GUSTAVO CAETANO-ANOLLÉS

# Protein evolution: Charting the universe of macromolecular structure

**ABSTRACT**

*Understanding how protein architecture evolves is paramount if we are to engineer protein targets and drugs with pharmacological properties, understand the epidemiology and evolution of disease, or model protein interactions in systems biology. In order to 'chart' the universe of protein architecture from an evolutionary perspective, structural similarities in protein design have been linked to evolutionary patterns of architectural occurrence in genomes using a general phylogenomic approach. Analysis of over 30 genomes revealed patterns of evolution in the overall structure of molecules and clear molecular transformation pathways.*

*This evolutionary structure uncovered in the world of proteomes and architectures can now be used to model evolutionary processes. This type of research uses principles drawn from disparate disciplines and addresses fundamental issues in biology such as the origin and diversification of life, the role of lateral gene transfer, and the network behavior of biological systems.*

Protein and nucleic acid molecules are the fundamental building blocks of life. The study of how these molecules are arranged in three-dimensional space is central to unraveling mechanisms and evolution of biological systems. This realization has spurred structural genomics, a new field that seeks the wide-scale acquisition of structural information from molecules (1,2). However, novel approaches are needed to understand i) how a genotype (the heritable repository of biological information) is mapped into a phenotype (the physical, organizational and behavioral manifestation of life); ii) how biological function and fitness condition at the molecular level the success of molecules, cells, organisms, species, and other subjects of natural selection; and iii) how evolution shapes the structure of macromolecules.

Focusing on nucleic acid molecules, we recently reconstructed evolutionary history directly from the structure of RNA, using cladistic principles and considerations in statistical mechanics (3).

This award-wining approach (4) "embeds structure and function directly into phylogenetic analysis" (5) producing hierarchical representations of ancestry (*phylogenies*; see definitions in Box 1) that are rooted and express the directionality of evolution's arrow. Evolutionary relationships were inferred on the basis of shared and derived characteristics in the structure of RNA molecules that had been well defined by crystallographic, functional, and comparative sequence analyses. Molecules were characterized by attributes describing topological features, thermodynamic properties, or statistical parameters that are capable of defining the geometry, stability and uniqueness of folded conformations. These attributes were then treated as linearly ordered *cladistic characters* (Box 1), and these characters were 'polarized' by fixing the direction of

**Characters:** Observable features that distinguish one object from another and constitute hypotheses of *primary homology*, i.e. they share modifications from a previous condition. Characters can display multiple numerical values and frequency distribution of values, called "character states".

**Cladistics:** A method of classification that groups taxa or objects hierarchically into nested sets.

**Modularity:** Ability to maintain structural integrity of autonomous components across a wide range of environments and genetic contexts.

**Phylogeny:** Hierarchical branching histories of inheritance (phylogenetic trees) in which network-like genealogies and non-hierarchical tokogenic relationships provide the foundation for evolutionary change. Phylogenetic trees constitute hypotheses of genealogical relationship.

**Protein domain:** A building block of a globular protein that has compact three-dimensional structure and acts as an evolutionary unit.

**Protein fold:** A folding topology shared by a number of proteins with common structure and function. Protein folds encompass proteins with common evolutionary origin that may or may not share significant sequence similarity.

**Synexpression:** Tight coordination of the expression of groups of genes functioning in a common process.

Box 1 – Useful definitions

evolutionary transformation towards molecular order. The approach complements classical methods of primary sequence comparison and uses statistical mechanics considerations (6) for character argumentation, but harbors several unique and valuable features. It produces rooted topologies capable of establishing direction of evolutionary change, a feature that can be very useful in the study of intractable problems such as the rooting of the universal tree of life. Furthermore and most importantly, the method enables a direct phylogenetic analysis of function embedded in molecular structure. For example, we were able to study the origin and diversification of ribosomal RNA (rRNA) directly at the structural level (7). The evolution of the complete repertoire of structural ribosomal characters was formally traced lineage-by-lineage in a universal tree that was reconstructed from the combined secondary structure of rRNA subunits. Character tracing revealed patterns of evolution in inter-subunit bridge contacts and transfer RNA (tRNA) binding sites that were consistent with the coupling of tRNA translocation and subunit movement during protein synthesis (8). This approach has inspired a similar evolutionary study, this time focusing on protein molecules (9).

Proteins display unique three-dimensional structures and contain *protein domains* (Box 1), minimal building blocks that share common ancestry (10,11). These building blocks can be unified into a comparatively small set of folding architectures (12,13). The universe of protein topology is therefore finite and structurally redundant, with *protein folds* (Box 1) being amongst the most conserved elements in biology. Several approaches have been used to characterize protein space, such as fold family trees (14-16) or taxonomies based on secondary structure (17). Recently, a metric comparison of structure similarity of proteins representing different protein fold categories provided measurements of distance between the different structures and a global representation of protein space (18). Four clear and separate groups representing the α/β, α+β, all-α, all-β protein classes were evident in this geometric representation. This result is important and shows it is possible to generate global views of protein structure. We have organized the universe of protein architecture at an evolutionary level, studying protein architectures in over 30 genomes that have been completely sequenced (9). Here, phylogenetic trees of organisms and architectures describe histories of architectural diversification of protein groups or entire protein complements (proteomes). Figure 1 shows a rooted phylogeny of protein architectures showing the course of architectural diversification in life. Interestingly, structural classes of globular proteins appeared early in evolution and in defined order, the α/β class being the first, followed by the α+β class, the all-α class, the all-β class and small (S) and multidomain (M) proteins in that order. This finding supports the idea that the most primitive proteins contained interspersed a-helical and b-sheet elements (as in the α/β class) that were segregated in the course of evolution, an idea that is consistent with the random origin hypothesis of protein architecture (19). The figure also shows overlapped a graphical representation of proteome diversification inferred from our phylogenomic analyses that suggest dramatic diversification events in the history of life and a common ancestor with a eukaryotic-like, gene-rich, and relatively modern organization (3,7,9). Proteome and architectural diversification may have started early in the RNA world and appears to have preceded organismal diversification (9; unpublished data), defined by the appearance of at least two of three organismal domains that were eukaryotic and prokaryotic-like in nature. In this regard, a viral origin of the DNA replication machinery has been postulated (20). Note however that the RNA or DNA nature of the last universal cellular ancestor remains highly controversial.

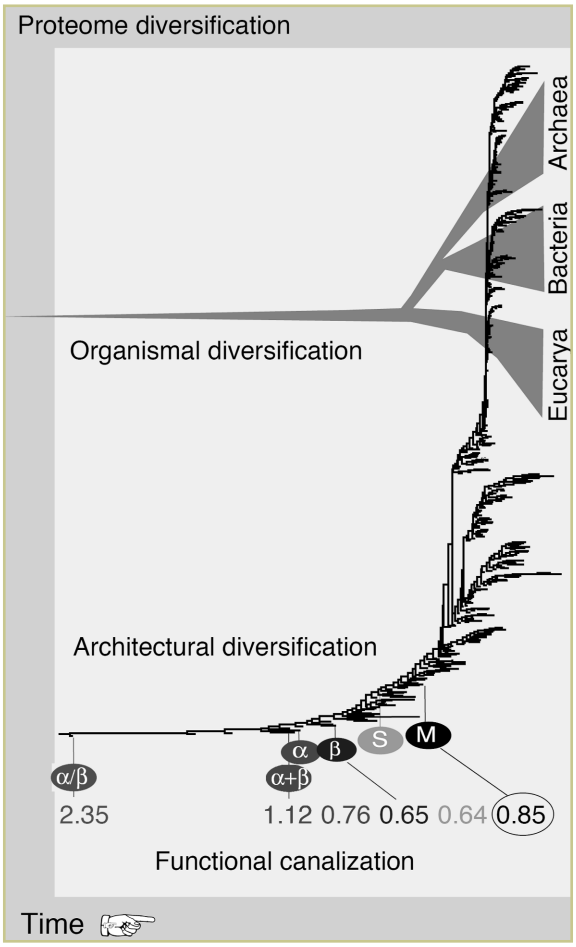

Figure 1 – Evolution of the universe of protein architecture. The phylogram shows the evolution of the different protein folds described in the Structural Classification Of Proteins (SCOP) database (release 1.59). These folds define a universe of protein architecture and are here arranged in a hierarchical representation that illustrates their evolution from a common ancestral architecture (in the base of the tree) to existant versions (in its branches). Note how the different classes of proteins arise in defined order, starting with the α/β class and ending with the small (S) and multidomain (M) classes, and how the average number of enzymatic functions associated with them appears to decrease with evolutionary time (with one exception). This tree is overlapped to a cartoon that illustrates the concurrent diversification of proteomes evident by the construction of whole genome trees. Results suggest that the diversification of proteomes and protein architectures preceded that of organisms (9; unpublished data).

The evolutionary sequence of architectural diversification here proposed entails a tendency towards *modularity* (Box 1) in the design of proteins. This tendency is consistent with patterns of modularity and simplification in molecular design (21,22) recently revealed in rRNA structure (7). Modularity appears to rise spontaneously in evolutionary systems in response to variation (23), may be pervasive, and may in itself increase organismal diversity (24). It is an intrinsic property of genetic systems, and expresses itself in processes such as the tight coordination of the transcription of genes linked by common function in a network (*synexpression*; Box 1) (25) or the modular gain and loss of functionally interacting genes (26). In proteins, theoretical arguments indicate that molecules that participate in many interactions are difficult to evolve (27). Consequently, the rise of modular blocks appears to foster lock-in mechanisms that suppress variation and favor the preservation of the modular structure. Our phylogenomic analysis supports this concept by showing that the number of enzymatic functions associated with individual architectural classes has decreased in the course of evolution (Figure 1). This 'functional canalization' appears to be a consequence of the rise of the individual modules. Interestingly, the multidomain (M) protein class that originated late in evolution had an increased number of associated functions. Proteins in this class result from the unique combination of domains that belong to different classes. Therefore, module combination appears to enhance functional diversity, but this trend seems to be incipient in the current protein universe. Overall results suggest that modularity constitutes a preferred outcome, and that its expression in protein architecture confers innovation advantages in function and design.

We are currently studying protein evolution on a broader scale. Functional, biophysical, and statistical features are being traced in 'universal' architectural trees. Models of molecular change are being inferred, using principles drawn from disparate disciplines (e.g. statistics, thermodynamics, molecular mechanics). Ultimately we want to further our understanding of our natural world, by concentrating on its history, focusing on patterns and processes, and predicting outcomes. The direct evolutionary linking of structure and function is key to our understanding of cellular functions and how these evolve. Consequently, research in this field will have lasting impact in biology and the biomedical sciences, benefiting important challenges such as the identification of drug targets or evolutionary processes in disease. Ultimately, we want to identify proteins that share evolutionary paths, trace functions that are related to structure in our evolutionary maps, and design motifs harboring defined functions.

## REFERENCES

1. S.K. Burley and J.B Bonanno; Structuring the universe of proteins. *Annu. Rev. Genomics Hum. Genet.* **3** 243-262 (2002).
2. C. Zhang and S.H. Kim; *Curr. Op. Chem. Biol.* **7** 28-32 (2003).
3. G. Caetano-Anollés; *J. Mol. Evol.* **54** 333-345 (2002).
4. The Zuckerkandl Prize; J. Mol. Evol. **56** 373-374 (2003).
5. D.D. Pollock; *J. Mol. Evol.* **56** 375-376 (2003).
6. W. Fontana; *BioEssays* **24** 1164-1177 (2002).
7. G. Caetano-Anollés; *Nucleic Acids Res.* **30** 2527-2587 (2002).
8. M.M. Yusupov, G.Z. Yusupova, A. Baucom, K. Lieberman, T.N. Earnest, J.H.D. Cate and H.F. Noller; *Science* **292** 883-896 (2001).
9. G. Caetano-Anollés and D. Caetano-Anollés, *Genome Res.*13 1563.
10. R.F. Doolittle; *Annu. Rev. Biochem.* **64** 287-314 (1995).
11. C.P. Ponting and R.R. Russell; *Annu. Rev. Biophys. Biomol. Struct.* **31** 45-71 (2002).
12. A. Murzin, S.E. Brenner, T. Hubbard and C. Clothia; *J. Mol. Biol.* **247** 536-540 (1995).
13. M.B. Swindells, C.A. Orengo, D.T. Jones, E.G. Hutchinson and J.M. Thornton; *BioEssays* **20** 884-891 (1998).
14. A.V. Efimov; *Proteins* **28** 241-260 (1997).
15. C. Zhang and S.H. Kim; *Proteins* **40** 409-419 (2000).
16. W.R. Taylor; *Nature* **416** 657-660 (2002).
17. T. Przytycka, R. Aurora and G.D. Rose; *Nature Struct. Biol.* **6** 672-682 (1999).
18. J. Hou, G.E. Sims, C. Zhang and S.H. Kim; *Proc. Natl. Acad. Sci. USA* 100:2386-2390 (2003).
19. S.H. White; *Annu. Rev. Biophys. Biomol. Struct.* **23** 407-439 (1994).
20. P. Forterre; *CR Acad. Sci. Paris Life Sci.* **324** 1067-1076 (2001).
21. L.H. Hartwell, J.J. Hopfield, S. Leibler and A.W. Murray; *Nature* 402:C47-C52 (1999).
22. L.W. Ancel and W. Fontana; *J. Exp. Zool. (Mol. Dev. Evol.)* 288 242-283 (2000).
23. H. Lipson, J.B. Pollack ad N.P. Suh; *Evolution* **56** 1549-1556.
24. A.S. Yang; *Evol. Develop.* **3** 59-72 (2001).
25. C. Niehrs and N. Pollet; *Nature* **402** 483-487 (1999).
26. T. Ettema, J van der Oost and M. Huynen; *Trends Genet.* **17** 485-487 (2001).
27. D. Waxman and J.R. Peck; *Science* **279** 1210-1213 (1998).

**Department of Crop Sciences, University of Illinois**
**Urbana, IL 61801, USA**