘Less similar’ does not imply ‘more distant’. To illustrate this point for structural similarity, 90% of all pairs of proteins that have 15% identical residues over their entire length have different structures; however, 90% of the pairs of proteins with similar structure have less than 15% identical residues [2,24,36].

Clustering without considering the domain problem

Some methods try to ignore the domain problem by applying very conservative thresholds. SYSTERS clusters proteins with BLAST E-values $< 10^{-40}$ by single-linkage [35]. At this level, many partial matches in multi-domain proteins are eliminated at the cost of small clusters. ProtoNet classifies proteins at different levels of confidence [34]. It begins at a high sequence similarity (E-values $< 10^{-100}$) with many small clusters. These initial clusters are then merged gradually at various levels of similarity [47^{*}]. Users can determine the wanted level of coarse-grained representation by dialling through different thresholds.

Domain-based clustering

A few methods explicitly predict domain boundaries from sequence information, in particular, through database searches [48^{*},49], concepts from protein folding [50], statistics [51], and neural networks [52,53]. None of these is well enough established yet for large-scale sequence analysis. Many proteins appear to have regions depleted of regular structure [54]; identifying such regions may assist the prediction of domain boundaries [55^{*}]. However, most methods that predict domain boundaries use alignment information and also classify the protein universe in two steps: first, by chopping proteins into domain-like fragments; and second, by clustering these fragments (ProDom [17], DOMO [18], GeneRAGE [41], CHOP [56]). ProDom applies the following algorithm [17]. First, stack all sequences in SWISS-PROT and TrEMBL. Then iterate by identifying the shortest sequence in the stack, finding related regions through PSI-BLAST, and then removing already clustered fragments from the stack. The algorithm terminates when no sequences are left. DOMO applies successive steps on the basis of similarity in amino acid composition, dipeptide composition, local sequence similarity, and multiple sequence alignment similarity to detect domain boundaries and then cluster the domains [18]. DOMO tends to propose longer regions than ProDom (Figure 1b). Picasso dissects and then clusters domain-like fragments by [57^{*}] defining close neighbours by pairwise BLAST, and then hierarchically merging the initial neighbours

through profile–profile comparisons. Domain borders are determined on the basis of overlapping maximal clusters (clusters that are not fully contained in any other cluster); unified families are defined as sets of clusters that share at least one common domain. GeneRAGE [41] detects multi-domain proteins through simple phylogenetic transitivity: if A is similar to B and C, and B is not similar to C, then A has at least two domains. The resulting fragments are clustered by single-linkage. Although GeneRAGE appeared adequate for bacterial genomes, its accuracy is insufficient for the complexity of eukaryotes [4*,43*].

Clustering with implicit domain information

A few methods that use graph theory attempt to avoid the explicit dissection into domains by embedding domain information into the clustering procedures. ProClust [44*] encodes partial alignments resulting from multi-domain proteins into the edge of similarity graphs. Instead of using symmetric Smith–Waterman scores corresponding to undirected edges, ProClust normalises the score by the length of the proteins, thus yielding two directed edges differentiated by protein length. The graphs are then partitioned into strongly connected components (SCCs) that constitute the final clusters. For about 55% of the data, the method is reported to achieve a high specificity (>99%) when tested against SCOP [19]. TRIBE-MCL expresses pairwise similarity through a particular matrix (Markov matrix) that is then clustered (by a Markov cluster algorithm) [43*]. The algorithm iterates over rounds of expansion and inflation to alter the matrix. Another graph-based method uses the normalised Ncut-algorithm to classify proteins through pairwise relations [58*]. The method was reported to reproduce COG families [30] accurately.

Conclusions

Grouping proteins into families is important both for biological and computational reasons. Expert-curated family databases such as Pfam [16*], TIGRFAMs [25], SMART [26] and COGs [30] have steadily increased their coverage of the protein universe over the past year. Obviously, these resources overlap to some extent (Figure 1a). Therefore, the first large-scale efforts toward integration of many resources are extremely important additions to the field of databases [31**,33*,59]. Structural genomics reveals the importance of identifying structural domains [4*,8,34,60–63]. Although the expert-driven family databases implicitly identify domains, the quality of this identification is quite mixed (Figure 1b). Furthermore, entire proteomes can currently only be clustered through automatic methods. Some methods try to avoid the domain problem by elaborate hierarchical clustering schemata [34,35,47*]. Others identify domains through alignments and then cluster these domains [17,18,41,56]. The first methods have been published that address the task of identifying domain boundaries directly [48*,49–53,55*]. Three methods implicitly combine domain-dis-

section and clustering through algorithms from graph-theory [43*,44*,58*]. At this point, most of these methods have not been compared with one another. A coarse-grained comparison suggests that some of the automatic methods may be able to compete with expert-driven annotations (Figure 1b). None of the existing clustering and domain-dissection methods appears to solve the problems conclusively. If we assume that structural domains constitute one candidate for ‘the atom of evolution’, we may hope to find the ‘final’ solution some day. Lupas and colleagues [64] speculated that proteins evolved through inserting and deleting fragments that are more like Blocks [13], PRINTS [15] or PROSITE [9–11] motifs than like structural domains. If true, methods that dissect proteins into domains on the basis of sequence similarity alone may be doomed to fail. Additional information, such as predicted secondary structure, may be needed to determine the domain border. One point is clear: we urgently need better tools to dissect proteins into domains and to cluster these domains.

Acknowledgements

Thanks to Henry Bigelow (Columbia University) for helpful comments and for critical proofreading. JL and BR were supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health (NIH). Last, but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Moore G: **Cramming more components onto integrated circuits.** *Electronics* 1965, **38**:114-117.
 2. Rost B: **Marrying structure and genomics.** *Structure* 1998, **6**:259-263.
 3. Moore G, Dillon P: **Chip “law” expands beyond its creator’s wildest expectations.** *Forbes* 2002, **25**(March):66.
 4. Liu J, Rost B: **Target space for structural genomics revisited.** *Bioinformatics* 2002, **18**:922-933.
 - Can we cluster sequence-space by grouping full-length proteins? The authors show that single linkage clustering is doomed to fail if proteins cannot be dissected into domains.
 5. Rost B: **Did evolution leap to create the protein universe?** *Curr Opin Struct Biol* 2002, **12**:409-416.
 6. Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol* 2002, **318**:595-608.
 7. Nair R, Carter P, Rost B: **NLSdb: database of nuclear localization signals.** *Nucleic Acids Res* 2002, in press.
 8. Nagano N, Orengo C, Thornton J: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.** *J Mol Biol* 2002, **321**:741.
 9. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219.
 10. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinf* 2002, **3**:265-274.
 11. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30**:235-238.

12. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R: **High-protein knowledge resource: SWISS-PROT and TrEMBL.** *Brief Bioinform* 2002, **3**:275-284.
13. 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.
14. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-197.
15. Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K *et al.*: **PRINTS and PRINTS-S shed light on protein ancestry.** *Nucleic Acids Res* 2002, **30**:239-241.
16. Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30**:276-280.
- Pfam is by far the most comprehensive manually curated protein family database. High quality is achieved by assuring the incorporation of accurate seed alignments and iterative refinement steps.
17. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains.** *Brief Bioinform* 2002, **3**:246-251.
18. Gracy J, Argos P: **DOMO: a new database of aligned protein domains.** *Trends Biochem Sci* 1998, **23**:495-497.
19. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomics.** *Nucleic Acids Res* 2002, **30**:264-267.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
21. Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM: **The CATH protein family database: a resource for structural and functional annotation of genomes.** *Proteomics* 2002, **2**:11-21.
22. Dietmann S, Holm L: **Identification of homology in protein structure classification.** *Nat Struct Biol* 2001, **8**:953-957.
23. Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH: **Comparison of sequence and structure alignments for protein domains.** *Proteins* 2002, **48**:439-446.
24. Yang AS, Honig B: **An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence.** *J Mol Biol* 2000, **301**:679-689.
25. 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.
26. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30**:242-244.
27. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29**:2994-3005.
28. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30**:52-55.
29. Vlahovicek K, Murvai J, Barta E, Pongor S: **The SBASE protein domain library, release 9.0: an online resource for protein domain identification.** *Nucleic Acids Res* 2002, **30**:273-275.
30. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22-28.
31. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**:37-40.
- InterPro provides a unified documentation resource for protein families, domains and functional sites by merging annotations from several motif/domain databases. All member databases are accessible and searchable through one common, intuitive interface.
32. Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC *et al.*: **The Protein Information Resource: an integrated public resource of functional annotation of proteins.** *Nucleic Acids Res* 2002, **30**:35-37.
33. Silverstein KA, Shoop E, Johnson JE, Retzel EF: **MetaFam: a unified classification of protein families. I. Overview and statistics.** *Bioinformatics* 2001, **17**:249-261.
- A unified protein family classification is built automatically on the basis of 10 original databases using set-theory. By combining different resources, proteins are added to families and conflicting annotations are identified.
34. Portugal E, Kifer I, Linal M: **Selecting targets for structural determination by navigating in a graph of protein families.** *Bioinformatics* 2002, **18**:899-907.
35. Krause A, Haas SA, Coward E, Vingron M: **SYSTEMS, GeneNest, SpliceNest: exploring sequence space from genome to protein.** *Nucleic Acids Res* 2002, **30**:299-300.
36. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12**:85-94.
37. Nair R, Rost B: **Sequence conserved for sub-cellular localization.** *Protein Sci* 2002, in press.
38. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**:1113-1143.
- Thorough overview of the structural backgrounds of enzymatic activity. All 31 enzyme super-families investigated exhibit functional diversity generated by local sequence variation and domain shuffling. Commonly, substrate specificity is diverse across a super-family, whilst the reaction chemistry is maintained.
39. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends Genet* 2001, **17**:429-431.
- The authors estimate the magnitude of possible annotation errors in the automatic transfer of functional classification. They conclude that the number of potential errors in the prediction of detailed functions is higher than is usually believed.
40. 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.
41. Enright AJ, Ouzounis CA: **GeneRAGE: a robust algorithm for sequence clustering and domain detection.** *Bioinformatics* 2000, **16**:451-457.
42. Yona G, Linal N, Linal M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucleic Acids Res* 2000, **28**:49-55.
43. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**:1575-1584.
- A novel method using a Markov Cluster algorithm was used to assign protein sequences into families. Multi-domain proteins are accounted for implicitly. The method is relatively fast.
44. Bolten E, Schliep A, Schneckener S, Schomburg D, Schrader R: **Clustering protein sequences — structure prediction by transitive homology.** *Bioinformatics* 2001, **17**:935-941.
- The authors present a graph-theory-based clustering algorithm. Length-normalised alignment scores are expressed as directed edges in the graph, and clustering is based on the concept of strongly connected components (SCC).
45. 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.
46. Yona G, Levitt M: **Within the twilight zone: a sensitive profile-profile comparison tool based on information theory.** *J Mol Biol* 2002, **315**:1257-1275.

The authors present a novel approach to profile–profile comparisons. The resulting new method appears significantly more sensitive in detecting distant homologies than PSI-BLAST and IMPALA. The resulting method is applied to cluster all protein sequences in BioSphere.

47. Sasson O, Linial N, Linial M: **The metric space of proteins-comparative study of clustering algorithms.** *Bioinformatics* 2002, **18**(Suppl. 1):S14-S21.
- ProtoNet provides a hierarchical view of the protein universe. It starts with many small clusters with very high similarity, and merges and clusters at different similarity levels. Different merging rules and termination rules are explored and compared to achieve optimal results.
48. George RA, Heringa J: **Protein domain identification and improved sequence similarity searching using PSI-BLAST.** *Proteins* 2002, **48**:672-681.
- DOMAINATION delineates structural domain-like fragments through analyzing iterative PSI-BLAST alignments. The overall accuracy is estimated to be around 50% for a set of 453 multi-domain proteins.
49. Kulikowski CA, Muchnik I, Yun HJ, Dayanik AA, Zhang D, Song Y, Montelione GT: **Protein structural domain parsing by consensus reasoning over multiple knowledge sources and methods.** *Medinfo* 2001, **10**:965-969.
50. George RA, Heringa J: **SnapDRAGON: a method to delineate protein structural domains from sequence data.** *J Mol Biol* 2002, **316**:839-851.
51. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries.** *Bioinformatics* 2000, **16**:613-618.
52. Miyazaki S, Kuroda Y, Yokoyama S: **Characterization and prediction of linker sequences of multi-domain proteins by a neural network.** *J Struct Funct Genom* 2002, **2**:37-51.
53. Murvai J, Vlahovicek K, Szepesvari C, Pongor S: **Prediction of protein functional domains from sequences using artificial neural networks.** *Genome Res* 2001, **11**:1410-1417.
54. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW *et al.*: **Intrinsically disordered protein.** *J Mol Graph Model* 2001, **19**:26-59.
55. Liu J, Tan H, Rost B: **Loopy proteins appear conserved in evolution.** *J Mol Biol* 2002, **322**:53-64.
- Long regions with no regular secondary structure (NORS) are abundant in 30 entirely sequenced organisms, particularly in eukaryotes. These regions are evolutionary conserved, and active in protein–protein interactions. They may constitute candidates for structural domains.
56. Carter P, Liu J, Rost B: **PEP: Predictions for Entire Proteomes.** *Nucleic Acids Res* 2002, in press.
57. Heger A, Holm L: **Picasso: generating a covering set of protein family profiles.** *Bioinformatics* 2001, **17**:272-279.
- Picasso clusters the protein universe starting with all-against-all BLAST alignments. The BLAST alignments are then merged on the basis of profile–profile comparison and set theoretic concepts. Structural domain-like fragments are identified from the final multiple alignments.
58. Abascal F, Valencia A: **Clustering of proximal sequence space for the identification of protein families.** *Bioinformatics* 2002, **18**:908-921.
- The authors propose a clustering strategy derived from minimum cut algorithm in graph theory. Application of the method to a COG dataset gives a similar result to COG classification itself.
59. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P *et al.*: **InterPro: an integrated documentation resource for protein families, domains and functional sites.** *Brief Bioinform* 2002, **3**:225-235.
60. Vitkup D, Melamud E, Moulton J, Sander C: **Completeness in structural genomics.** *Nat Struct Biol* 2001, **8**:559-566.
61. Montelione GT: **Structural genomics: an approach to the protein folding problem.** *Proc Natl Acad Sci USA* 2001, **98**:13488-13489.
62. Hurlley JH, Anderson DE, Beach B, Canagarajah B, Ho YS, Jones E, Miller G, Misra S, Pearson M, Saidi L *et al.*: **Structural genomics and signaling domains.** *Trends Biochem Sci* 2002, **27**:48-53.
63. Frishman D: **Knowledge-based selection of targets for structural genomics.** *Protein Eng* 2002, **15**:169-183.
64. Lupas AN, Ponting CP, Russell RB: **On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?** *J Struct Biol* 2001, **134**:191-203.