Clustering and analysis of protein families
Evgenia V Kriventseva, Margaret Biswas and Rolf Apweiler *

Various sequence-motif and sequence-cluster databases have been integrated into a new resource known as InterPro. Because the contributing databases have different clustering principles and scoring sensitivities, the combined assignments complement each other for grouping protein families and delineating domains. InterPro and new developments in the analysis of both the phylogenetic profiles of protein families and domain fusion events improve the prediction of specific functions for numerous proteins.

Introduction
In the past few years, the technology of sequencing has developed to the stage at which the sequencing of a complete genome can be contemplated as a practical and routine possibility. The complete sequences of more than 55 genomes have been published and at least 100 more are known to be nearing completion. These projects produce large amounts of sequence data lacking experimental determination of the biological function of the predicted gene products. One challenge of the genome era is to predict molecular functions and biological roles for the predicted gene products. Most approaches for the tentative assignment of functions to predicted proteins are based on pairwise sequence similarity searches against known proteins using sequence comparison programs such as FASTA [1] and BLAST [2]. However, the currently used methods, especially if automated, have various drawbacks [3]. Many proteins are multifunctional multidomain proteins, for which the assignment of a single function results in loss of information and outright errors. Also, with more and more predicted proteins from genome projects being added to the protein sequence databases, the best hit in pairwise sequence similarity searches is frequently a hypothetical protein or one that is poorly annotated or simply has a different function; thus, the propagation of wrong annotation is widespread.

To overcome these and other known limitations of functional annotation based on pairwise sequence similarity searches, the use of resources concerning protein families and domains gains more and more importance. These resources allow the assignment of functions to uncharacterised or predicted proteins by selecting proteins that belong to the same group of proteins as a given uncharacterised protein, extracting the annotation shared by all functionally characterised proteins of this group and assigning this common annotation to the unannotated protein [4].

In recognition of the growing importance of protein family and domain resources, we will focus in this review on current developments in the clustering and analysis of protein families. We will start by considering printed reviews of protein families, move on to manually curated protein family and domain databases, and from there discuss sequence-cluster databases. It is also of importance to discuss resources that combine sequence alignments with structural information and to point to recent work on phylogenetic profiles, domain fusion events and their role in predicting functional interactions. We will end this review with a discussion of how to use these valuable resources for the assignment of molecular functions to uncharacterised proteins.

Protein profiles
Comprehensive and accessible information on major groups of proteins is provided by the Protein Profile series published by Oxford University Press (OUP) (Table 1). Each printed volume is focused on a single family or subfamily of proteins, and contains a wealth of information, coupled with an extensive bibliography. From a collaboration between the SWISS-PROT group at the European Bioinformatics Institute (EBI) and OUP, SWISS-PROT and TrEMBL protein sequence data [5] and alignments (Figure 1) for the protein families covered in published and forthcoming volumes have been made available in electronic form (http://www.ebi.ac.uk/sp/proteinprofiles/).

Databases of protein signatures for families, domains and sites
A number of databases that use different methodologies and a varying degree of biological information on well-characterised protein families, domains and sites to derive protein signatures are available and are used to characterise new protein sequences. There are two main approaches: sequence-motif methods and sequence-cluster databases.

Sequence-motif methods
The sequence-motif methods rely on multiple sequence alignments from which conserved regions are used to provide characteristic signatures. PROSITE [6] is the oldest of the sequence-motif databases and includes extensive documentation on many protein families, as defined by regular expressions or generalised profiles. Regular expression patterns describe a consensus in which all but the most...
significant residue information is discarded. The profiles or weight matrices take into account the variable regions between conserved motifs by defining which residues are allowed at given positions, which positions are highly conserved and which positions can tolerate insertions. Other databases in which proteins are grouped, using various algorithms, by sequence motifs include PRINTS [7], Pfam [8], SMART [9], BLOCKS [10] and eMotif [11].

PROSITE, PRINTS, Pfam and SMART all store motifs that are derived from searches of SWISS-PROT and TrEMBL. PRINTS motifs are fingerprints in which all the residue information is retained within a set of frequency matrices. Pfam and SMART use hidden Markov models (HMMs), probabilistic models derived from alignments. BLOCKS is based on InterPro entries, with cross-references to PROSITE signatures, and stores motifs as aligned, clustered, untagged blocks. eMotif stores regular expressions derived automatically from motifs in BLOCKS and PRINTS. These databases have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. For example, regular expressions are likely to be unreliable in the identification of members of highly divergent superfamilies; fingerprints perform relatively poorly in the diagnosis of very short motifs; and profiles and HMMs are less likely to give specific subfamily diagnoses. Although all of the resources share a common interest in protein sequence classification, some focus on divergent domains (e.g. Pfam), some focus on functional sites (e.g. PROSITE) and others focus on families. These last specialise in hierarchical definitions from superfamily down to subfamily levels in order to pinpoint specific functions (e.g. PROSITE). The last specialise in hierarchical definitions from superfamily down to subfamily levels in order to pinpoint specific functions (e.g. PROSITE). The last specialise in hierarchical definitions from superfamily down to subfamily levels in order to pinpoint specific functions (e.g. PROSITE). These methods and their role in sequence analysis have been reviewed recently [12**]. As no one method alone is ideal, most analysis strategies endeavour to combine a range of secondary protein databases. InterPro [13**], an integrated documentation resource, was created to combine information from the PROSITE, PRINTS, Pfam, ProDom and, recently, SMART database projects. InterPro is a very useful resource for the computational functional classification of newly determined sequences (Figure 2) and for comparative analysis of whole genomes [14**]. InterPro has already been used for the
Sequences and topology

A section of the multiple sequence alignment for matrix metalloproteinasises (MMPs) that corresponds to the metal-binding site and active site of these proteins. Asterisks mark the three zinc-binding sites and the circumflex marks the active site. The SWISS-PROT and TrEMBL accession numbers are listed on the left. Data are from http://www.ebi.ac.uk/sp/proteinprofiles/. The protein families that are currently included in this data set are actins, unconventional collagens, helix-loop-helix transcription factors, lysosomal cysteine proteinases, MMPs, myosins, peptidyl-prolyl cis/trans isomerases and tyrosine phosphoprotein phosphatases. Additional families for forthcoming inclusion are the cadherins, EF-hand proteins, gelsolins, integrins and thermolysins.

Sequence-cluster databases
Unlike sequence-motif databases, sequence-cluster databases are derived automatically from sequence databases using different clustering algorithms. Because they do not depend on manual crafting and validation of family discriminators, these databases are relatively comprehensive, although the biological relevance of clusters can be ambiguous and can sometimes be an artefact of particular thresholds.

A group of family-cluster methods is connected to the Protein Information Resource (PIR) [17]. PIR-ALN is a database of protein sequence alignments derived from sequences and annotations in the PIR protein sequence database. Information about PIR protein families, protein superfamilies and homology domains can be found in ProtFam, a database including the alignments of all PIR superfamilies and homology domains [18]. The ProClass database is a resource organised according to family relationships, as defined collectively by PIR superfamilies and PROSITE patterns [19]. iProClass is a recent extension of ProClass.

Other sequence-cluster databases include ProDom [20], SBASE [21], DOMO [22], SYSTERS [23], ProtoMap [24] and CluSTr [25]. ProDom, SBASE and DOMO provide automated analysis of protein domains. ProDom assumes that the shortest full-length sequence in the SWISS-PROT and TrEMBL database is a single-domain protein and uses PSI-BLAST [26] to find homologous domains, which are then clustered in the same ProDom entry. DOMO uses the relative positions of homologous segment pairs within the same protein (for repeats) or within homologous proteins, with regard to the N or C terminus, to define domain boundaries.

SYSTERS, ProtoMap and CluSTr work with complete protein sequences. SYSTERS applies iterative gapped BLAST searches to cluster proteins in the SWISS-PROT and PIR databases. For each seed protein, all BLAST hits higher than a preset threshold are retained and the lowest scoring sequence is used for the next query. The process repeats until no new sequences above the cutoff value are found. In the ProtoMap methodology, high-resolution clusters are obtained by combining Smith–Waterman [27], FASTA and BLAST searches. A statistical algorithm is then used to identify groups of possibly related clusters. Within these groups, clusters are interactively merged when the similarity between the closest clusters is higher than the threshold. The CluSTr methodology uses the Smith–Waterman algorithm and Z-score statistics [28], which allows the data to be updated incrementally, avoiding time-consuming recalculations. CluSTr, SYSTERS and ProtoMap provide a hierarchical organisation of protein clusters by performing analysis at different levels of sequence similarity.

Structural alignment and cluster databases
Structural alignment databases combine protein sequence alignments with structural information obtained from the Protein Data Bank (PDB) [29]. HSSP (Homology-derived Secondary Structure of Proteins) [30], for example, is a database of the alignments of the sequences of proteins with known structure with sequences of all close homologues. The sequence-pattern-embedded discrete state-space models (pDSMs) [31] combine information about functionally conserved sequence patterns with information about structure context.

Structural domain databases attempt to align or cluster proteins at the three-dimensional level. MMDB, the three-dimensional structure data in Entrez [32], includes a complete comparison of all pairs of protein structures using an automatic procedure. CATH is a classification of protein domains [33]. Each protein structure in the PDB is cut into its constituent domains and each domain is classified separately. 3Dee [34] contains structural domain definitions for all protein chains in the PDB with 20 or more residues. In addition, the domains have been clustered on the basis of sequence similarity and structural similarity. The resulting families are stored as a hierarchy.

The Structural Classification of Proteins (SCOP) database [35] is constructed manually by the visual inspection and comparison of structures, but with the assistance of
automatic sequence and structure comparison tools. The classification is at the domain level for many proteins but, in general, a protein is only split into domains when there is a clear indication that the individual domains may have existed as independent proteins. FSSP (Families of Structurally Similar Proteins) [36] is a complete comparison of all pairs of protein structures in the PDB using an automatic procedure. It is the basis for the Dali Domain Dictionary, a structural classification of protein domains [37]. Domains are delineated automatically and similar structures are clustered hierarchically into fold classes or structural neighbourhood. The I-sites library of sequence structure motifs derived from HMMs for local sequence structure correlations in proteins is implemented to predict the local structure of a protein from its amino acid sequence [38,39]. A recent review compares the protein structure classifications of SCOP, CATH and FSSP [40••]. The authors conclude that, despite employing different methods and basing their systems on different rules of protein structure and taxonomy, SCOP, CATH and FSSP agree on the majority of their classifications.

**Phylogenetic classifications**

With the availability of complete proteomes, clustering in the phylogenetic space gains a lot of interest. Analysis of the phylogenetic profiles of protein families and of domain fusion events helps to predict many functional interactions and deduce specific functions for numerous proteins. A phylogenetic classification of proteins encoded in more than 34 complete genomes representing 26 major phylogenetic lineages can be found in the Clusters of Orthologous Groups of proteins (COGs) database [41]. Each COG consists of orthologous proteins or orthologous sets of paralogues from at least three species. The COGs are created by starting with an all-against-all comparison using gapped BLAST and identifying the obvious paralogues; finding triangles of orthologous proteins taking into account paralogues, and then merging triangles that share a side. An additional step is carried out to split COGs that were merged because of the existence of multidomain proteins. Another comparative genomics approach that recently has received considerable attention is the analysis of protein and domain fusion (and fission) [42•,43•,44,45]. The basic assumption is straightforward: a fusion is only maintained during evolution when it facilitates kinetic coupling of consecutive enzymes in pathways or other forms of functional interactions between proteins. Therefore, those proteins that are fused in some species are likely to interact, physically or at least functionally, in other organisms. Such proteins have been called ‘Rosetta stone’ proteins because they might be the linking element between proteins or domains with known function and those with so far unknown function. Using this methodology, a number of protein–protein interactions have been predicted [42•,43•]. Phylogenetic profiles also
can enable researchers to elucidate the subcellular location [46].

Conclusions

Very recently, some major advances in the clustering and analysis of protein families have occurred. InterPro, which integrates various sequence motif and cluster databases (PROSITE, PRINTS, Pfam, and ProDom), and the new algorithms for the analysis of both the phylogenetic profiles of protein families and domain fusion events are very powerful resources for the computational functional classification of newly determined sequences and the comparative analysis of whole genomes. The potential of these new methods will be of even more value for the life science community if they are combined with the work of the Gene Ontology (GO) consortium [47**]. The goal of the GO consortium is to produce a dynamic, controlled vocabulary for biological process, molecular function and cellular component. The GO ontologies produce a controlled vocabulary that can be used to ensure dynamic maintenance and interoperability among genome databases. GO, combined with the new developments in the clustering and analysis of protein families that we have tried to illustrate in this review, will significantly improve the prediction of specific functions for numerous proteins. This will eventually lead to a better understanding of the biology of living cells, now achievable because of the accelerating availability of molecular sequences of complete genomes.

Acknowledgements

This work was supported, in part, by grant B104-CT97-2099 from the European Commission.

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


A very interesting review focusing on databases of protein signatures for families and domains. The authors provide an excellent overview of the current status of these databases, outlines the methods behind them and discusses their diagnostic strengths and weaknesses.


The authors describe the creation of InterPro, a new integrated documentation resource for protein families, domains and functional sites that was developed as a means of rationalising the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database projects. Merged annotations from PRINTS, PROSITE and Pfam form the InterPro core. Each combined InterPro entry includes functional descriptions and literature references, and links are made back to the relevant member database(s), allowing users to see at a glance whether a particular family or domain has associated patterns, profiles, fingerprints and so on. In the task of sequence characterisation, InterPro will provide a more reliable, concerted method for identifying protein family traits and for inheriting functional annotation. This is especially important given the dependence on automatic methods for assigning functions to the raw sequence data issuing from genome projects.


This paper presents an extensive comparative analysis of the eukaryotic genomes of S. cerevisiae, C. elegans and recently sequenced D. melanogaster. The authors analysed the ‘core proteome’ (the number of distinct protein families each genome encodes), the distribution of genes encoding these protein families and the number of shared genes among flies, worms, yeast and mammals. An InterPro-based analysis of the protein family composition and the organisation of protein domains within the proteomes of fly, worm and yeast is also performed. The analysis of genes that have been directly implicated as causative agents of human diseases has shown that the fly has orthologues of 177 of the 289 human disease genes searched for. In addition, some fundamental cellular processes, such as the cell cycle, cell structure, cell adhesion, cell signalling, apoptosis, neuronal signalling and the immune system, underwent comparative analysis.


22. Gracy J, Argos P: Automated protein sequence database classification. I. Integration of compositional similarity search,


A very well conceived study of several methods of structural classification. This review provides an excellent systematic comparison of three unique methods of classifying protein structures: purely manual, a combination of manual and automated, and purely automated.


This paper discusses a computational method for inferring protein-protein interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologues fused into a single protein chain in another organism. Analysis of sequences from various genomes revealed 6899 such putative protein-protein interactions in E. coli and 45,502 in yeast. On the basis of three independent tests, many members of these pairs were confirmed as functionally related.


This paper presents a sequence comparison method that identifies gene fusion events in complete genomes. This method is used to predict functional associations of proteins. 215 genes or proteins in the complete genomes of E. coli, H. influenzae and M. jannaschii are shown to be involved in 64 unique fusion events.


The authors describe the creation of a dynamic, controlled vocabulary that can be applied to all eukaryotes, even as knowledge of gene and protein roles in cells is accumulating and changing. To this end, the Gene Ontology (GO) consortium has constructed three independent ontologies: biological process, molecular function and cellular component. The GO concept is intended to make possible, in a flexible and dynamic way, the annotation of homologous and gene protein sequences in multiple organisms using a common vocabulary that results in the ability to query and retrieve genes and proteins on the basis of their shared biology. The GO ontologies produce a controlled vocabulary that can be used to ensure dynamic maintenance and interoperability among genome databases.