

# Did evolution leap to create the protein universe?

Burkhard Rost

The genomes of over 60 organisms from all three kingdoms of life are now entirely sequenced. In many respects, the inventory of proteins used in different kingdoms appears surprisingly similar. However, eukaryotes differ from other kingdoms in that they use many long proteins, and have more proteins with coiled-coil helices and with regions abundant in regular secondary structure. Particular structural domains are used in many pathways. Nevertheless, one domain tends to occur only once in one particular pathway. Many proteins do not have close homologues in different species (orphans) and there could even be folds that are specific to one species. This view implies that protein fold space is discrete. An alternative model suggests that structure space is continuous and that modern proteins evolved by aggregating fragments of ancient proteins. Either way, after having harvested proteomes by applying standard tools, the challenge now seems to be to develop better methods for comparative proteomics.

## Addresses

CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168<sup>th</sup> Street, BB217, New York, NY 10032, USA and Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA; e-mail: rost@columbia.edu

**Current Opinion in Structural Biology** 2002, 12:409–416

0959-440X/02/\$ – see front matter  
© 2002 Elsevier Science Ltd. All rights reserved.

## Abbreviations

**ADS** antecedent domain segment  
**NORS** no regular secondary structure  
**ORF** open reading frame  
**rmsd** root mean square deviation

## Introduction

*Natura non facit saltus* (Nature does not make leaps)

Attributed to Tito Lucrezio Caro (Titus Lucretius Carus), 34 AD ?;  
Gottfried Wilhelm von Leibniz, 1698;  
Charles Darwin, 1859.

Some perceive New York City as a place jammed with humans. However, for those who live there, the metropolis feels like a village because life evolves around neighborhoods. Scientists explaining our behavior from a ‘systems’ perspective can argue that we belong to various groups defined by our address, our work or even the shape of our noses. We may possibly feel that such classifications do not explain who we really are. In analogy, we cannot expect catalogues of entirely sequenced genomes to convey understanding of protein structure, function or evolution. The literature bursts with examples of detailed studies of how protein structure and function co-evolve [1–7,8\*,9].

Although most of these analyses utilize a wealth of biological data, they are not explicitly based on the fact that we have entire genome sequences from representatives of all three kingdoms of life: eukarya, bacteria (prokaryotes) and archaea. What do we learn from generating lists of parts of the whole [10,11\*]?

Here, I focus on findings from methods that endeavor to capture overall features for entire organisms. I challenge the assumption that bioinformatics is slowly but steadily approaching the point at which we can smoothly move through neighborhoods of protein relationships in order to generate an atlas of the fates and functions of proteins in the context of the cell.

## Overview of proteomes: catalogs of structure and function

### Eukaryotes have many very long proteins

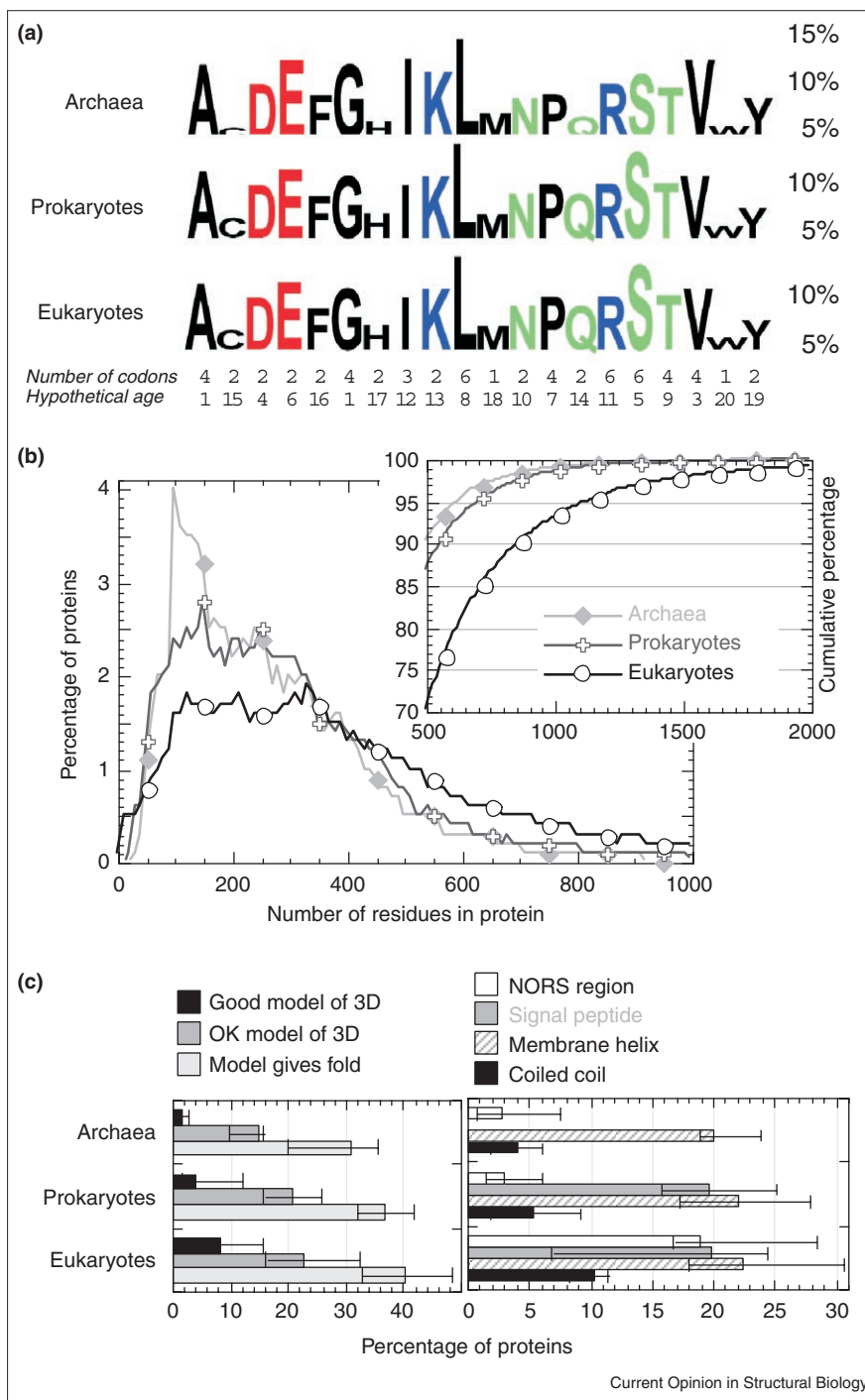
Genomes differ significantly in their nucleotide composition [12\*]. In contrast, the amino acid compositions of the entire proteomes of 28 organisms from all three kingdoms are similar ([13\*]; Figure 1a). The length of proteins differs significantly between the three kingdoms (Figure 1b); in particular, about 7% of eukaryotic proteins are longer than 1000 residues, whereas less than 2% of all proteins in archaea and prokaryotes are that long.

We now realize that many proteins were overlooked in the initial annotation of genome projects, as demonstrated by the annual growth of the estimated number of proteins in worm [14], yeast [15] and human. Some groups specialize in hunting for these overlooked proteins [16\*]. In microbial genomes, however, the number of proteins appears significantly over-estimated [12\*]. If so, the differences in the protein length distributions for short proteins might not hold up (Figure 1b). Nevertheless, the corrections in the number of proteins, as suggested by Skovgaard, Krogh and colleagues [12\*], alter the distributions for long proteins only marginally (data not shown). Some of the incorrectly annotated short proteins might actually be short RNAs [17] and others might be pseudo-genes [18].

### Despite the complete set of sequences, comparisons remain guesses

Comparing proteomes on the basis of residue composition or protein length constitutes an extremely dumb realization of comparative genomics. Unfortunately, more meaningful comparisons across kingdoms typically require focusing on subsets of the complete proteomes for which we have annotations or involve predictions of limited accuracy, or both. For instance, statements about protein families, folds and functions either are restricted to some arbitrary subset of proteins for which we can infer features by homology or are limited by the accuracy of prediction methods.

Figure 1



Comparing proteomes. **(a)** Amino acid usage in each of the three kingdoms. The height of the letter is proportional to the frequency of the respective amino acid (overall percentages given on the right). The lower rows show the number of codons for each of the amino acids and the age rank (1 is oldest, 20 newest), as estimated by Trifonov *et al.* [88]. The variation of the frequency within the kingdoms is as insignificant as that between the kingdoms [13\*]. **(b)** Distribution of protein length. The data are binned in intervals of ten residues. The inset gives the cumulative length; it illustrates the significantly higher proportion of long proteins in eukaryotes. Both (a) and (b) are based on an analysis of 238 326 proteins in 63 complete proteomes (12 archaea, 46 prokaryotes and 5 eukaryotes [89]). **(c)** The left panel shows the percentage of proteins for which we can predict structure through comparative modeling. Three types of models are distinguished by applying different cut-offs for the required level of sequence similarity [90]. At levels above 70% pairwise sequence identity, models reach an average accuracy around 2 Å rmsd ('good models', dark bars). At a level of sequence similarity that corresponds to 32% pairwise identical residues over 100 residues aligned ('HSSP' distance of 0 [13\*]), models differ about 3–5 Å ('OK models', gray bars). At a PSI-BLAST threshold of  $10^{-3}$ , most models identify the fold correctly ('model gives fold', white bars). The right panel shows the percentage of proteins predicted with NORS regions [51] (white bars), signal peptides (gray bars, note: all these proteins are extracellular), transmembrane helices (striped bars) and coiled-coil helices (dark bars). The graphs in (c) are based on 30 proteomes, as given in [13\*]. Note that the lines give the spread between the minima and maxima found in the respective kingdom.

Consider the idea to investigate whether or not particular folds are used more often in some organisms than in others. Firstly, we know neither which fraction of all existing folds we know already (estimates range from 10 to 50% [3,13\*,19\*,20\*,21]) nor whether the types of folds we know are representative. Secondly, we can, at best, infer the folds for half of all proteins in entirely sequenced genomes (Figure 1c) [9,10,19\*,22–24,25\*,26]. Thirdly, we have no way to predict novel folds in the context of entire

proteomes [4]. It is much easier to infer aspects of structure from sequence similarity than to infer aspects of function. Thus, the classification of proteomes by function relies even more on guesses than classification by structure.

#### Popular folds also most populated in proteomes

Gerstein pioneered the structural census of proteomes [10,27] by analyzing which folds are most often used in proteins of different organisms. When analyzing the

universe of protein sequences, we observe that some types of proteins belong to larger families than others [13<sup>\*</sup>,28–33]. As expected, folds used in these large protein families also dominate proteomes [2,3,10,11<sup>\*</sup>,34–37]. The major difference between the types of proteins that populate the largest sequence-based family and those that populate the largest fold-based family was the under-representation of membrane proteins with known structures [25<sup>\*</sup>,26].

### **Kingdoms have a similar percentage but different types of membrane proteins**

Contrary to speculation before completing the genome sequences of two animals and one plant, it seems that the percentage of membrane helical proteins differs less between multicellular and unicellular organisms than between different organisms within each of the three kingdoms [13<sup>\*</sup>]. Overall, about 16–26% of all proteins have membrane helices [13<sup>\*</sup>] (Figure 1c). Most membrane helical proteins have fewer than four helices and about half of all membrane proteins have no globular regions of considerable length [13<sup>\*</sup>]. One reason why membrane helix predictions are so valuable in the context of proteomes is that the number of helices is typically related to the type of protein and its function. Proteins with seven transmembrane helices (e.g. G-protein-coupled receptors) are significantly over-represented in worm and human, whereas proteins with six and twelve helices (e.g. transporters) are over-represented in most prokaryotes [13<sup>\*</sup>]. Surprisingly, we found relatively few seven transmembrane helix proteins in fly and many in worm. This finding may be explained by the immense difference in the number of olfactory receptors alone: worm appears to contain 1000 smell receptors, whereas fly has less than 100 [38]. Interestingly, the conservation of protein type does not always span all kingdoms: membrane protein families spanning all kingdoms do not necessarily have the same number of membrane helices, suggesting that proteins can add or remove helices over the course of evolution [13<sup>\*</sup>].

Although many methods predict  $\alpha$ -helical membrane proteins, few methods have addressed the prediction of proteins that insert  $\beta$ -strand barrels into the membrane [39]. Recently, methods have reached reasonable levels of accuracy in predicting  $\beta$  strands in membranes from sequence alignments [40<sup>\*</sup>]. Furthermore, a simple statistical model predicted about 105 potential  $\beta$ -barrel membrane proteins in the two Gram-negative bacteria *Escherichia coli* and *Pseudomonas aeruginosa* [41<sup>\*</sup>].

### **Coiled-coil proteins are significantly over-represented in eukaryotes**

Secondary structure correlates with function [42] and we can learn about evolution from taxonomy of secondary structure [43]. However, the content of secondary structure is only of limited value in the context of comparative proteomics [44]. One exception is the presence of coiled-coil helices: eukaryotes appear to have significantly more coiled-coil proteins than all other kingdoms (Figure 1c) [13<sup>\*</sup>].

Proteins with coiled-coil regions are often insoluble because such regions are frequently responsible for aggregation; they often indicate structural proteins. However, the high fraction of coiled-coil proteins in eukaryotes might originate from the role of coiled-coil regions in protein–DNA interactions and in regulation, transcription and translation. The problem with this explanation is that, although we tend to assume that eukaryotes utilize a more complex machinery to control their protein repository, we have no solid data confirming this assumption on the scale of the entire proteome (see below).

### **Eukaryotes have significantly more ‘loopy’ proteins**

The most outstanding difference between eukaryotes and other kingdoms in terms of protein structure is the high fraction of proteins that appear to be unlike typical globular structures. It has long held true that proteins fold into a unique three-dimensional structure and that this structure determines protein function. Over the past few years, evidence has gathered such that we have to reassess this paradigm of structural biology: many long regions appear essentially unstructured in isolation [45<sup>\*</sup>,46]. Such regions could introduce particular flexibility in that they could adopt different shapes through binding (induced fit) [47–50].

We have recently analyzed proteins with long regions (>70 residues) that appear to have no regular secondary structure (NORS regions; [51]). Confirming neural-network-based predictions of disordered regions [45<sup>\*</sup>,52], we found that eukaryotes had, on average, 3–5 times more proteins with NORS regions than organisms from other kingdoms (Figure 1c) [51]. Many of these ‘loopy’ proteins appear to be involved in gene regulation. However, experimental results are needed to shed more light on this new class of protein.

### **Too many functionally unclassified proteins hampered comparing function**

Using EUCLID [53], we could classify about 45–65% of all proteins from 30 complete proteomes into one of 13 classes of cellular function at a level reported to yield about 70% correct classifications [54,55<sup>\*</sup>]. When grouping the 13 classes into three superclasses (i.e. energy, information and communication), we found similar compositions within the archaean and eukaryotic kingdoms [13<sup>\*</sup>]. In contrast, the composition varied significantly among prokaryotic organisms [13<sup>\*</sup>]. In more detail, we found the following differences: proteins in biosynthesis and energy metabolism were abundant in prokaryotes, whereas human seemed to have a larger portion of the classes related to transport, binding and regulatory functions [13<sup>\*</sup>]. The significant variations between prokaryotic proteomes might reflect the very different environments in which these organisms dwell. However, the most important result is that, although accepting classification errors of 30% or more, we still can classify only about half of all proteins. Thus, conclusions about the meaning of the relative proportions remain highly speculative, at best. We also could not verify earlier

findings that the subset of proteins with homology to known structures differs in their cellular function from proteins of unknown structure (J Liu, B Rost, unpublished data).

### **Detailed analysis of evolution: domains, pathways and orphans**

#### **Working hypothesis: domains constitute the atoms of protein structures**

Structural genomics aims at experimentally determining one structure for every fold in nature [9,56–58]. Proteins of unknown fold are identified by clustering protein sequences [3,19\*,22,29,59]. Intuitively, the goal is to group proteins with similar ‘structural elements’ and to separate these clusters from proteins not containing those structural elements. What is the smallest structural element? Structural biologists tend to take structural domains to be the ‘atoms of structure’. Thus, proteins have to be chopped into domains before clustering the protein universe; all methods that do this use evolutionary relationships and thereby implicitly connect ‘atoms of structure’ and ‘atoms of evolution’ [2,28,32,34,60,61,62\*,63,64,65\*]. The principal assumption is that, if protein A is similar to B and to C, but B is not similar to C, then B and C constitute domains of A. By applying this scheme to all complete eukaryotes, we found over 17 000 ‘domain-like’ fragment clusters [22]. This lower bound, based on complete data sets, generally confirmed earlier estimates based on extrapolations from representative data sets [19\*]. Even if structural domains are not the atoms of evolution, analyses based on domains are more accurate than those based on entire sequences.

#### **Domain combinations are evolutionarily conserved**

Like many other biological relationships, domain family relationships follow scale-free network relationships [37,66–68,69\*,70,71\*,72]; that is, these relationships are explained by statistical models that do not require assumptions about biology. In particular, only large families engage in many types of domain combinations, whereas small families engage in only a few types of domain combinations. The majority of domain combinations AB involve families of domains A and B spanning across all kingdoms of life [71\*,72]. Teichmann and colleagues [71\*,72] suggest that evolution creates novel functions predominantly by combining existing domains. There are more repeats of similar domains adjacent to one another in eukaryotes than in other kingdoms [71\*,72]; the most extreme example of this is the giant protein titin [73]. Interestingly, the sequential order of different domains appears to be evolutionarily conserved [71\*,72,74].

#### **Mapping domains onto pathways suggests the image of a mosaic**

Although we have information about structure for less than half of all proteomes, almost 90% of the enzymes from the 106 small molecule metabolic pathways in *E. coli* have domains of known structure [75\*,76]. A particular fold is typically used only once in a given pathway. In other words, more homologues are distributed across pathways

than within pathways. Interestingly, 75% of all enzymes in metabolic pathways of *E. coli* appear to be enzymes known to catalyze a single enzymatic reaction and the majority of enzymes used in any metabolic pathway are specific to that particular pathway [77\*]. The authors concluded that pathways use enzyme mosaics [75\*,76], that is, they are taken from a limited set of protein families and there are no discernible repetitions.

As established in an excellent study of the structural conservation of enzymatic activity [8\*], enzyme families of small molecule metabolic pathways also often conserve their catalytic or cofactor-binding properties, whereas their substrate recognition properties seem rarely conserved [75\*,76]. About half of all protein–protein interactions are between domains from their own family [69\*]. Obviously, the ultimate goal of analyzing proteomes is to learn ways of refining our database searches. One particular application that uses the completeness of proteomes combines sequence analysis with structure prediction to find all disulfide oxidoreductases in yeast [78\*].

#### **Some folds might have been realized only once in nature**

Although the term ‘fold’ is not well defined, it intuitively refers to subunits between 30 and >700 residues long that let structural biologists recognize a particular protein structure. A few protein folds are used by many different protein sequences; they are often referred to as superfolds [2,34]. Presumably, the superfolds are more energetically favorable [79]. However, the most surprising result from the advent of genome sequencing is the observation that there is a constant rise in the number of known proteins that have no homologue of very similar sequence (i.e. each species uses some very specific proteins [23], often referred to as orphans). If we believe that cross-species evolution was a major event, we could argue that we simply fail to recognize the similarity between a particular kinase in aquifex and human, and therefore incorrectly classify that kinase as an orphan. In other words, we could argue that there is another protein that adopts the same fold, thus using a similar mechanism to realize function, but that we simply fail to find it because it has diverged too far in evolution.

Recently, Coulson and Moulton [20\*] proposed a somewhat shocking conclusion: most folds are specific to one species (i.e. the aquifex and the human kinase have different structures). They propose a model that assumes three separated regions: unifolds (realized only once in nature), superfolds (repeated many times) and mesofolds (between unifolds and superfolds). Coulson and Moulton estimate that there are over 10 000 folds in nature. Most of these are unifolds corresponding to orphan families. Note that this estimate is about three [13\*] to ten times [21] higher than previous estimates not considering the reality of orphans. At the other extreme, the model suggests that 80% of all sequence families adopt one of 400 superfolds, most of which are already known. If this proposition were true, we could argue cynically that the fastest way to a high yield

from structural genomics initiatives might be to simply identify the corresponding superfolds for the proposed tens of thousands of targets [19\*,22]. Obviously, structural genomics will not achieve a broad coverage of fold space using this short cut to producing high-throughput structure determination. This is one reason why the recommendations from the National Institutes of Health in the USA emphasized the goal of selecting new folds as targets.

#### Domains might not constitute the evolutionary atom

Did nature really separate three fold types (uni/meso/super) or are the separations based on a lack of the complete picture? If we believe that structural domains constitute the atoms of evolution, the concept of 'folds' and of the three types of distinct folds appears reasonable. However, there is evidence that the working hypothesis of folds or structural domains at the basis of evolution might not be the last word. First, the structurally most conserved regions often appear to be skeletons of active sites [80\*]. Technically, we can explore this observation by searching known structures with such three-dimensional motifs in order to find similarities obfuscated in sequence [7]. Second, the attempt to cluster sequence space based on putative structural domains results in large clusters of proteins that are connected through a ladder of 10–30 overlapping residues [22]. Third, particular stretches of 10–40 residue fragments are observed often in protein structures [81]. This leads to successes in predicting protein structure based on such fragments [4,82,83].

All these findings might be explained by a rather challenging hypothesis argued for in detail by Lupas, Ponting and Russell [84\*]: the diversity of today's folds might have evolved from peptide ancestors referred to as 'antecedent domain segments' (ADSs). The authors explain how ancient protein structures could have been formed by self-assembling aggregates of short polypeptides. They speculate that subsequently, and perhaps concomitantly with the evolution of higher fidelity DNA replication and repair systems, single polypeptide domains arose from the fusion of ADS genes. Although the authors provide ample details for the feasibility of their assumptions, we may never be able to falsify their model. Clearly, however, the hypothesis explains why it is so difficult to find similarities and why the same functional motif is often realized by many structures. The model implies that some modern proteins may have evolved by fusing multiple ADSs or by recombining domains that contain structurally compatible ADSs; these proteins are essentially of poly-phyletic origin. Thus, we would also understand why phylogenetic trees do not always agree [36,85]. We find many internal repeats that are shorter than structural domains; such repeats might constitute an evolutionary advantage in that they can adopt many functions easily [86\*]. The study of such repeats supports the ADS model [86\*]. Overall, the ADS model is appealing in the number of observations it explains given a minimum set of assumptions. Unfortunately, we still have to technically solve the difficult problem of identifying these ADSs from sequence and sequences drift easily, thus possibly erasing the ancient signal.

## Conclusions

Is protein structure and/or sequence space continuous, or has nature leaped when inventing folds and functions? If proteins were assembled from fragments, does this imply modularity of sequences and folds, as, for example, seen in short peptide fragments that regulate the targeting of proteins through the cell? Does the existence of short motifs or modules imply fragment assembly? I doubt that we have the data to unambiguously answer these questions. In fact, the evidence from analyses of entirely sequenced organisms is equally spread between pro and con '*natura non facit saltus*'.

The age of comparative proteomics has just begun. Already researchers have harvested many fruits by combining tools that had been developed in the past decade. Most papers reviewed here give examples of combining state-of-the-art methods and databases to explore protein function, structure and evolution. Today, the network of databases and methods generated by computational biology and bioinformatics approaches the complexity of organisms. However, we are still a long way from an atlas mapping the activity of a cell in terms of space (localization) and time (interaction history of each protein). Nakai [87\*] recently reviewed tools that capture some aspects about the '*in vivo* fates' of proteins. In a more abstract way, we can describe this objective by the following concept. First, describe all proteins by neighborhoods in terms of sequence families, structural families, sequence motifs, functional classes, pathways, expression profiles, interaction networks and subcellular compartments. Second, extend the similarity measure to a measure of distance. Third, combine these distances to enable a database search that simultaneously considers multiple classes of neighborhoods to find similarities between two proteins. If today's proteins really evolved by fragment assembly [84\*], methods that merge different features will be essential for comparative proteomics.

## Acknowledgements

Thanks to Jinfeng Liu (Columbia) for computer assistance, for the collection of genome data sets and for providing preliminary data. Also thanks to Henry Bigelow (Columbia) for helpful comments on the manuscript. The work of B Rost was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health. Last, but not least, thanks to all those who deposit their experimental data in public databases and to those who maintain these databases.

I refrained from quoting older publications, even if these would have been more appropriate; my apologies to those who should have been quoted. In highlighting papers, I considered only original articles and tended to ignore papers describing methods and databases, although these were the basis of the papers that were highlighted.

## 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. Lichtarge O, Sowa ME: **Evolutionary predictions of binding surfaces and interactions.** *Curr Opin Struct Biol* 2002, **12**:21-27.
  2. 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.
3. Dietmann S, Holm L: **Identification of homology in protein structure classification.** *Nat Struct Biol* 2001, **8**:953-957.
  4. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Curr Opin Struct Biol* 2001, **294**:93-96.
  5. Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LJ, Fetrow JS: **Enhanced functional annotation of protein sequences via the use of structural descriptors.** *J Struct Biol* 2001, **134**:232-245.
  6. Gaasterland T, Oprea M: **Whole-genome analysis: annotations and updates.** *Curr Opin Struct Biol* 2001, **11**:377-381.
  7. Thornton JM: **From genome to function.** *Science* 2001, **292**:2095-2097.
  8. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**:1113-1143.
- The authors present a thorough overview of the structural background of enzymatic activity. All 31 enzyme superfamilies investigated exhibit functional diversity generated by local sequence variation and domain shuffling. Commonly, substrate specificity is diverse across a superfamily, whereas the reaction chemistry is maintained. In many superfamilies, the position of catalytic residues can vary despite playing equivalent functional roles in related proteins.
9. Teichmann SA, Murzin AG, Chothia C: **Determination of protein function, evolution and interactions by structural genomics.** *Curr Opin Struct Biol* 2001, **11**:354-363.
  10. Qian J, Stenger B, Wilson CA, Lin J, Jansen R, Teichmann SA, Park J, Krebs WG, Yu H, Alexandrov V *et al.*: **PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information.** *Nucleic Acids Res* 2001, **29**:1750-1764.
  11. Hegyi H, Gerstein M: **Annotation transfer for genomics: measuring functional divergence in multi-domain proteins.** *Genome Res* 2001, **11**:1632-1640.
- The authors present a survey of annotation transfer for multidomain proteins. They find that the error in transferring functional classifications for multi-domain proteins is two times higher than that for single-domain proteins. On the other hand, if two multidomain proteins contain the same combination of two structural superfamilies, the probability of their sharing the same function increases significantly.
12. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A: **On the total number of genes and their length distribution in complete microbial genomes.** *Trends Genet* 2001, **17**:425-428.
- On the basis of length distributions of well-studied proteins and of a function between ORF length and stop-triplet frequency, the authors estimate the number of genes for a number of entirely sequenced microbial genomes. They suggest that too many short genes are annotated as ORFs and challenge that, for example, *E. coli* has only 3800 instead of 4300 genes.
13. Liu J, Rost B: **Comparing function and structure between entire proteomes.** *Protein Sci* 2001, **10**:1970-1979.
- Analysis of structural and functional features for 30 organisms that have had their genomes completely sequenced. Note, many of the findings are discussed in more detail in this review.
14. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of *Caenorhabditis elegans*.** *Nucleic Acids Res* 2001, **29**:82-86.
  15. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkötter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30**:31-34.
  16. Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB, Snyder M: **An integrated approach for finding overlooked genes in yeast.** *Nat Biotechnol* 2002, **20**:58-63.
- The authors integrate methods of gene-trapping, microarray-based expression analysis and genome-wide homology searching to unravel 137 previously overlooked proteins in yeast.
17. Rivas E, Klein RJ, Jones TA, Eddy SR: **Computational identification of noncoding RNAs in *E. coli* by comparative genomics.** *Curr Biol* 2001, **11**:1369-1373.
  18. Harrison PM, Echols N, Gerstein MB: **Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome.** *Nucleic Acids Res* 2001, **29**:818-830.
  19. Vitkup D, Melamud E, Moulton J, Sander C: **Completeness in structural genomics.** *Nat Struct Biol* 2001, **8**:559-566.
- The authors estimate the number of targets for structural genomics by extrapolating from the Pfam database. The number of targets is estimated to be above 17 000. However, the authors point out in detail how much this number increases if target selection in structural genomics is not coordinated.
20. Coulson AF, Moulton J: **A unifold, mesofold, and superfold model of protein fold use.** *Proteins* 2002, **46**:61-71.
- The authors suggest a three-tier model separating folds into three types: unifolds, mesofolds and superfolds. Superfolds are realized by many proteins of very diverged sequence; unifolds are confined to one particular protein; mesofolds occupy the region in between these two extremes. If the model is correct, it implies that the vast majority of all protein folds are unifolds, that there are over 10 000 different folds in nature and that 80% of all sequence families have one of about 400 folds, most of which are already known.
21. Wolf YI, Grishin NV, Koonin EV: **Estimating the number of protein folds and families from complete genome data.** *J Mol Biol* 2000, **299**:897-905.
  22. Liu J, Rost B: **Target space for structural genomics revisited.** *Bioinformatics* 2002, in press.
  23. Fischer D, Eisenberg D: **Predicting structures for genome proteins.** *Curr Opin Struct Biol* 1999, **9**:208-211.
  24. Pawlowski K, Rychlewski L, Zhang B, Godzik A: **Fold predictions for bacterial genomes.** *J Struct Biol* 2001, **134**:219-231.
  25. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313**:903-919.
- SUPERFAMILY is a collection of hidden Markov models (HMMs) for all known folds. The authors refine the original SAMT99 software to accomplish what appears to be one of the fastest, most accurate and most sensitive threading programs.
26. Gough J, Chothia C: **SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments.** *Nucleic Acids Res* 2002, **30**:268-272.
  27. Gerstein M, Levitt M: **A structural census of the current population of protein sequences.** *Proc Natl Acad Sci USA* 1997, **94**:11911-11916.
  28. Yona G, Linial N, Linial M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucleic Acids Res* 2000, **28**:49-55.
  29. Linial M, Yona G: **Methodologies for target selection in structural genomics.** *Prog Biophys Mol Biol* 2000, **73**:297-320.
  30. Bejerano G, Yona G: **Variations on probabilistic suffix trees: statistical modeling and prediction of protein families.** *Bioinformatics* 2001, **17**:23-43.
  31. 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.
  32. 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.
  33. Kriventseva EV, Biswas M, Apweiler R: **Clustering and analysis of protein families.** *Curr Opin Struct Biol* 2001, **11**:334-339.
  34. 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.
  35. Pearl FM, Martin N, Bray JE, Buchan DW, Harrison AP, Lee D, Reeves GA, Shepherd AJ, Sillitoe I, Todd AE *et al.*: **A rapid classification protocol for the CATH domain database to support structural genomics.** *Nucleic Acids Res* 2001, **29**:223-227.
  36. Lin J, Gerstein M: **Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.** *Genome Res* 2000, **10**:808-818.
  37. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313**:673-681.

38. Vosshall LB, Wong AM, Axel R: **An olfactory sensory map in the fly brain.** *Cell* 2000, **102**:147-159.
39. Schulz GE:  **$\beta$ -Barrel membrane proteins.** *Curr Opin Struct Biol* 2000, **10**:443-447.
40. Jacoboni I, Martelli PL, Fariselli P, De Pinto V, Casadio R: **Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor.** *Protein Sci* 2001, **10**:779-787.
- A neural-network-based method predicting  $\beta$  strands in membranes from sequence alignments. Although the performance reported by the authors appears impressively high, the method fails to find all  $\beta$ -strand membrane proteins in entire proteomes.
41. Wimley WC: **Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures.** *Protein Sci* 2002, **11**:301-312.
- This paper provides a detailed description of known  $\beta$ -barrel membrane proteins. Statistical preferences are employed to detect over 105 new putative  $\beta$ -barrel membrane proteins in *E. coli* and *P. aeruginosa*.
42. Andersen CAF, Palmer AG, Brunak S, Rost B: **Continuous assignment of secondary structure correlates with protein flexibility.** *Structure* 2002, **10**:175-184.
43. Przytycka T, Aurora R, Rose GD: **A protein taxonomy based on secondary structure.** *Nat Struct Biol* 1999, **6**:672-682.
44. Rost B: **Protein secondary structure prediction continues to rise.** *J Struct Biol* 2001, **134**:204-218.
45. 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.
- The authors review their method predicting 'disordered' regions in proteins. They survey functional mechanisms that might be related to disordered proteins. By comparing predictions for 29 complete proteomes, the authors note a significant difference between eukaryotes and all other kingdoms.
46. Dunker AK, Obradovic Z: **The protein trinity-linking function and disorder.** *Nat Biotechnol* 2001, **19**:805-806.
47. Wright PE, Dyson HJ: **Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.** *J Mol Biol* 1999, **293**:321-331.
48. Zetina CR: **A conserved helix-unfolding motif in the naturally unfolded proteins.** *Proteins* 2001, **44**:479-483.
49. Uversky VN, Gillespie JR, Fink AL: **Why are 'natively unfolded' proteins unstructured under physiologic conditions?** *Proteins* 2000, **41**:415-427.
50. Namba K: **Roles of partly unfolded conformations in macromolecular self-assembly.** *Genes Cells* 2001, **6**:1-12.
51. Liu J, Tan H, Rost B: **Loopy proteins appear conserved in evolution.** *J Mol Biol* 2002, in press.
52. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK: **Sequence complexity of disordered protein.** *Proteins* 2001, **42**:38-48.
53. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics* 1998, **14**:542-543.
54. Devos D, Valencia A: **Practical limits of function prediction.** *Proteins* 2000, **41**:98-107.
55. 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 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.
56. Thornton J: **Structural genomics takes off.** *Trends Biochem Sci* 2001, **26**:88-89.
57. Shapiro L, Harris T: **Finding function through structural genomics.** *Curr Opin Biotechnol* 2000, **11**:31-35.
58. Brenner SE: **A tour of structural genomics.** *Nature* 2001, **2**:801-809.
59. Mallick P, Goodwill KE, Fitz-Gibbon S, Miller JH, Eisenberg D: **Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: validating automated fold assignment methods by using binary hypothesis testing.** *Proc Natl Acad Sci USA* 2000, **97**:2450-2455.
60. Wu CH, Xiao C, Hou Z, Huang H, Barker WC: **iProClass: an integrated, comprehensive and annotated protein classification database.** *Nucleic Acids Res* 2001, **29**:52-54.
61. Reddy BV, Li WW, Shindyalov IN, Bourne PE: **Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins.** *Proteins* 2001, **42**:148-163.
62. Heger A, Holm L: **Picasso: generating a covering set of protein family profiles.** *Bioinformatics* 2001, **17**:272-279.
- Picasso starts from highly overlapping sequence neighborhoods revealed by all-on-all pairwise Blast alignment. Overlaps are reduced by merging sequences or parts of sequences into multiple alignments. Picasso groups functionally related proteins into about 10 000 unified domain families.
63. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28**:267-269.
64. Enright AJ, Ouzounis CA: **GeneRAGE: a robust algorithm for sequence clustering and domain detection.** *Bioinformatics* 2000, **16**:451-457.
65. 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 to be significantly more sensitive in detecting distant homologies than PSI-BLAST and IMPALA. The resulting method is applied to cluster all protein sequences in BioSphere.
66. Rzhetsky A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome.** *Bioinformatics* 2001, **17**:988-996.
67. Wuchty S: **Scale-free behavior in protein domain networks.** *Mol Biol Evol* 2001, **18**:1694-1702.
68. Lappe M, Park J, Niggemann O, Holm L: **Generating protein interaction maps from incomplete data: application to fold assignment.** *Bioinformatics* 2001, **17**:S149-S156.
69. Park J, Lappe M, Teichmann SA: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast.** *J Mol Biol* 2001, **307**:929-938.
- A detailed study of 8151 interacting pairs of protein domains in yeast. Almost half of all known families are found in interactions with domains from their own family. The authors also observe that the repertoires of interactions of domains within and between polypeptide chains overlap mostly for two specific types of protein families: enzymes and same-family interactions. They conclude that different types of protein interaction repertoires exist for structural, functional and regulatory reasons.
70. Wolf YI, Karev G, Koonin EV: **Scale-free networks in biology: new insights into the fundamentals of evolution?** *Bioessays* 2002, **24**:105-109.
71. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310**:311-325.
- The phylogenetic distribution of domain combinations is surveyed to establish the extent of common and kingdom-specific combinations. Of the kingdom-specific combinations, significantly more combinations consist of families present in all three kingdoms than of families present in one or two kingdoms. The authors conclude that recombination between common families, as compared to the invention of new families and recombination among these, has also been a major contribution to the evolution of kingdom-specific and species-specific functions in organisms in all three kingdoms.
72. Apic G, Gough J, Teichmann SA: **An insight into domain combinations.** *Bioinformatics* 2001, **17**:S83-S89.
73. Amodeo P, Fraternali F, Lesk AM, Pastore A: **Modularity and homology: modelling of the titin type I modules and their interfaces.** *J Mol Biol* 2001, **311**:283-296.
74. Bashton M, Chothia C: **The geometry of domain combination in proteins.** *J Mol Biol* 2002, **315**:927-939.
75. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*.** *J Mol Biol* 2001, **311**:693-708.
- This paper provides a detailed, thorough analysis of the domain compositions and family relationships for 90% of the enzymes in 106 small molecule

metabolic pathways of *E. coli*. The authors find that recruitment of domains across pathways is very common, but that there is little regularity in the pattern of domains. They describe this finding as an enzyme mosaic.

76. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **Small-molecule metabolism: an enzyme mosaic.** *TIBTECH* 2001, **19**:482-486.
77. Tsoka S, Ouzounis CA: **Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*.** *Genome Res* 2001, **11**:1503-1510.  
Detailed analysis of the enzymes in all metabolic pathways of *E. coli*. Most enzymes are found to be specific to one particular pathway. Only the most complex and extensively studied pathways are found to span more than ten enzyme families.
78. Fetrow JS, Siew N, Di Gennaro JA, Martinez-Yamout M, Dyson HJ, Skolnick J: **Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight?** *Protein Sci* 2001, **10**:1005-1014.  
Sequence analysis was combined with structure prediction to find all disulfide oxidoreductases in yeast. The method detected 24 possible candidates for unknown enzymes; at least three of these have since been verified experimentally.
79. Rykunov DS, Lobanov MY, Finkelstein AV: **Search for the most stable folds of protein chains: III. Improvement in fold recognition by averaging over homologous sequences and 3D structures.** *Proteins* 2000, **40**:494-501.
80. Irving JA, Whisstock JC, Lesk AM: **Protein structural alignments and functional genomics.** *Proteins* 2001, **42**:378-382.  
The authors find that the structurally most conserved regions in structural alignments of similar folds with different sequences are active sites. In the region of small common substructures, reduced aligned subsets define active sites and can be used to suggest the locations of active sites in homologous proteins.
81. Bystroff C, Thorsson V, Baker D: **HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins.** *J Mol Biol* 2000, **301**:173-190.
82. Jones DT: **Predicting novel protein folds by using FRAGFOLD.** *Proteins* 2001, **45**(suppl 5):S127-S132.
83. de La Cruz X, Sillitoe I, Orengo C: **Use of structure comparison methods for the refinement of protein structure predictions. I. Identifying the structural family of a protein from low-resolution models.** *Proteins* 2002, **46**:72-84.

84. 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.

This paper presents and discusses evidence suggesting how the diversity of domain folds in existence today might have evolved from peptide ancestors. The authors explain how ancient protein structures could have been formed by self-assembling aggregates of short polypeptides. Modern protein domains may thus be a product of fusing ADSs. If true, modern proteins may have a polyphyletic origin; consequently, phylogeny may be a difficult nut to crack.

85. Grishin NV, Wolf YI, Koonin EV: **From complete genomes to measures of substitution rate variability within and between proteins.** *Genome Res* 2000, **10**:991-1000.

86. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats: structures, functions, and evolution.** *J Struct Biol* 2001, **134**:117-131.

The authors review many different examples of internal repeats. They suggest that internal repeats constitute an evolutionary advantage because they enlarge the available binding surface area. They argue that multiple repeats that show strong structural and functional interdependencies could have evolved from a single repeat ancestor.

87. Nakai K: **Review: prediction of *in vivo* fates of proteins in the era of genomics and proteomics.** *J Struct Biol* 2001, **134**:103-116.

Even after a nascent protein emerges from the ribosome, its fate is still controlled by its own amino acid sequence information. Namely, it may be co-/post-translationally modified (e.g. phosphorylated, *N/O*-glycosylated and lipidated); it may be inserted into the membrane, translocated to an organelle or secreted to the outside milieu; it may be processed for maturation or selective degradation; finally, its fragment may be presented on the cell surface as an antigen. This paper reviews prediction methods that address such fates.

88. Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN: **Distinct stages of protein evolution as suggested by protein sequence analysis.** *J Mol Evol* 2001, **53**:394-401.

89. PEP: database with Predictions for Entire Proteomes on World Wide Web URL: <http://cubic.bioc.columbia.edu/pep>

90. Eyrich V, Marti-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B: **EVA: continuous automatic evaluation of protein structure prediction servers.** *Bioinformatics* 2001, **17**:1242-1243.