Published in final edited form as:
Curr Opin Struct Biol. 2009 June ; 19(3): 329–334. doi:10.1016/j.sbi.2009.03.011.

Nothing about Protein Structure Classification Makes Sense Except in the Light of Evolution

Ruben Valas, **Song Yang**, and **Philip E. Bourne**

Abstract

In this the 200th anniversary of Charles Darwin's birth and the 150th anniversary of the publication of the Origin of Species it is fitting to revisit the classification of protein structures from an evolutionary perspective. Existing classifications use homologous sequence relationships, but knowing that structure is much more conserved that sequence creates an iterative loop from which structures can be further classified beyond that of the domain, thereby teasing out distant evolutionary relationships. The desired classification scheme is then one in which a fold is merely semantics and structure can be classified as either ancestral or derived.

Introduction

In 1980 the Protein Data Bank (PDB; [1]) contained less than 100 structures and structural biologists had studied and could name most if not all of them. Today the PDB contains approximately 55,000 macromolecular structures of proteins, DNA, RNA, and complexes thereof, often combined with a variety of small molecules [2]. No human can assimilate such a breadth of information and so it is only natural, as has happened in so many areas of science with positive consequences, that we attempt an act of reductionism. Thus, the classification of protein structures is an attempt at reductionism from which biological function can be better interpreted. In its purest form reductionism would imply that the application of a simply theory could take a subset of structures, the unique set, and generate all others from it. Clearly this cannot be done completely, Nature is far too tricky, but the notion of generating all structures from a parts list [3] has persisted. Two parts are considered the same if they can be superimposed in 3-dimensions. This raises at least three issues. What constitutes a part; what metric defines two parts as the same, and most importantly, does that sameness convey any biological meaning? Stating the problem a different way, the parts list approach could be considered a bottom-up approach, whereas a consideration of the biological context a top-down approach. The issues then become how well do the two approaches mesh in the middle and what constitutes the biological context?

Already we have introduced a very significant set of issues, yet enormous scientific progress has been made through existing classification schemes. Let us briefly consider some of these schemes in the context of the bottom-up versus top-down approaches. This will serve as an introduction to why we believe the future calls for a more detailed classification which only makes sense in an evolutionary framework.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Bottom-up Approaches to Protein Structure Classification

A large variety of protein structure comparison algorithms have been developed over the past 20 years (see [4] for a review). While they use different methods of protein representation, different algorithms for comparison and different scoring functions, in the majority of cases the end result is a geometric comparison which results in a superposition of the structures according to a root means-square deviation (RMSD), length of alignment, number of gaps, and a score of the statistical significance. As was shown a number of years ago [5] and again more recently [6] there is rarely a unique answer and at a fine level of detail (the devil is often in the details) certainly leads to misalignments by failing to capture the biological relevance. Nevertheless, these methods lead to a reductionism which provides a non-redundant structural set as originally exemplified by Dali [7] and the FSSP database [8], with a number of other databases of classified protein structures following [4]. In the majority of cases the comparison is between protein domains and beyond that has little biological context.

Top-down Approaches to Protein Structure Classification

Top-down approaches are exemplified by CATH [9] and SCOP [10], today's gold standards for protein structure classification. While the sheer volume of data to classify requires automation (CATH more than SCOP), human expertise is still used since difficult cases require manual inspection. Much has been written about CATH and SCOP and comparisons have been made between these classification schemes [11] [12] and there is no need to go into further detail here. Both methods involve a consideration of protein domains and incorporate the biological context primarily through detecting homologous sequence relationships. This later point implies that evolution is already a consideration in structure classification; here we suggest that this needs to be taken further. How extant proteins emerged from smaller building blocks, the role of gene duplication, convergence versus divergence, and co-evolution in a functional context are examples of evolutionary considerations that need to be incorporated into future protein structure classification schemes as we shall see subsequently. In this context we would argue that the end goal of protein classification is to describe the evolutionary pathways between all protein structures.

The Protein Domain as Today's Unit of Structural Classification

Protein domains, as independent folding units, are the modular building blocks of proteins and most current protein structure classification schemes, whether top-down or bottom-up, are based on domains. Protein domain definition from 3D structure is not a fully solved problem [13,14] which explains some of the differences between existing classification schemes. Since many proteins are multi-domain proteins, and multi-domain proteins are more common in eukaryotes than prokaryotes, we already have a hint for the role evolution can play in an extended protein structure classification scheme. Some domains have high sequence similarity and are evolutionarily related; others are distantly related, sharing obvious structure similarity but not sequence similarity; others have similar topologies, but not to the point where there is clear evidence of common ancestry. Taking SCOP as an example, the first two groups are further classified into the family and superfamily levels, forming a hierarchical scheme. There lies a fundamental problem, a domain can be thought of as both an evolutionary and non-evolutionary unit. Difficulties with current schemes are further compounded by the notion of folds (all or part of a domain) which are considered discreet components in current top-down classification schemes. Folds are not considered from an evolutionary perspective, but they may be related. Folds do change during evolution to give rise to new folds [15,16]. Grishin proposed that it is possible for an all-alpha fold to evolve into an all-beta fold by sequential secondary structure flip-over [17]. Similarly, recent work attempted to create two short peptides with high sequence similarity but distinct folds [18]. They achieved this goal with two 50 amino acid peptides with 88% sequence identity, but totally different structure and function. Finally, another case which is difficult for the current classification schemes to embrace are chameleon sequences which can adopt multiple folds [19]. If one accepts the notion of gradual structural variation at the fold level, how can protein structures be classified this way? One notion is the use of smaller fragments [20], but as we shall propose subsequently, this too only makes sense in the light of evolution. In summary, whether or not two proteins are in the same fold is really semantics, whereas describing which is ancestral and which is derived truly captures their relationship. Unfortunately this is a harder problem than simply clustering similar structures. In part it is harder since first you need to identify that protein within extant species and second you need to know the relationship between those species and their ancestors. Ironically, the first problem is addressed well using existing classification schemes.

Domain Distribution in Complete Genomes

The recent accumulation of genomic and structural data as well as improvements in homology detection algorithms has led to the reliable prediction of the protein domain content of all completed genomes using both SCOP and CATH domain definitions [21,22]. These protein domain distributions are the starting point for the investigation of protein domain evolution in the genomic era [23–27].

The work of assessing the distribution of domain content across the tree of life began shortly after the completion of the first genomes from each of the three superkingdoms [28]. As the number of structures and the number of genomes accumulated a power law distribution of domains [29] and domain combinations [30] emerged. Several models have been proposed to explain this distribution [31,32]. To illustrate this point, according to SCOP 1.73 which contains 1087 folds, 692 folds contain only one family (and hence one superfamily). Therefore, the majority of folds correspond to one homologous family that covers a very tiny portion of sequence space. Conversely, the ferredoxin-like fold (SCOP d.58) is found in 55 superfamilies, comprising 123 families. This imbalance is undoubtedly the result of evolution as can be seen by considering the power law relationship with respect to the complexity of the organism.

Two independent groups compared domain abundance to features representing complexity, namely genome size [33] and numbers of cell types [34]. Ranea et al. [33] clustered domain families into three categories in terms of their relationship to genome size: unrelated (mainly translation and biosynthesis), linearly-related (mainly metabolism) and non-linearly-related (mainly involved in gene regulation). Vogel et al. [34] compared domain family abundance with cell type numbers in different eukaryote species. About 10% of domain families have a strong correlation with complexity. Half of these superfamilies are involved in extracellular processes and regulation. Such results infer subtle structure-function relationships of protein domains during evolution leading to the current protein structure repertoire.

An important evolutionary consideration is not just the abundance of domains, but their organization. Over 70% of proteins in eukaryotes and over 50% of proteins in prokaryotes contain more than one domain [23]. These multi-domain proteins are represented by linear combinations of domains; the domain architecture [35]. Domain architectures arise through domain shuffling, domain duplication, and domain insertion and deletion (see [36,37] for a review) leading to new functions [38]. Baus et al. [39] defined “promiscuous” domains as those that occur in diverse domain architectures. The authors provided a measurement of promiscuity of domains based on the frequency of their coexistence with different domain partners. A systematic comparative genomic analysis in 28 eukaryotes resulted in 215 strongly promiscuous domains. It is not surprising that most are involved in protein-protein interactions, especially in signal transduction pathways. Vogel et al. [40] observed an over-representation of some two-domain or three-domain combinations in complete genomes and termed them “supra-domains.” Those supra-domains (described here as macrodomains) have stable internal domain architectures that are conserved over long evolutionary distances, acting like a single domain in combination with other domain partners. About 1400 macrodomains have been identified with diverse functions, indicating that the preferred association of certain domains is universal and evolutionarily advantageous. These two examples show that domain combinations are determined by functional constraints and evolutionary selection, not just random processes [29]. As such, domain combinations are an important aspect of any protein classification scheme.

A logical extension of these findings is to map domain combinations to presumed phylogenetic relationships derived by other means as exemplified by Snell et al. [41]. Kummerfeld et al. [42] counted the distribution of various types of single domain and multi-domain proteins across the tree of life and predicted that fusion is four times more common than fission in domain combinations. Fong et al. [43] viewed the domain architecture in multi-domain proteins as the rearrangement of existing architectures, acquisition of new domains or deletion of old domains, and proposed a parsimony model to derive the evolutionary pathways by which extant domain architectures may have evolved. Guided by the evolutionary information in phylogenetic trees, Ekman et al. [44] studied the rate of multi-domain architecture formation across different branches of the phylogenetic tree and found that there are elevated rates of domain rearrangement in Metazoa, whereas creation of domains was more frequent in early evolution. Similarly, Itoh et al. [45] observed a large number of group-specific domain combinations in animals and investigated the difference in domain combinations among different phylogenetic groups. Yang et al. [46] aimed to derive the entire evolutionary history of each domain and domain combination throughout the tree of life by mapping current domain content onto the species trees. This approach reveals the origin of each protein domain as well as evolutionary processes such as horizontal gene transfer among more distant species.

Is the Domain the Correct Unit of Classification?

The discussion thus far has focused on the protein domain as the best single level for classifying protein structure, but it is by no means the only one. Just as Ford Doolittle has argued the shortcomings of tree representations to illustrate the relationship between species [47], calling for a pluralistic approach where no one tree maps all species, we propose a pluralistic approach to protein structure classification incorporating domains, subdomains, macrodomains, and both convergent and divergent evolution. Subdomain Features

There are currently several available tools for comparing proteins at the subdomain level. Fragnostic is a database that defines relationships in the PDB based on shared fragments between structures [48]. These fragments share both structural and sequence similarity. They can be varying sizes from 5 to 20 residues. Each of these edges is ambiguous (not defined as divergent or convergent evolution) and directionless. However, combining this information with other sources of information could polarize and test some of these edges as a hypothesis for structural evolution.

Another subdomain unit is the closed loop. Most protein structures are composed of loops that come back around on themselves every 25–30 residues [49]. Domain Hierarchy and closed Loops (DHcL) is a web server that decomposes protein structures into domains and closed loops based on van der Waals energies [50]. The protein modules that are the most conserved since the last universal common ancestor (LUCA) correspond to closed loops [51]. Recently all prokaryotic proteins were decomposed into 20 residue fragments (possible closed loops) and clustered based on an identity threshold [52]. The authors found that fragments that corresponded to closed loops were more likely to form large clusters. It is possible to walk between clusters because some have small connections. The authors propose this description is superior to a domain based one because it represents a finer view of protein function. Closed loops of a common origin in different superfamilies could be evidence for a common ancestor between those superfamilies. Functional sites are another subdomain feature that could be used for classification. Many distinct superfamilies bind the same ligand. It is possible that these superfamilies share a common ancestor that bound that ligand, but diverged in global structure while the site that binds the ligand is conserved. SMAP [53] finds such binding pockets with both sequence and structural conservation, so these are probably the result of divergent evolution. However, it is also possible that two superfamilies could converge on the same ligand. The PROCOGNATE database defines what superfamilies bind what ligand using structural information from the PDB [54]. A combination of these approaches could create a ligand based classification for domains that encompasses both convergent and divergent evolutionary events.

Macrodomain Features

A protein-protein interaction site is an example of a macro feature conserved from an evolutionary perspective. The interface is conserved while the composite proteins form new superfamilies. A comparison between all protein-protein interfaces in the PDB revealed several examples of highly similar interfaces between different pairs of superfamilies [55]. MAPPIS is a tool for aligning protein-protein binding sites [56]. This level of classification is best done using quaternary structure. 3D complex is a database that classifies protein structures by their quaternary structure [57]. Homomeric complexes evolve in a stepwise fashion from monomers to structures with cyclic symmetry and then to structures with dihedral symmetry [58]. This information can be used to establish evolutionary relationships between homomers. As an example consider the SCOP family N-acylglucosamine (NAG) epimerase (48222). SCOP 1.73 has two structures in this family; N-acyl-D-glucosamine 2-epimerase(1fp3) and NAG isomerase (2afa). N-acyl-D-glucosamine 2-epimerase is a dimer with cyclic symmetry (C2) and NAG isomerase is a hexamer with dihedral symmetry (D3) according to the 3D complex database [57]. This implies that NAG isomerase must be derived from N-acyl-D-glucosamine 2-epimerase which evolved from one of the many monomeric structures found in this superfamily. It should be noted there may be structural intermediates that have not yet been solved.

Quaternary structure can also define the evolution of some heteromeric complexes. The simplest case is when a heteromer is composed of the same chains as a homomer. The heteromer is almost certainly derived via gene duplication. There are many examples in SCOP where proteins in the same family or superfamily have different quaternary structures. We propose that this information must be incorporated in a classification scheme. A domain based scheme would simply say these proteins share a common ancestor, while a system that includes quaternary structure defines them more explicitly. In summary a domain based classification implies common ancestry, but a macrodomain and subdomain analysis implies an evolutionary hypothesis.

Putting it all Together

We are proposing a pluralistic (some would say fuzzy) approach to protein structure classification that depends much more on evolution than simply defining homologous relationships between sequences as used in current top-down approaches. Yet these existing schemes form the basis from which pluralism is possible. Pluralism still proposes the domain as a fundamental evolutionary unit, yet encompasses the notion of subdomains and macrodomains.

The scheme needs to be dynamic since many phylogenetic relationships upon which the classification is based will change. For example, there are currently several proposed branching orders for the major taxonomic groups [59,60]. In the Cavalier-Smith scheme [59], archaea and eukaryotes are both derived superkingdoms, so if there is a link between a protein in bacteria and another found in only archaea, the archaeal protein must be derived. The tree of life infers polarity in the evolution of proteins, but the classification of proteins can also polarize the tree of life. Ideally the two would eventually converge to a solution that captures the history of species as well as proteins. Difficulties arise with our pluralistic scheme since convergence of structure reflects independent evolutionary invention of similar structural folds. Although convergent evolution of structure is rare, it does occur and thus can we really know if promiscuous folds, such as the TIM beta/alpha barrel fold, did not emerge several times independently in evolution? How many cases are there like this?

In our pluralistic scheme any relationship can be defined as divergent, convergent, or ambiguous. What would the map of protein classification/evolution look like when it is complete? It would likely consist of a series of views at different levels of structural granularity where each feature in a given structure could be mapped to equivalent features in other structures and mapped to its presence or absence in extant organisms and by inference common ancestors. The ancestry of modern proteins would reveal the history of their domains and domain combinations as well as similar and dissimilar micro and macro features. The architecture of the classification scheme would depend on the level it was being explored. Domains would exist as part of a directed acyclic graph if their ancestry was established or as undirected graphs for convergent or ambiguous events.

If such an integrated scheme were in place, and it is a big if, we could contemplate protein evolution in and before LUCA. The superfamily content of the last universal common ancestor (LUCA) has been estimated to contain over 140 different superfamilies [61], although we argue this is an underestimate (in preparation)). It has also been proposed that the oldest fold is the P-loop containing nucleoside triphosphate hydrolase [62]. But how did this fold arise? If we are to root a classification based on evolution we need to explain how to get from that fold to 140 different superfamilies. This is not possible by simply comparing sequences or even structures of whole domains. Protein evolution probably began with structures smaller than what we would consider a domain. It has been proposed that the earliest proteins were created by trans-splicing RNAs that code for protein modules and the origin of genes is much later, independent in archaea and bacteria [63]. Understanding the relationship between the modules that composed LUCA is essential to testing this idea and other hypothesis' about LUCA. This will only be possible by classifying protein structures based on an evolutionary scheme at all levels of protein structure.

The possibility of a pluralistic scheme of protein structure classification is only possible by virtue of the foresight and hard work that has gone into creating our existing bottom-up and top down approaches. Notwithstanding, if improvements in important areas such as functional annotation and structure prediction are to be made new insights are needed. Further use of what evolution can teach us would seem to be required. In so doing Nature's reductionism will become the reductionism that helps science advance.

References

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542. [PubMed: 875032]
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]
3. 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. [PubMed: 11292848]

4. Marti-Renom, M.; Capriotti, E.; Shindyalov, I.; Bourne, P. Structure Comparison and Alignment. Vol. 2. Gu, J.; Bourne, P., editors. New York: Wiley-Blackwell; 2009.

5. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338. [PubMed: 8819165]

6. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188. [PubMed: 15701525]

7. Holm L, Kaariainen S, Wilton C, Plewczynski D. Using Dali for structural comparison of proteins. Curr Protoc Bioinformatics. 2006Chapter 5:Unit 5 5.

8. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234. [PubMed: 9016542]

9. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 2008

10. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36:D419–425. [PubMed: 18000004]

11. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112. [PubMed: 10508779]

12. Jefferson ER, Walsh TP, Barton GJ. A comparison of SCOP and CATH with respect to domain-domain interactions. Proteins 2008;70:54–62. [PubMed: 17634986]

13. Holland TA, Veretnik S, Shindyalov IN, Bourne PE. Partitioning protein structures into domains: why is it so difficult? J Mol Biol 2006;361:562–590. [PubMed: 16863650]

14. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward consistent assignment of structural domains in proteins. J Mol Biol 2004;339:647–678. [PubMed: 15147847]

15. Goldstein RA. The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol 2008;18:170–177. [PubMed: 18328690]

16. Taylor WR. Evolutionary transitions in protein fold space. Curr Opin Struct Biol 2007;17:354–361. [PubMed: 17580115]

17. Grishin NV. Fold change in evolution of protein structures. J Struct Biol 2001;134:167–185. [PubMed: 11551177]

18**. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc Natl Acad Sci U S A 2007;104:11963–11968. [PubMed: 17609385]Could proteins with sequence identity higher than 80% not be homologous? This is shown to be true for engineered proteins. Although the two designed protein peptides are only about 50 amino acids long and this scenario is not likely to be common, this report requires we rethink protein sequence-structure relationship, protein folding, protein classification and protein evolution.

19. Andreeva A, Murzin AG. Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 2006;16:399–408. [PubMed: 16650981]

20. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins 2000;38:247–260. [PubMed: 10713986]

21. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C. Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res 2008;36:D414–418. [PubMed: 18032434]

22. Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res 2007;35:D308–313. [PubMed: 17098927]

23. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science 2003;300:1701–1703. [PubMed: 12805536]

24. Copley RR, Doerks T, Letunic I, Bork P. Protein domain analysis in the era of complete genomes. FEBS Lett 2002;513:129–134. [PubMed: 11911892]

25. Doolittle RF. Evolutionary aspects of whole-genome biology. Curr Opin Struct Biol 2005;15:248–253. [PubMed: 15963888]

26. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218–223. [PubMed: 12432406]

27. Orengo CA, Thornton JM. Protein families and their evolution-a structural perspective. Annu Rev Biochem 2005;74:867–900. [PubMed: 15954844]

28. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J Mol Biol 1997;274:562–576. [PubMed: 9417935]

29. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. Genome Res 1999;9:17–26. [PubMed: 9927481]

30. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310:311–325. [PubMed: 11428892]

31. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci U S A 2002;99:14132–14136. [PubMed: 12384571]

32. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2002;2:18. [PubMed: 12379152]

33. Ranea JA, Buchan DW, Thornton JM, Orengo CA. Evolution of protein superfamilies and bacterial genome size. J Mol Biol 2004;336:871–887. [PubMed: 15095866]

34. Vogel C, Chothia C. Protein family expansions and biological complexity. PLoS Comput Biol 2006;2:e48. [PubMed: 16733546]

35. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem 1995;64:287–314. [PubMed: 7574483]

36. Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. Trends Biochem Sci 2008;33:444–451. [PubMed: 18656364]

37. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 2004;14:208–216. [PubMed: 15093836]

38. Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure 2007;15:85–99. [PubMed: 17223535]

39. Basu MK, Carmel L, Rogozin IB, Koonin EV. Evolution of protein domain promiscuity in eukaryotes. Genome Res 2008;18:449–461. [PubMed: 18230802]

40. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA. Supra-domains: evolutionary units larger than single protein domains. J Mol Biol 2004;336:809–823. [PubMed: 15095989]

41**. Snel B, Bork P, Huynen MA. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 2002;12:17–25. [PubMed: 11779827]The first attempt at using phylogenetic relationship to derive the gene contents of hypothetical ancestor species from the genomes of contemporary organisms, and to analyze the evolution of genomes and individual genes.

42. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet 2005;21:25–30. [PubMed: 15680510]

43*. Fong JH, Geer LY, Panchenko AR, Bryant SH. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol 2007;366:307–315. [PubMed: 17166515]The authors combine domain data and a species tree to infer the most probable order of events that leads to the current distribution of domain combinations. They find that domain fusion is much more probable than domain fission

44. Ekman D, Bjorklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol 2007;372:1337–1348. [PubMed: 17689563]

45. Itoh M, Nacher JC, Kuma KI, Goto S, Kanehisa M. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. Genome Biol 2007;8:R121. [PubMed: 17588271]

46. Yang S, Bourne PE. The Evolutionary History of Protein Domains Viewed by Species Phylogeny. PLoS Comput Biol. To be published

47*. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A 2007;104:2043–2049. [PubMed: 17261804]A sanity to check the field of evolution. This paper questions whether Darwin’s metaphor of a tree of life is the correct representation for the evolution of species. The ideas raised by this paper apply to any system used to represent evolution and should be considered when considering how to describe the evolution of proteins.

48. Friedberg I, Godzik A. Fragnostic: walking through protein structure space. Nucleic Acids Res 2005;33:W249–251. [PubMed: 15980462]

49. Berezovsky IN, Grosberg AY, Trifonov EN. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 2000;466:283–286. [PubMed: 10682844]

50. Koczyk G, Berezovsky IN. Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure. Nucleic Acids Res 2008;36:W239–245. [PubMed: 18502776]

51. Sobolevsky Y, Trifonov EN. Protein modules conserved since LUCA. J Mol Evol 2006;63:622–634. [PubMed: 17075700]

52**. Frenkel ZM, Trifonov EN. From protein sequence space to elementary protein modules. Gene 2008;408:64–71. [PubMed: 18022768]The authors divide all prokaryotic proteins into small fragments, and cluster the fragments based on sequence identity. They find fragments corresponding to closed loops are in large clusters. They propose these large clusters reflect functional modules that are reused in different proteins. This a subdomain feature of protein structure which may describe function.

53. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A 2008;105:5441–5446. [PubMed: 18385384]

54. Bashton M, Nobeli I, Thornton JM. Cognate ligand domain mapping for enzymes. J Mol Biol 2006;364:836–852. [PubMed: 17034815]

55. Mintz S, Shulman-Peleg A, Wolfson HJ, Nussinov R. Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. Proteins 2005;61:6–20. [PubMed: 16184518]

56. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ. MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. Nucleic Acids Res 2008;36:W260–264. [PubMed: 18467424]

57. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2006;2:e155. [PubMed: 17112313]

58**. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. Nature 2008;453:1262–1265. [PubMed: 18563089]The authors investigate the evolutionary pathways of homomeric protein complexes. They propose a stepwise model for evolution of quaternary structure. They experimentally verify that complexes assemble and dissemble in the same the routes that the authors propose the complexes evolved

59*. Cavalier-Smith T. Rooting the tree of life by transition analyses. Biol Direct 2006;1:19. [PubMed: 16834776]Cavalier-Smith analyzes 13 transitions which he claims polarizes the tree of life. Many of these transitions are based on protein structure. This is probably the most detailed description of the evolution of the major taxa to date. Any classification scheme that incorporates evolution must incorporate the ideas included in this work.

60. Gupta RS, Griffiths E. Critical issues in bacterial phylogeny. Theor Popul Biol 2002;61:423–434. [PubMed: 12167362]

61. Ranea JA, Sillero A, Thornton JM, Orengo CA. Protein superfamily evolution and the last universal common ancestor (LUCA). J Mol Evol 2006;63:513–525. [PubMed: 17021929]

62. Ma BG, Chen L, Ji HF, Chen ZH, Yang FR, Wang L, Qu G, Jiang YY, Ji C, Zhang HY. Characters of very ancient proteins. Biochem Biophys Res Commun 2008;366:607–611. [PubMed: 18073136]

63. Di Giulio M. The origin of genes could be polyphyletic. Gene 2008;426:39–46. [PubMed: 18706983]