

# CLASSIFICATION SCHEMES FOR PROTEIN STRUCTURE AND FUNCTION

Christos A. Ouzounis\*, Richard M. R. Coulson\*, Anton J. Enright<sup>‡</sup>, Victor Kunin\* and José B. Pereira-Leal\*

We examine the structural and functional classifications of the protein universe, providing an overview of the existing classification schemes, their features and inter-relationships. We argue that a unified scheme should be based on a natural classification approach and that more comparative analyses of the present schemes are required both to understand their limitations and to help delimit the number of known protein folds and their corresponding functional roles in cells.

## METRIC

A criterion or set of criteria that are stated in quantifiable terms.

## DISTANCE-BASED

### HIERARCHICAL CLUSTERING

Clustering is the process of grouping objects on the basis of their similarity. Distance-based hierarchical clustering is used to construct a tree of nested clusters on the basis of the proximity (or distance) between data points.

\**Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK.*

<sup>‡</sup>*Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York 10021, USA. Correspondence to C.A.O. e-mail: ouzounis@ebi.ac.uk doi:10.1038/nrg1113*

Biology has a long tradition of classification, which is the definition and naming of groups<sup>1</sup>. Classification is based on the comparative analysis of groups, which was pioneered by Aristotle (384–322 BC) who, for the first time, understood the challenge of grouping living organisms into meaningful classes on the basis of their anatomy and physiology<sup>2</sup>. This tradition was rediscovered and further expanded 20 centuries later by a series of great naturalists, including John Ray (1628–1705), Carl von Linné (1707–1778), Jean Baptiste de Monet, Chevalier de Lamarck (1744–1829) and Charles Darwin (1809–1882)<sup>2</sup>. This long and brilliant tradition of natural classification schemes for biological species had a significant impact on the modern molecular biology of the last century, in particular on the organization and further classification of molecular systems, including metabolic pathways and protein families or folds.

The recent revolution in high-throughput technologies in molecular biology has produced vast amounts of information about the structure, function and evolution of biological macromolecules at the genome scale<sup>3</sup>. To deduce possible clues about the action and interactions of these molecules in the cell, it is necessary to classify them into meaningful categories that are collectively linked to existing biological knowledge. In technical terms, clustering and classification refer to certain elements to be classified (for example, protein structure and function), the features (or attributes) of these

elements (for example, fold types or enzyme properties) that are used for the classification, the definition of a METRIC (similarity or distance, which are derived on the basis of the features), a set of algorithms that generate the metric and perform clustering (for example, DISTANCE-BASED HIERARCHICAL CLUSTERING) and, finally, the interpretation of intra- and inter-cluster relationships, which are usually tightly linked to the performance evaluation of the entire procedure.

Several recent reviews on the classification of proteins<sup>4–6</sup> discuss the properties and contents of a set of resources for the analysis of protein structure. Here, we provide a critical assessment of these resources and expand the scope by including an overview of the functional classifications of proteins. Our main point is that the scope of all classification schemes is similar and that a more rigorous comparative analysis of these schemes might ultimately lead to a single widely acceptable taxonomy of molecular types and families for all living organisms.

## Approaches to classification

We begin by describing in detail several approaches for the classification of proteins and their functions that have been developed over the past few decades. We provide both the original and latest citations for each classification or resource, to give an indication of how long the corresponding project has been established.

Table 1 | Resources for structural classifications of proteins

Scheme	Description	Automation (algorithm)	Type	Comments
SCOP	Structural classification of proteins	Low	Fold level and below	High quality, highly curated, possible problems with scalability
CATH	Class, architecture, topology, homology	Low/medium (SSAP)	Fold, topology similarity	Comprehensive, some degree of human intervention
FSSP	Fold assignment using DALI	High (DALI)	Fold, topology similarity	Highly automated, highly scalable
DSSP	Database of secondary structure of proteins	High (DSSP)	Secondary structure of proteins	The first algorithmic definition of secondary structure
HSSP	Homology and secondary structure of proteins	High (MaxHom)	Secondary structures plus sequence alignments	One of the first databases of sequence alignments
Pfam	Domain-level classification of proteins	Medium (HMMER)	Domain database, sensitive detection of homologues by hidden Markov models	Human intervention required for the seed set, highly comprehensive collection of domains
PRINTS	Fingerprint information for protein sequences	Low	Motif database, multiple levels of sequence similarity	Human intervention required, limited set with specific families, highly curated
SMART	Mobile domains in proteins	Low/medium	Short motif database	Impressive collection of motifs, especially for nuclear, signalling and extracellular proteins
PROSITE	Motif definition	Low	Pattern and sequence profile database	Contributions by the community, frequently updated
TIGRFAMS	Protein family database	Medium (HMMER)	Annotated protein family database, pointers to gene ontology	Concentrates on genomes published by TIGR
PRODOM	Protein domain database	Medium (BLAST)	Protein domain database	Recent versions encompass completed genome sequences
BLOCKS	Multiple-alignment blocks	High (BLIMPS, LAMA, BLOCKSSEARCH)	Multiple-alignment database	Seed set supplied by PROSITE has been extended to encompass sequences from other databases
eMOTIF	Protein motif database derivative	High (MOTIFMAKER)	Motif database, derived from PRINTS and BLOCKS	Powerful pattern definition, although database has not been updated recently
Bio-Dictionary	Pattern-based classification	High (Teiresias)	Pattern discovery and detection	Highly scalable, unsupervised motif discovery, high coverage
SYSTEMS	Protein families	Medium	Protein sequence family database	Comprehensive resource, recently extended to encompass expressed sequence tags
ClustR	Clusters of related proteins	High (Smith-Waterman dynamic programming)	Protein family database	Highly comprehensive, extensively cross-referenced, computationally demanding construction of protein families
COGS	Clusters of orthologous groups	Low/medium, BLAST similarity	Attempt to construct lists of orthologous genes across genomes	COGS (families) classified into functional classes, difficult to scale up
ProtoMap	Hierarchical classification of proteins	Medium/high	Protein families detected by a graph representing sequence-similarity relationships	Hierarchical clustering of SwissProt
MetaFam	A meta-database of protein families	Medium/high	An integrated database for several protein family databases	Detailed schema and powerful query capabilities
TRIBES	Protein family database	High (TRIBE-MCL)	Protein sequence family database for complete genomes and SwissProt	Scales well owing to an efficient clustering algorithm, problems with high granularity

**Fundamental distinctions.** Classification schemes can be broadly divided into curated versus automatic, and structural versus functional, in a continuum of semi-automatic and structure-to-function types of approach. Here, we define this distinction more precisely, because it is useful for our analysis, comparison and evaluation of these different approaches.

**Curation versus automation.** The curated classifications take into account human expertise, guided by computer analyses, to identify similarities between proteins

and assign them to particular groups. The automatic classifications rely on the execution of an algorithm that generates metrics for similarity or distance, which are subsequently processed to identify these groups. One advantage of curation is the high quality of the clustering; however, the disadvantage is that the end result might not be reproducible and scalable to high volumes of incoming data. Conversely, automation might generate more inaccurate assignments; yet, this approach is fully reproducible and should be scalable, given sufficient computational resources. Ideally, automation should be

Table 2 | Resources for functional classifications of proteins

Scheme	Description	Automation (algorithm)	Type	Comments
EC	Enzyme classification	Low	Enzyme Commission hierarchical classification	Refers to reaction types, not explicitly to proteins
YPD*	Functional classification for yeast proteins, also localization	Low	Species-specific function classification database	Derivatives are species specific
SGD	Functional classification for yeast proteins, also localization, mutants and so on	Low	Species-specific function classification database, community based	One of the most comprehensive species-specific resources available so far
MIPS	Munich Information Center for Protein Sequences function database	Medium (PEDANT)	Functional classifications derived manually, highly curated, inferred by similarity	Initially a database of functional analysis of the yeast genome, later expanded to encompass many other species
WIT/ERGO*	Acronym for 'what is there?'	Low/medium	Functional associations discovered by gene clusters	Encompasses a wide range of genomes
STRING	Search tool for the retrieval of interacting genes/proteins	Medium/high	Functional associations discovered by gene clusters, gene fusions and phylogenetic profiles	Server provides seamless access to the principal methods for the detection of associations using genome constraints
AllFuse	Differential fusion	High (DiffFuse)	Functional associations discovered by gene clusters, gene fusions and phylogenetic profiles	Scalable, computation intensive
Predictome	Putative functional links between proteins	Medium/high	Functional associations discovered by computational methods	Contains information about conserved gene clusters, gene fusions and phylogenetic profiles
Riley	Riley's functional classification scheme for <i>Escherichia coli</i>	Low	Details a number of functional classes, explicit classifications of gene products	The first functional classification of an entire genome, that of <i>E. coli</i>
GeneQuiz	Large-scale automatic annotation system	High (EUCLID)	Shallow hierarchy of three superclasses and fourteen classes	Highly automated, opts for low coverage and high precision, unique among this group
GO	Gene ontology	Low	A classification for molecular function, biological process and cellular component for <i>Drosophila melanogaster</i>	Expanded to encompass a large number of species
BioCyc	EcoCyc, MetaCyc and derivatives	Medium (PathTools)	A database architecture for genome and pathway information	Performs automatic metabolic reconstruction using genome-derived sequence annotations
KEGG	Kyoto encyclopaedia of genes and genomes	Low	An integrated database of gene and metabolic pathway information for many species	Contains implicit information about functional associations of genes in terms of pathway and reaction participation
DIP	Database of interacting proteins	Low/medium	A collection of interacting proteins from different species, highly curated, aided by some text mining from Medline abstracts	The first protein-interaction database, already includes thousands of entries. Not clear if scalable, but certainly a good gold-standard for computation with protein-interaction data
YPL.db	Localization database for yeast	Low/medium	A database of localization for yeast proteins, includes images	Not densely populated at present, ultimate goal is to annotate all yeast proteins
TRIPLES	Transposon-insertion phenotypes, localization and expression in <i>Saccharomyces</i>	Low	Information on mutants and subcellular localization for the yeast genome	Impressive contents for mutation analysis
MINT	Molecular interactions database	Low	A molecular-interaction database, containing information both for genetic and protein interactions	Not densely populated at present
BIND	Biomolecular interactions database	Low/medium	Detailed schema for molecular interactions	A sophisticated database schema for molecular interactions, not highly populated, scalable
PIM*	Protein interaction manager	Medium (PIMRider)	Prototyped to facilitate analysis of two-hybrid interaction analysis of <i>Helicobacter pylori</i>	Probably extensible to other species, geared towards two-hybrid experiments
CellZome db*	CellZome data	Medium	Information extracted from TAP-MS experiments	Prototyped on the yeast genome, applicable to other organisms

\*Commercial, no access without registration and/or license fee. For further information on these resources see links in online links box. SGD, *Saccharomyces* Genome Database; TAP-MS, tandem affinity purification-mass spectrometry; YPD, Yeast Proteome Database.

the ultimate goal of all classifiers, building on experience from manual curation on a smaller scale. Indeed, many curation projects use a host of different algorithms that assist human experts in the classification process. Another more subtle distinction is that curated classifications are 'supervised', in the sense that the classes are usually defined by experts, whereas automatic classifications are usually 'unsupervised' — the number of classes is determined by the algorithm and is not necessarily fixed.

**Structure versus function.** The distinction between structure and function is less obvious. The traditional view has distinguished systems that classify protein sequences into sequence families, as distinct from the classification of protein structures into structure folds. However, in our view, this distinction is not fundamentally different: sequence families and structure folds represent different levels of the protein-structure hierarchy — the primary and tertiary level, respectively. According to this view, classifications that take into account secondary structure should also be considered as part of this approach. The defining factor for structure-based classifications is the ability to derive measures of molecular similarity exclusively on the basis of different levels of protein structure. The function classification systems do (or should) not, in principle, take structure into account. They draw on the features that characterize the individual elements from other lines of evidence, including experimental information about participation in specific pathways, networks or phases of the cell cycle, sets of constraints obtained from genome structure and evolution, textual information in terms of supporting literature and database records, or simply *a priori* schemes of biological roles for macromolecules. In summary, function classification should be independent from structure attributes, and allow structurally (or evolutionarily) unrelated molecular types to be assigned to the same functional class on the basis of shared cellular roles or other features.

### Overview of classification schemes

Below, we describe the classification schemes for protein structure and function, and simultaneously assess their particular features according to the distinctive elements mentioned earlier. We then present some examples of proteins that have been classified with different levels of success, and explore possible avenues for future improvement and integration of these classifications. Our examples include cases for individual structure or (where possible) function associations against the available schemes.

**Structural classifications.** Structural classifications derive groups on the basis of molecular similarity in terms of primary or tertiary structure (TABLE 1). Three of the most widely known tertiary structural classifications are the Structural Classification of Proteins (SCOP)<sup>7,8</sup>, the Class /Architecture/Topology/Homology (CATH) database<sup>9,10</sup> and the Fold Classification based on Structure–Structure Alignment of Proteins (FSSP) database<sup>11,12</sup>.

SCOP is based on the visual inspection of protein folds and manual curation of the corresponding groups.

This is a hierarchical classification with different levels, folds (protein domains of similar topology and structure without detectable sequence similarity), superfamilies (similar structures with weak sequence similarity) and families (in cases in which sequence similarity is readily detectable)<sup>8</sup>.

CATH follows a similar principle and classifies proteins according to four main classes ( $\alpha$ ,  $\beta$ ,  $\alpha$ - $\beta$  and 'other' in terms of secondary-structure content), multiple architectural classes (based on the orientations of secondary-structure elements ignoring their sequence connectivity), topology (in which sequence connectivity is taken into account) and finally, homology classes (with detectable sequence similarity)<sup>10</sup>. CATH relies both on manual inspection and the SSAP structure-comparison algorithm<sup>13</sup>.

Finally, the FSSP database relies on an exhaustive all-against-all comparison of protein structures<sup>12</sup>, which is performed by the DALI algorithm<sup>14,15</sup>. Of these three databases, FSSP is the most automated and possibly most scalable, whereas SCOP contains the highest amount of manual curation. These databases allow the detection of common and unusual folds in protein-structure space<sup>16,17</sup> and represent invaluable tools for STRUCTURAL GENOMICS<sup>18</sup>.

Other databases of protein structure include the Dictionary of Secondary Structures of Proteins (DSSP) database<sup>19</sup> and the Homology-derived Secondary Structure of Proteins (HSSP) database<sup>20,21</sup>. DSSP uses a set of three-dimensional (3D) coordinates as input and defines the secondary-structure elements on the basis of hydrogen-bonding patterns and other geometrical features<sup>19</sup>. HSSP uses a POSITION-WEIGHTED DYNAMIC PROGRAMMING METHOD for sequence profile alignment, which is called MaxHom<sup>20</sup>. By aligning protein sequences with definitions of secondary structure that are derived from DSSP, HSSP enhances the set of known structures by adding sequence-similarity information<sup>21</sup>.

The primary structure (protein sequence) classification schemes focus on the detection of homologues in sequence databases and the accurate definition of protein families. Some of the best-known classifications include motif or domain databases — for example, the Protein Families database (Pfam)<sup>22,23</sup>, Protein Fingerprints (PRINTS)<sup>24,25</sup>, the Simple Modular Architecture Research Tool (SMART)<sup>26,27</sup>, the PROSITE database of protein families and domains<sup>28,29</sup>, the TIGRFAMS protein-family database<sup>30,31</sup>, the ProDom database<sup>32,33</sup>, the BLOCKs database<sup>34,35</sup> and the eMOTIF collection<sup>36,37</sup> — and protein-family databases, such as the Bio-Dictionary resource<sup>38</sup>, the SYSTEMS database<sup>39</sup>, ClustR<sup>40,41</sup>, COGS<sup>42,43</sup>, ProtoMap<sup>44,45</sup>, MetaFam<sup>46,47</sup> and TRIBES<sup>48</sup>.

The Pfam database is constructed by SEED ALIGNMENTS that are subsequently extended by using HIDDEN MARKOV MODEL (HMM) searches against the full protein database<sup>23</sup>. The PRINTS database is a highly curated database of motifs that are diagnostic of the functional properties of proteins at different levels of sequence similarity<sup>25</sup>. The SMART system describes mobile domains, which are principally found in eukaryotes<sup>26</sup>, on the basis of sensitive database searches and multiple alignment<sup>27</sup>.

#### STRUCTURAL GENOMICS

Initiatives to solve the structures of proteins that are encoded in an entire genome by high-throughput methods.

#### POSITION-WEIGHTED DYNAMIC PROGRAMMING

Dynamic programming is an algorithmic approach to solve sequential or multi-stage decision problems, such as finding optimal protein-sequence alignments. The position-weighted dynamic-programming method incorporates a matrix of substitution frequencies between amino acids, weighted by the degree of conservation of particular residues.

#### SEED ALIGNMENTS

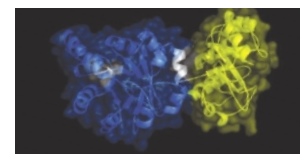
Hand-edited multiple sequence alignments that incorporate sequences that are described in the literature as belonging to the same family. From these seed alignments, hidden Markov models can be created that can in turn be used to search databases and identify new members of the family.

#### HIDDEN MARKOV MODEL

(HMM). A pattern-recognition approach that is used in bioinformatics for DNA/protein feature detection and sequence comparison. HMMs are based on transition probabilities for discrete states. These probabilities are usually derived from training sets such as seed alignments.

Box 1 | **Best practice example for protein annotation**

An archetype of consistent annotation of the well-characterized *Escherichia coli* protein TrpCF tryptophan biosynthesis enzyme IGPS-PRAI is shown below. This example represents a case of ‘best practice’ for protein annotation. The figure shows a three-dimensional view of the enzyme, in which the two domains are highlighted in different colours.



Database	Entry	Description
SWISSPROT	TRPC_ECOLI	Tryptophan biosynthesis protein trpCF [Includes: Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS); N-(5'-phospho-riboseyl)anthranilate isomerase (EC 5.3.1.24) (PRAI)]
INTERPRO, BLOCKS, E-MOTIF (IPB001468); PROSITE (PS00614); PROTOMAP (Cluster #737), Tribes (TR-0000000188), PFAM (PF00218); COG (COG0134)		Indole-3-glycerol phosphate synthase
PDB, CATH, SCOP, FSSP, DSSP, HSSP	1PII	N-(5'-phosphoribosyl)anthranilate isomerase synthase
PFAM (PF00697); COG (COG0135)	PRAI	N-(5'-phosphoribosyl)anthranilate (PRA) isomerase
ENZYME	4.1.1.48	Indole-3-glycerol-phosphate synthase. 1-(2-carboxyphenylamino)-1-deoxy- $\beta$ -D-ribose 5-phosphate = 1-(indol-3-yl)glycerol 3-phosphate + CO(2) + H(2)O
ENZYME	5.3.1.24	Phosphoribosylanthranilate isomerase: N-(5-phospho- $\beta$ -D-riboseyl)-anthranilate = 1-(2-carboxyphenylamino)-1-deoxy- $\beta$ -D-ribose 5-phosphate
PRINTS, TIGRFAMS, CLUSTR, PREDICTOME, MINT, BIND		No entries available
TRIPLES, YPLDB, CELLZOME		Species specific, not covering <i>Escherichia coli</i>

The PROSITE database is one of the oldest curated motif databases with contributions from the community<sup>28</sup>, which has now been extended to encompass a large number of patterns and sequence profiles<sup>29</sup>. The TIGRFAMS database is a collection of manually curated protein families that are identified by HMMs. It contains comments and functional classification assignments<sup>31</sup>. The ProDom database contains a comprehensive set of protein domain families that are automatically generated from sequence databases<sup>32</sup> and complete genome sequences<sup>33</sup>. These six motif or domain collections have been integrated into a single resource, the **InterPro** database<sup>49</sup>. The BLOCKS database contains a set of gap-free multiple alignments of protein families that have been contained in PROSITE<sup>34</sup>. The eMOTIF database is a derivative of the BLOCKS and PRINTS databases<sup>37</sup>, under a unified language for amino-acid similarity and pattern definition<sup>36</sup>.

The Bio-Dictionary resource detects the presence of sequence patterns using Teiresias, which is an unsupervised COMBINATORIAL PATTERN-DISCOVERY algorithm<sup>50</sup>, and subsequently allows the automated annotation of the proteins that contain these patterns<sup>38</sup>. The SYSTERS database performs automated clustering and classification of protein sequences into non-overlapping families and superfamilies<sup>39</sup>. ClustR represents the effort to hierarchically classify protein sequences into families using the Smith–Waterman dynamic-programming alignment algorithm at different levels of sequence similarity<sup>40,41</sup>.

COGS is a resource that documents the ORTHOLOGOUS genes across entire genomes<sup>42</sup>. ProtoMap is a comprehensive resource of protein families, which are created using similarity graphs that are generated from different sequence-similarity detection methods<sup>44,45</sup>. MetaFam is a resource that allows the querying and integration of various family databases<sup>46,47</sup>. Finally, TRIBES is a database of protein families<sup>48</sup> that are detected by the TRIBE-MCL algorithm for entire genomes and **Swiss-Prot**<sup>51</sup>.

**Functional classifications.** Functional classifications derive groups on the basis of functional similarity in terms of enzyme reaction mechanisms, participation in biochemical pathways, functional roles and cellular localization (TABLE 2). Following our definitions, we consider the sets of functional classifications that make no reference to structural similarities, to obtain a more balanced view of the classification approaches. In other words, we do not consider structural classifications with associated functional information, because related proteins often have related functions. Instead, we attempt to examine the available schemes for the classification of function independently from structural information.

One of the oldest classification schemes is the Enzyme Commission (EC) hierarchical classification, which defines six principal classes of enzymes<sup>52,53</sup>. This scheme groups reactions into categories with similar

COMBINATORIAL PATTERN DISCOVERY

An approach that produces all patterns in any given data set in an efficient way that avoids the explicit enumeration of the entire pattern space.

ORTHOLOGUES

Genes of common origin that have diverged through speciation rather than duplication. This term is sometimes ambiguously used to denote functionally equivalent genes that are of common origin in different organisms.

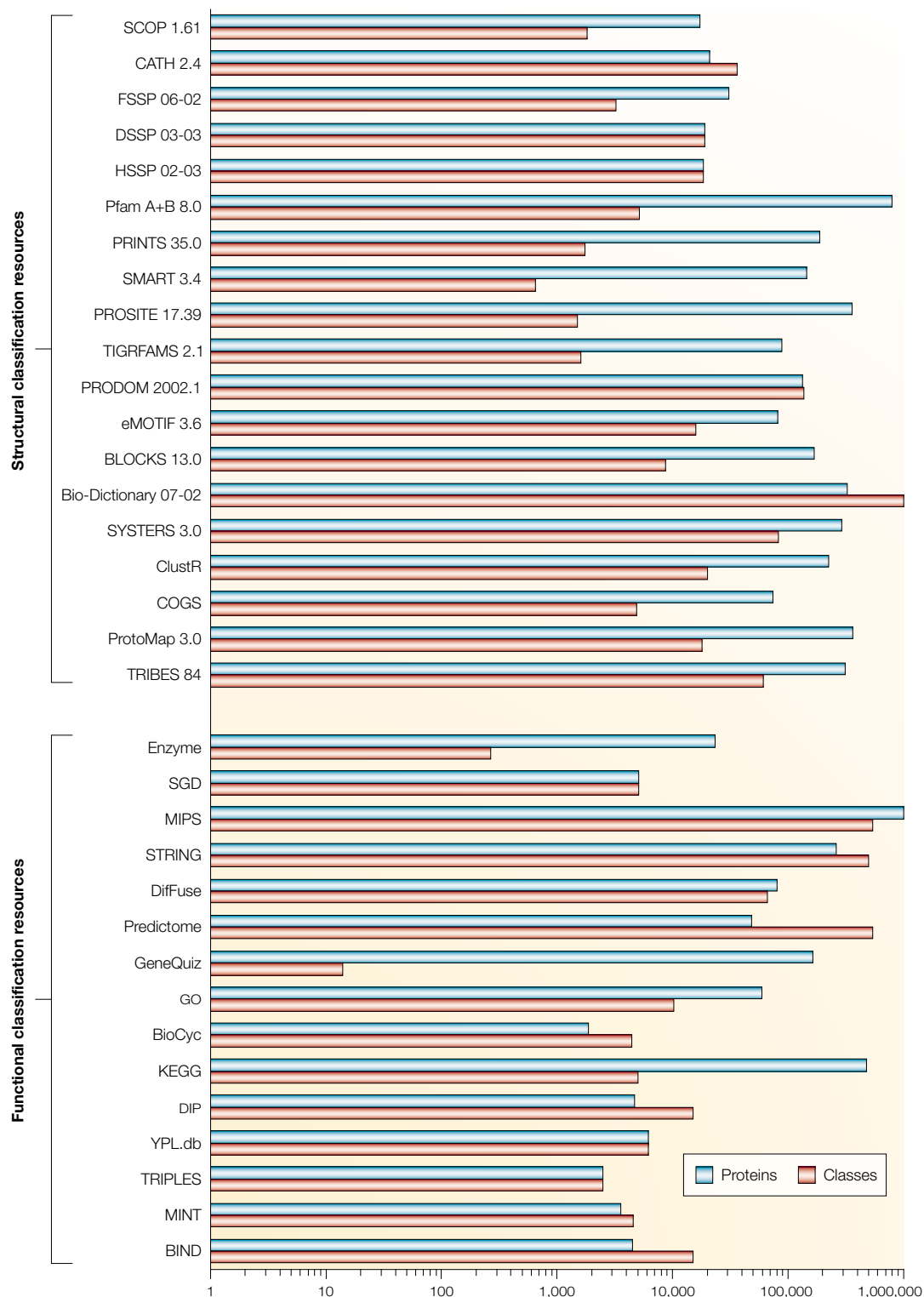


Figure 1 | **Coverage of protein sequence space by structural versus functional classifications.** Coverage of the known protein sequence space in terms of entries (blue) and classes (red) for structural and functional classification schemes. The x-axis corresponds to the number of entries in logarithmic scale, clipped at one million entries. The y-axis indicates the resource or classification scheme (version numbers, for example, SCOP 1.61, or release dates, for example, FSSP 06-02, are provided where possible). All structural classification schemes from TABLE 1, with the exception of MetaFam, are listed; only publicly available functional classification schemes are listed. The number of classes for functional classification schemes are not necessarily comparable, for example, ENZYME corresponds to the total number of three-level classifications, EcoCyc and KEGG classes correspond to enzymatic reactions, GeneQuiz and Gene Ontology (GO) contain functional classes, and protein interactions are listed as raw counts. It is clear by comparing the total number of protein entries (represented by the blue bars) that structural classifications provide a better coverage of protein sequence space than functional classifications, which are less automated.

properties, and further sub-classifies them with respect to reaction mechanisms, reactants and products, and specificities, by assigning them EC numbers. Although these widely used EC numbers are associated with proteins (enzymes), it is worth noting that they refer to reactions and not proteins (different proteins might have similar EC numbers and *vice versa*).

Another widely used database with a significant amount of curation is the Yeast Proteome Database (YPD)<sup>54,55</sup>, which was originally created for the *Saccharomyces cerevisiae* genome and later expanded to other species, before it was acquired by Incyte Genomics. The definitive public-domain resource for *S. cerevisiae* is, at present, the *Saccharomyces* Genome Database (SGD), which contains a wealth of information about yeast proteins<sup>56,57</sup>, including mutant and

subcellular localization information. Another significant resource that provides a widely used functional classification scheme is the Munich Information Center for Protein Sequences (MIPS)<sup>58,59</sup>. A similar resource was What Is There? (WIT)<sup>60</sup>, which provided information on conserved gene clusters and metabolic reconstructions. Similar to YPD, it was also commercialized, by Integrated Genomics, Inc., and renamed ERGO.

Several other resources provide genome-wide functional associations that are detected by comparative genomics<sup>3</sup>. The most comprehensive of those is the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)<sup>61,62</sup>, which was initially developed for conserved gene clusters<sup>61</sup> and later expanded<sup>62</sup> to cover gene fusions<sup>63</sup> and phylogenetic profiles<sup>64</sup>. Other more specialized databases that classify proteins

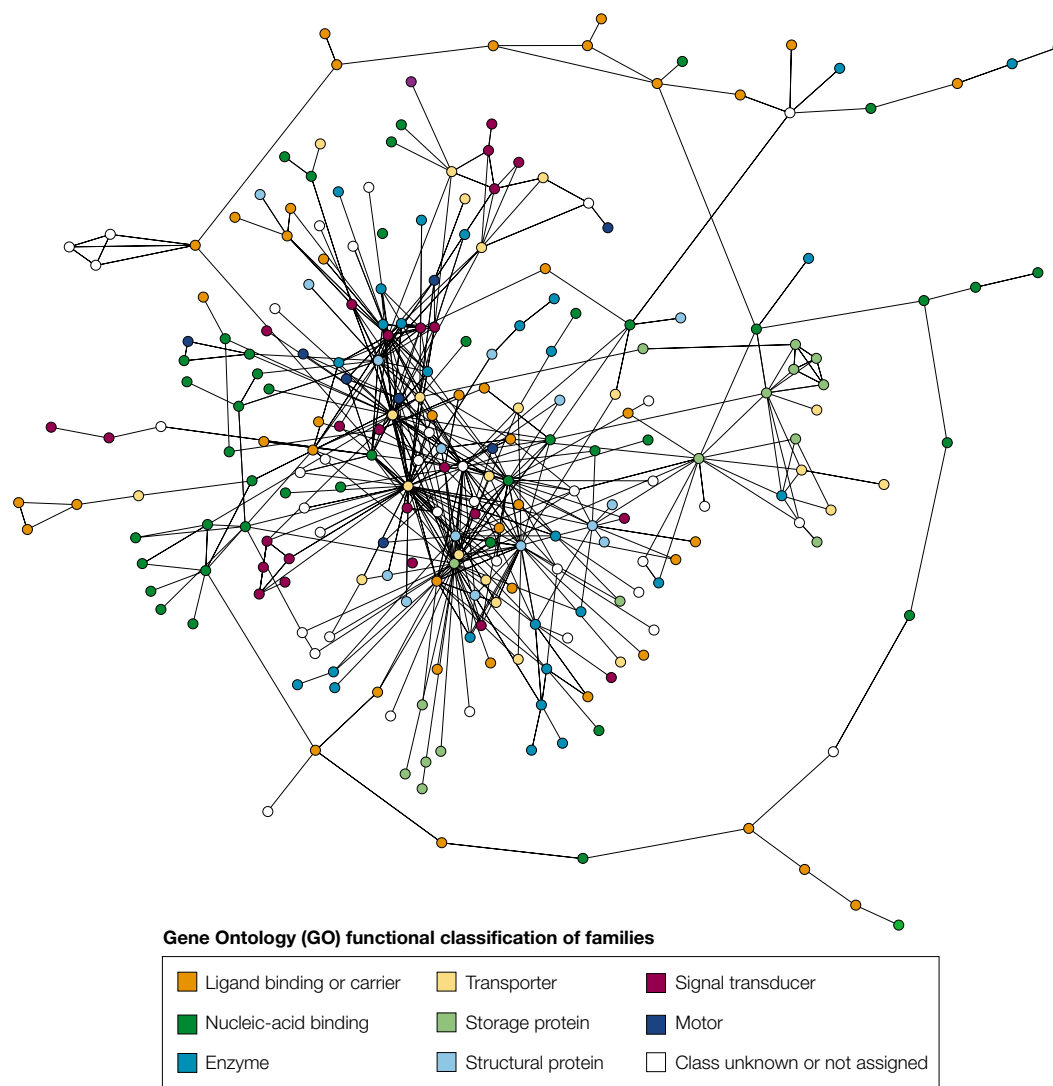
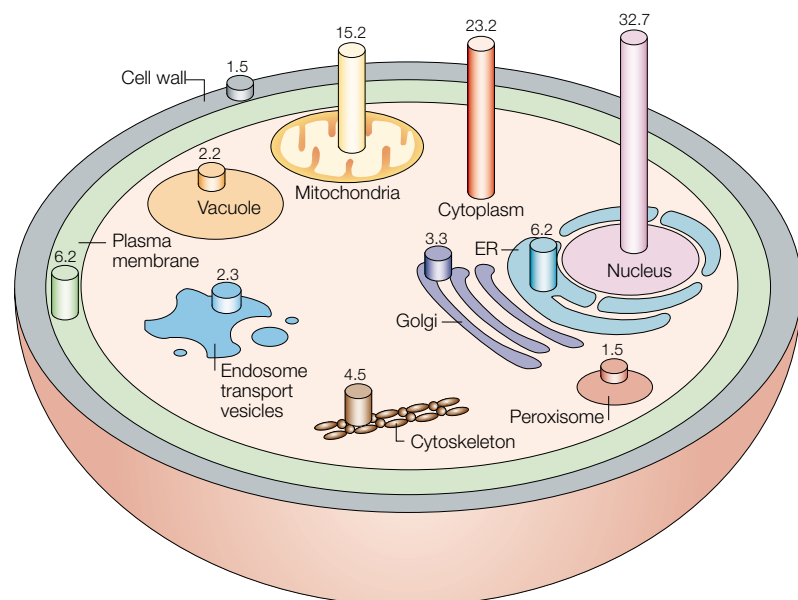


Figure 2 | **Overlaying structural and functional classifications for a group of interconnected proteins.** The largest interconnected group of protein families from the Swiss-Prot protein database (237 protein families; 21,727 sequences in total) is shown. Circles represent protein families. Lines show sequence similarities between families. Circles are coloured according to the GeneOntology functional classes<sup>73</sup> (where available). This is an example of two independent classifications, one based on sequence similarity (structural) and the other based on functional categorization. Figure modified with permission from REF. 51.



**Figure 3 | Relative frequencies of subcellular localization labels for *Saccharomyces cerevisiae* gene products of known function.** Subcellular localization represents a functional classification scheme that is independent of structural information, such as sequence similarity. Data obtained from the Munich Information Center for Protein Sequences (MIPS) Comprehensive Yeast Genome Database (CYGD)<sup>59</sup>. ER, endoplasmic reticulum.

into functionally associated groups are the AllFuse resource<sup>65</sup>, which is based on the detection of gene-fusion events<sup>66</sup>, and Predictome<sup>67</sup>, which is similar in content to STRING<sup>68</sup>.

One of the most influential schemes of *a priori* functional classification was a hierarchy of properties for the gene products of *Escherichia coli*<sup>69</sup>, which was later extended to include multifunctional classifications<sup>70</sup>. Inspired by this classification, a shallow hierarchy of protein function was devised for the automated genome-annotation system GeneQuiz<sup>71</sup>, based on the keyword mapping of protein families to 14 functional classes, called EUCLID<sup>72</sup>. Another complex classification scheme, which comprises three unique sub-schemes on molecular function, biological process and cellular component, is the Gene Ontology (GO) classification<sup>73</sup>.

A different way of classifying genes and proteins is by their participation or association with metabolic pathways. One such system is the EcoCyc database, an encyclopaedia of *E. coli* genes and metabolism<sup>74</sup>, which has also been used for metabolic predictions of *Haemophilus influenzae*<sup>75</sup> and other species by the incorporation of other pathways in the MetaCyc database<sup>76</sup>. Another important metabolic database is the Kyoto Encyclopaedia of Genes and Genomes (KEGG)<sup>77,78</sup>.

Finally, another set of implicit classifications can be derived from protein interaction and cellular-localization information. The groups of proteins that are defined as interacting can be viewed as classes in a functional classification scheme, the classes of which have not necessarily been defined. Resources here include the Database of Interacting Proteins (DIP)<sup>79,80</sup>, the Yeast Protein Localization database (YPL.db)<sup>81</sup>, the TRIPLES database<sup>82,83</sup>, the Molecular INTeraction database (MINT)<sup>84</sup>, the BIND database<sup>85,86</sup>, the Protein Interaction Manager (PIM) resource for two-hybrid experiments in *Helicobacter pylori*<sup>87</sup> and the CellZome yeast-interaction database<sup>88</sup>. Of the above, DIP, MINT and BIND are public-domain resources that allow the recording and annotation of protein interactions from various species, which greatly facilitates pattern discovery in protein-association networks<sup>89</sup>.

### General features of classifications

These classification schemes cover a vast spectrum of biological properties (TABLES 1,2). Structural classifications range from short motifs and promiscuous domains to full-length protein-sequence families to secondary-structure libraries, alignments and protein folds. Functional classifications range from cellular roles and localization to phenotype data to biochemical pathways and protein-interaction networks. Although this diversity is welcome in principle, it is not useful in the absence of a theme that will unite these classifications under one roof, as the scope of these schemes is, in fact, similar. Well-characterized proteins indicate the ultimate goal of all classification schemes: the generation of consistent and ideally complementary levels of structural and functional information (see BOX 1 for an example of best practice).

### Box 2 | Coverage of structure and function classification schemes

Examples of four randomly selected proteins from the *Chlamydia trachomatis* serovar D genome sequence<sup>97</sup> and their annotations. The coverage for the 20 structural and functional classification schemes is shown as a percentage. For example, in the case of CT313, one-half of the structural classification schemes list this protein, compared with only one-third of the functional classification schemes. It is evident from this short list that proteins from genome projects are more likely to be included in structural classification schemes than in functional schemes.

Protein ID	Consistent annotations (%)		Annotation
	Structural	Functional	
CT080	25	15	Late transcription unit B hypothetical protein
CT094	45	25	tRNA pseudouridine synthase EC 4.2.1.70
CT313	50	30	Transaldolase EC 2.2.1.2
CT664	45	25	FHA domain-containing protein

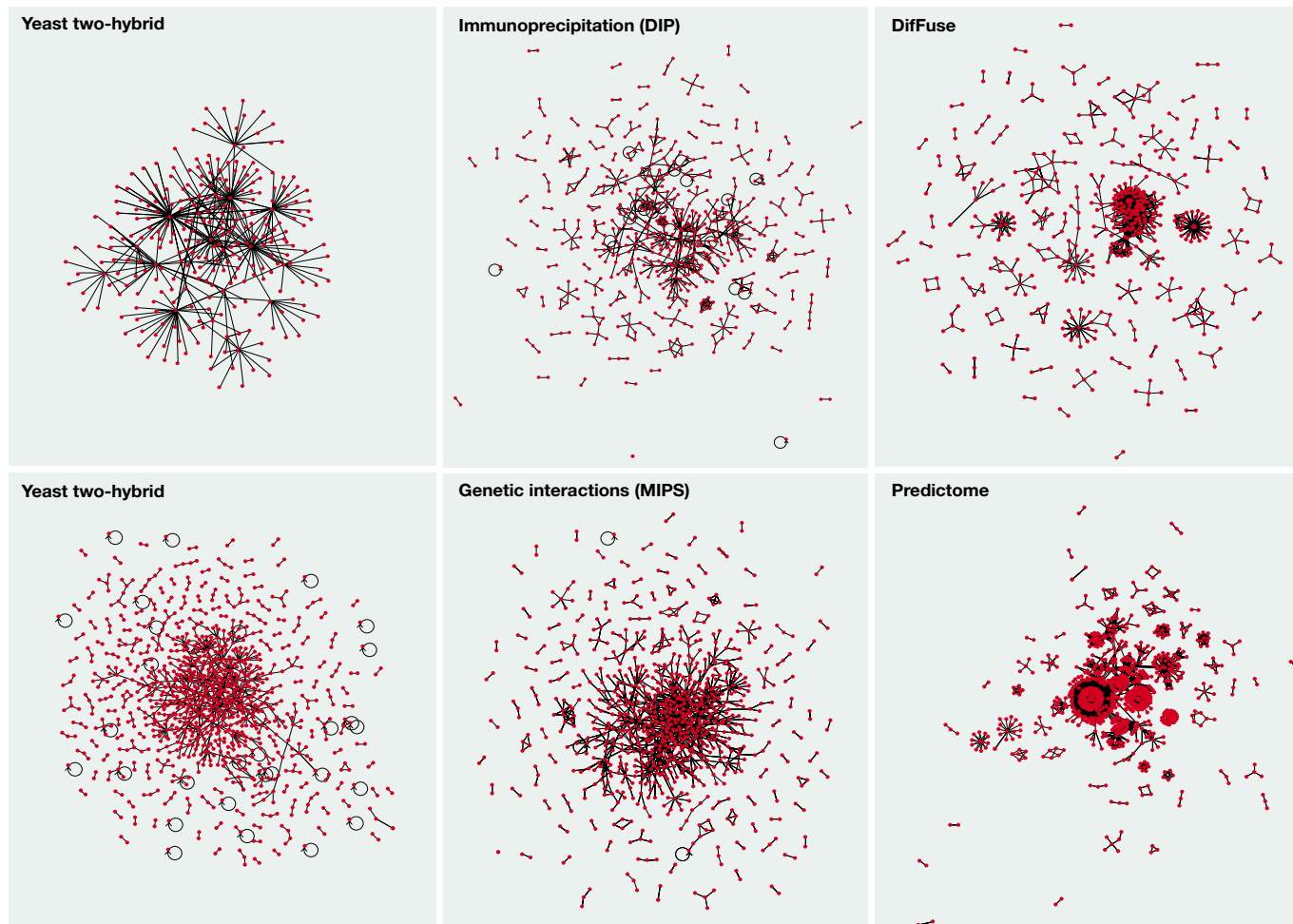


Figure 4 | **Topology and network structure of known protein interactions from yeast.** The different resources include two yeast two-hybrid experiments (upper left panel<sup>98</sup> and lower left panel<sup>99</sup>), immunoprecipitation or genetic interactions (middle), and computational predictions (right). Notice the diversity of these groupings, both in terms of the number of proteins involved in interactions and the structure of the relationships, which might be attributable to the low coverage of interaction space for the yeast cell. Lines represent interactions; circles represent proteins.

Certain elements of these classifications become apparent in this survey (TABLES 1,2). Structural classifications contain fold, motif and protein-family information, whereas functional classifications refer to biochemical and cellular roles, metabolic pathways, subcellular localization and molecular interactions. Evidently, the coverage of protein space might be different and sometimes difficult to obtain (FIG. 1). Most of these data collections were created more than a decade ago, a fact that also makes them susceptible to certain legacy requirements (such as the format of identifiers and modes of distribution).

Although structural classifications are probably easier to define on the basis of molecular-similarity criteria, their overlap is surprisingly limited<sup>6</sup>. Conversely, functional classifications encompass many processes and elements, which range from pathways to cellular compartments, and have been shown to overlap considerably with each other<sup>90</sup>. This is exemplified by four gene products that were randomly selected from the *Chlamydia trachomatis* serovar D genome<sup>91</sup> (BOX 2).

It is worth noting that structural classifications can be generated automatically because the similarity relationships are, in principle, obtained by algorithms that facilitate clustering and further annotation, whereas functional classifications are labour intensive, as they are often derived *a priori* or from large-scale experimentation. Consequently, structural classifications are comparatively more shallow owing to the limitations of algorithms to detect any relationships beyond the protein family (or motif) level<sup>92</sup>. By contrast, functional classifications, which are less constrained by this type of resolution limit, imply deep hierarchical classifications that potentially span all levels of molecular and cellular processes (FIG. 2).

It is also conceivable that, at least on a species-specific level, all high-throughput experiments result in a set that encompasses functional classifications on a genome-wide scale. According to our definitions, these associations of proteins are not a result of molecular similarity (that is structural classes), as they reflect the various biological processes as snapshots of cellular activity. For instance, experiments that provide information about

the subcellular localization of gene products should be considered as such classifications (FIG. 3). These resources should be viewed as individual classification schemes that will grow rapidly with time, reflecting the functional properties of molecular networks in the cell (FIG. 4). Not unlike crystallographic or NUCLEAR MAGNETIC RESONANCE (NMR) experiments that deduce the 3D structure of proteins — or DNA sequencing experiments that decipher the structure of genomes — large-scale experimentation for the detection of subcellular localization, protein interactions, phenotypic analysis and the like, is the basis for further, sometimes implicit, functional classifications of proteins.

The analogy goes further: in structural analysis, the experimental conditions (for example, temperature and pH) are strictly defined and perturbations involve the creation of artificial constructs (for example, by point mutation) that might affect molecular structure. The fact that molecular structure is robust to such changes allows the reliable inference of properties by structural similarity. The time-dependent element is the phylogeny of molecules, which extends to millions of years. By contrast, in functional analysis the experimental conditions are much harder to define because they involve the study of an entire biological system, not a DNA sequence or an atomic crystal. The molecule or crystal is at a state of low ENTROPY compared with a snapshot of cell physiology. This ‘entropy difference’ makes the comparison of functional experiments more difficult. Also, perturbations of physiological conditions involve system parameters that are difficult to control, with unexpected results. The time-dependent element is the life cycle or developmental time of the cell, which is orders of magnitude shorter than evolutionary time. Therefore, all snapshots of biological systems that use functional genomics provide a view of the underlying complexity of life that has not been available by the examination of molecular ‘relics’.

From a computational perspective, this time-dependent element makes functional classifications particularly prone to a lack of consistency, an issue that has already been discussed extensively in the literature<sup>93,94</sup>.

#### Challenges for classification schemes

Given the number of classification schemes, it would be desirable to unify them under a single theoretical framework. In practical terms, that would require the re-engineering of these resources and their integration, so that experimental biologists can find them more transparent and user-friendly. The problem, in our opinion, is that we lack the highly desirable fundamental conceptual framework for the classification of protein structure and function that would facilitate the creation of a single meaningful classification. Moreover, to achieve this goal, research into the comparative analysis of these schemes is required. Certain steps towards achieving the goals of unification and comparison have already been taken<sup>6,90</sup>.

**Natural classification, unification.** The first problem is the lack of natural classification schemes (most of which rely on directly derived metrics that do not necessarily have a physical meaning). For example, the

clustering of protein families cannot take the evolution of proteins into account because ‘molecular phylogeny’ can only be inferred. As another example, the clustering of gene-expression patterns cannot, at present, trace the production of the corresponding mRNAs in real time, which is a process that might be termed ‘molecular ONTOGENY’. Although the molecular ontogeny or phylogeny criterion is desirable, our present knowledge of these processes is insufficient and, therefore, the classifications are compromised by heuristic criteria that only partly reflect the physical reality.

It is an open problem to what extent the metrics that are derived by our algorithms reflect a natural process and, therefore, to what extent our final classifications reflect a more objective natural classification scheme. Note that a similar debate over the classification of species has been continuing for decades<sup>95</sup>. Once a natural classification scheme is obtained, the challenge of unifying these schemes under a single theoretical framework will be far more feasible.

**Comparative analysis.** The second problem is that the multitude of classifications requires consistency that has not yet been achieved, perhaps because this endeavour is still at an exploratory stage. To achieve a high level of consistency, more elaborate database schemas are required, which are known as ‘ONTOLOGIES’ in computer science. These schemas necessitate strict definitions of the group (or ‘class’) definitions (for example, what is a protein family?), in otherwise disparate database implementations (each scheme might have a different idea of what a protein family is). We argue that more research is required to realign the diverse collection of protein classifications and achieve a desirable consensus. Integration projects such as InterPro<sup>49</sup>, MetaFam<sup>46</sup> and GO<sup>73</sup> point in the right direction, although a rigorous theoretical framework is still missing.

#### Future perspectives

From the earlier discussion, it seems that there is a need for a meta-classification as a base on which more refined classifications (and ontologies) can be built. This classification must encompass all of the subtleties and definitions of existing classification schemes to ensure the communication and interchangeability of information between them, while also reflecting our genuine level of biological understanding.

The structural classification schemes are varied and disparate, although many use the same terminology, including motifs, domains and sequence families. It will be interesting to see whether a more formal ontology develops that encompasses these entities under formal definitions. In that sense, functional classifications have been more successful under GO and other systems, and have generated a wide-ranging dictionary that can be used across species and molecular processes. It is also interesting that many functional genomics experiments use functional classes from GO to annotate groups of genes without committing to more specific functional assignments<sup>96</sup>.

#### NUCLEAR MAGNETIC RESSONANCE

(NMR). An analytical chemistry technique that is used to study molecular structure and dynamics, which explores spectrum differences that are caused by the differential alignment of atomic spins in the presence of a strong magnetic field.

#### ENTROPY

A measure of the disorder or unavailability of energy within a closed system.

#### ONTOGENY

The development and life cycle of a single organism.

#### ONTOLOGY

An explicit formal specification of how to represent the objects, concepts and other entities within a domain of discourse, and the relationships among them. Ontologies are designed to create agreed vocabularies for exchanging information.

**Concluding remarks**

The wealth of structure information (protein sequence and fold) has been used to derive dynamic models (for example, molecular dynamics from mutants, evolutionary threading from folds, subfamily determinants from interactions and genome sequences from evolution). This field of exploring the evolutionary dynamics of proteins

will expand with the availability of further primary and tertiary structures of proteins and their functional properties. The most obvious applicability of dynamic models, however, comes from functional analyses, in which the short timescales and amenability to highly detailed experimentation render these systems ideally suited for the development of time-dependent models.

1. Ridley, M. in *Philosophy of Biology* (ed. Ruse, M.) 167–179 (Macmillan Publishing Co., New York, 1989).
2. Asimov, I. *A Short History of Biology* (Thomas Nelson & Sons Ltd., London, 1964).
3. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in the post-genomic era. *Nature* **405**, 823–826 (2000).
4. Swindells, M. B., Orengo, C. A., Jones, D. T., Hutchinson, E. G. & Thornton, J. M. Contemporary approaches to protein structure classification. *Bioessays* **20**, 884–891 (1998).
5. Heger, A. & Holm, L. Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.* **73**, 321–337 (2000).  
**A comprehensive analysis of strategies and resources for protein-sequence clustering and protein-family identification.**
6. Liu, J. & Rost, B. Domains, motifs and clusters in the protein universe. *Curr. Opin. Chem. Biol.* **7**, 5–11 (2003).  
**An overview of present methods for protein-sequence clustering.**
7. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
8. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264–267 (2002).
9. Orengo, C. A. *et al.* CATH: a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
10. Pearl, F. M. *et al.* The CATH database: an extended protein family resource for structural and functional genomics. *Nucl. Acids Res.* **31**, 452–455 (2003).
11. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698 (1992).
12. Holm, L. & Sander, C. Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316–319 (1998).
13. Orengo, C. A. & Taylor, W. R. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635 (1996).
14. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
15. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480 (1995).
16. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–602 (1996).
17. Brenner, S. E., Chothia, C. & Hubbard, T. J. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376 (1997).
18. Burley, S. K. & Bonanno, J. B. Structuring the universe of proteins. *Ann. Rev. Genomics Hum. Genet.* **3**, 243–262 (2002).
19. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
20. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68 (1991).
21. Dodge, C., Schneider, R. & Sander, C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313–315 (1998).
22. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).
23. Bateman, A. *et al.* The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280 (2002).
24. Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. PRINTS — a database of protein motif fingerprints. *Nucl. Acids Res.* **22**, 3590–3596 (1994).
25. Attwood, T. K. *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucl. Acids Res.* **31**, 400–402 (2003).
26. Schultz, J., Milpets, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
27. Letunic, I. *et al.* Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**, 242–244 (2002).
28. Bairoch, A. PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **19**, 2241–2245 (1991).
29. Falquet, L. *et al.* The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238 (2002).
30. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucl. Acids Res.* **29**, 41–43 (2001).
31. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucl. Acids Res.* **31**, 371–373 (2003).
32. Corpet, F., Gouzy, J. & Kahn, D. The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323–326 (1998).
33. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucl. Acids Res.* **28**, 267–269 (2000).
34. Henikoff, S. & Henikoff, J. G. Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565–6567 (1991).
35. Henikoff, S., Henikoff, J. G. & Pietrovski, S. Blocks-*r*: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**, 471–479 (1999).
36. Nevill-Maning, C. G., Wu, T. D. & Brutlag, D. L. Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA* **95**, 5865–5871 (1998).
37. Huang, J. Y. & Brutlag, D. L. The EMOTIF database. *Nucl. Acids Res.* **29**, 202–204 (2001).
38. Figoutos, I., Huynh, T., Floratos, A., Parida, L. & Platt, D. Dictionary-driven protein annotation. *Nucl. Acids Res.* **30**, 3901–3916 (2002).
39. Krause, A., Haas, S. A., Coward, E. & Vingron, M. SYSTEMS, GeneNet, SpliceNest: exploring sequence space from genome to protein. *Nucl. Acids Res.* **30**, 299–300 (2002).
40. Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M. & Apweiler, R. CluSTR: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucl. Acids Res.* **29**, 33–36 (2001).
41. Kriventseva, E. V., Servant, F. & Apweiler, R. Improvements to CluSTR: the database of SWISS-PROT+TrEMBL protein clusters. *Nucl. Acids Res.* **31**, 388–389 (2003).
42. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
43. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **31**, 28–33 (2003).
44. Yona, G., Linnal, N. & Linnal, M. ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* **37**, 360–378 (1999).
45. Yona, G., Linnal, N. & Linnal, M. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucl. Acids Res.* **28**, 49–55 (2000).
46. Silverstein, K. A., Shoop, E., Johnson, J. E. & Retzel, E. F. MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics* **17**, 249–261 (2001).
47. Shoop, E., Silverstein, K. A., Johnson, J. E. & Retzel, E. F. MetaFam: a unified classification of protein families. II. Schema and query capabilities. *Bioinformatics* **17**, 262–271 (2001).
48. Enright, A. J., Kunin, V. & Ouzounis, C. A. Protein families and TRIBES in genome sequence space. *Nucl. Acids Res.* (in the press).
49. Mulder, N. J. *et al.* The InterPro database, 2003 brings increased coverage and new features. *Nucl. Acids Res.* **31**, 315–318 (2003).
50. Rigoutos, I. & Floratos, A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* **14**, 55–67 (1998).
51. Enright, A. J., van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* **30**, 1575–1584 (2002).
52. Bairoch, A. The ENZYME data bank. *Nucl. Acids Res.* **22**, 3626–3627 (1993).
53. Bairoch, A. The ENZYME database in 2000. *Nucl. Acids Res.* **28**, 304–305 (2000).
54. Garrels, J. I. YPD — a database for the proteins of *Saccharomyces cerevisiae*. *Nucl. Acids Res.* **24**, 46–49 (1996).
55. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucl. Acids Res.* **27**, 69–73 (1999).
56. Cherry, J. M. *et al.* SGD: *Saccharomyces* Genome Database. *Nucl. Acids Res.* **26**, 73–79 (1998).
57. Dwight, S. S. *et al.* *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucl. Acids Res.* **30**, 69–72 (2002).
58. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **27**, 44–48 (1999).
59. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **30**, 31–34 (2002).
60. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucl. Acids Res.* **28**, 123–125 (2000).
61. Snel, B., Lehmann, G., Bork, P. & Huynh, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucl. Acids Res.* **28**, 3442–3444 (2000).
62. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucl. Acids Res.* **31**, 258–261 (2003).
63. Marcotte, E. M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
64. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).  
**This paper introduces the concept of phylogenetic profiles, and the idea that similar phylogenetic profiles indicate functional association between genes.**
65. Enright, A. J. & Ouzounis, C. A. Functional associations of proteins in entire genomes via exhaustive detection of gene fusion. *Genome Biol.* **2**, 0031–0037 (2001).
66. Enright, A. J., Iliopoulos, I., Kypides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
67. Yanai, I., Derti, A. & DeLisi, C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci. USA* **98**, 7940–7945 (2001).  
**This paper is a 'proof of principle' that gene-fusion events can be used to infer functional associations, as proposed in references 63 and 65.**
68. Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J. & DeLisi, C. Predictome: a database of putative functional links between proteins. *Nucl. Acids Res.* **30**, 306–309 (2002).
69. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952 (1993).  
**The original comprehensive functional-classification scheme, developed for the gene products of the *E. coli* genome.**
70. Serres, M. H. & Riley, M. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics* **5**, 205–222 (2000).
71. Andrade, M. A. *et al.* Automated genome sequence analysis and annotation. *Bioinformatics* **15**, 391–412 (1999).
72. Tamames, J., Ouzounis, C., Casari, G., Sander, C. & Valencia, A. EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics* **14**, 542–543 (1998).

73. Ashburner, M. A. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000). **This paper describes the development of a dynamic controlled vocabulary for the functional annotation of eukaryotic gene products.**
74. Karp, P. D., Riley, M., Paley, S. M. & Pellegrini-Toole, A. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **24**, 32–39 (1996).
75. Karp, P. D., Ouzounis, C. & Paley, S. HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 116–124 (1996).
76. Karp, P. D., Riley, M., Paley, S. M. & Pellegrini-Toole, A. The MetaCyc database. *Nucl. Acids Res.* **30**, 59–61 (2002).
77. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucl. Acids Res.* **30**, 42–46 (2002).
78. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* **27**, 29–34 (1999).
79. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucl. Acids Res.* **28**, 289–291 (2000).
80. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303–305 (2002).
81. Habeler, G. *et al.* YPL.db: the Yeast Protein Localization database. *Nucl. Acids Res.* **30**, 80–83 (2002).
82. Kumar, A. *et al.* TRIPLES: a database of gene function in *Saccharomyces cerevisiae*. *Nucl. Acids Res.* **28**, 81–84 (2000).
83. Kumar, A. *et al.* The TRIPLES database: a community resource for yeast molecular biology. *Nucl. Acids Res.* **30**, 73–75 (2002).
84. Zanzoni, A. *et al.* MINT: a Molecular INteraction database. *FEBS Lett.* **513**, 135–140 (2002).
85. Bader, G. D. *et al.* BIND — the Biomolecular Interaction Network Database. *Nucl. Acids Res.* **29**, 242–245 (2001).
86. Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* **31**, 248–250 (2003).
87. Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001). **The only genome-wide protein–interaction map, so far, to be constructed for a prokaryote.**
88. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002). **This paper describes the identification of yeast–protein complexes using large-scale tandem-affinity purification coupled to mass spectrometry.**
89. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
90. Rison, S. C., Hodgman, T. C. & Thornton, J. M. Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics* **1**, 56–69 (2000). **An in-depth analysis and comparison of present functional classification schemes.**
91. Iliopoulos, I. *et al.* Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **19**, 717–726 (2003).
92. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
93. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
94. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* **1**, 349–356 (2002).
95. Mayr, E. Biological classification: toward a synthesis of opposing methodologies. *Science* **214**, 510–516 (1981).
96. Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21–28 (2001). **An automated analysis of the biomedical literature that identifies large-scale functional associations between thousands of human genes.**
97. Stephens, R. S. *et al.* Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759 (1998).
98. Fromont-Racine, M. *et al.* Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95–110 (2000). **This paper describes the first large-scale use of two-hybrid arrays to identify protein interactions in yeast.**
99. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).

## Acknowledgements

R.M.R.C. would like to acknowledge the Medical Research Council (UK) for support. J.P.L. would like to acknowledge the Foundation for Science and Technology (Portugal). We also thank C. Von Mering, I. Rigoutsos and colleagues at the European Bioinformatics Institute (EBI) for providing information on Figure 1.

 Online Links

## FURTHER INFORMATION

AllFuse: <http://maine.ebi.ac.uk:8000/services/allfuse>  
 BIND: <http://cbm.bio.uniroma2.it/mint>  
 BioCyc: <http://biocyc.org>  
 Bio-Dictionary: <http://www.research.ibm.com/bioinformatics/metadata.phtml.html>  
 BLOCKS: <http://www.blocks.fhrc.org>  
 CATH: <http://www.biochem.ucl.ac.uk/bsm/cath>  
 CellZome: <http://www.cellzome.com>  
 ClustR: <http://www.ebi.ac.uk/cluster>  
 COGS: <http://www.ncbi.nlm.nih.gov/COG>  
 CYGD: <http://mips.gsf.de/proj/yeast/CYGD/db>  
 DIP: <http://dip.doe-mbi.ucla.edu>  
 DSSP: <http://www.sander.ebi.ac.uk/dssp>  
 eMOTIF: <http://motif.stanford.edu/emotif>  
 ERGO: [http://www.integratedgenomics.com/ergo\\_light/ergo\\_overview.html](http://www.integratedgenomics.com/ergo_light/ergo_overview.html)  
 FSSP: <http://www.ebi.ac.uk/dali/fssp>  
 Gene Ontology: <http://www.geneontology.org>  
 GeneQuiz: <http://jura.ebi.ac.uk:8765/ext-genequiz>  
 HSSP: <http://www.sander.ebi.ac.uk/hssp>  
 InterPro: <http://www.ebi.ac.uk/interpro>  
 ISCB: <http://www.iscb.org>  
 KEGG: <http://www.genome.ad.jp/kegg/kegg2.html>  
 MetaFam: <http://metafam.ahc.umn.edu>  
 MINT: <http://cbm.bio.uniroma2.it/mint>  
 MIPS: <http://mips.gsf.de>  
 Pfam: <http://www.sanger.ac.uk/Software/Pfam/index.shtml>  
 PIM: <http://pim.hybrigenics.com/pimrider/pimriderlobby/PimRiderLobbyHpFull.jsp>  
 Predictome: <http://predictome.bu.edu>  
 PRINTS: <http://bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html>  
 ProDom: <http://prodes.toulouse.inra.fr/prodom/2002.1/html/home.php>  
 PROSITE: <http://us.expasy.org/prosite>  
 ProtoMap: <http://protomap.cornell.edu/index.html>  
 SCOP: <http://scop.mrc-lmb.cam.ac.uk/scop>  
 SGD: <http://www.yeastgenome.org>  
 SMART: <http://smart.ox.ac.uk>  
 STRING: <http://www.bork.embl-heidelberg.de/STRING>  
 Swiss-Prot: <http://us.expasy.org/sprot>  
 SYSTEMS: <http://systems.molgen.mpg.de>  
 TIGRFAMS: <http://www.tigr.org/TIGRFAMS>  
 TRIBES: <http://maine.ebi.ac.uk:8000/services/tribes>  
 TRIPLES: <http://ygac.med.yale.edu/triples/triples.htm>  
 YPL.db: <http://ypl.tugraz.at>  
**Access to this interactive links box is free online.**