Evolution of the Protein Repertoire

Cyrus Chothia,[1] Julian Gough,[2] Christine Vogel,[1] Sarah A. Teichmann[1]

Most proteins have been formed by gene duplication, recombination, and divergence. Proteins of known structure can be matched to about 50% of genome sequences, and these data provide a quantitative description and can suggest hypotheses about the origins of these processes.

During the course of evolution, forms of life with increasing complexity have arisen. What are the mechanisms that have produced the increases in protein repertoires that underlie the evolution of more complex forms of life? How are proteins organized to form pathways? Answers to such questions at the molecular level began to appear 40 years ago (*1*), but it is only with the advent of complete genome sequences that we have begun to get a comprehensive view.

Proteins consist of domains. A domain, as the term is used here, is an evolutionary unit whose coding sequence can be duplicated and/or undergo recombination. Small proteins contain just one domain. Large proteins are formed by combinations of domains. A domain family contains small proteins, and/or parts of larger ones, that descend from a common ancestor. Domains typically have 100 to 250 residues, though smaller and larger domains do occur.

It is now clear that the dominant mechanisms that produce increases in protein repertoires are (i) duplication of sequences that code for one or more domains; (ii) divergence of the duplicated sequences by mutations, deletions, and insertions to produce modified structures that may have useful new properties and be selected; and, in some cases, (iii) recombination of genes that results in novel arrangements of domains. These mechanisms have long been believed to be the source of new proteins, and rates at which they occur have been calculated recently (*2*). The new findings discussed here come from the use of structural information to analyze genome sequences. This provides for the first time a quantitative view of the nature and extent of these processes.

It is difficult to detect distant protein family relationships and the presence of different domains by direct comparisons of sequences. However, the presence or absence of domains and their family relationships can usually be determined if the three-dimensional structures of the proteins are known. This means that we only clearly know the family relationships and domain structures of those proteins that either have a known structure or are homologous to proteins of known structure. At present, close to 50% of the sequences in the currently known genomes are homologous to proteins of known structure (*3*). We describe how analyses of this half of the protein repertoire have given us a detailed picture of its evolution. Important discoveries have also been made from the analyses of sequences alone, without the use of structural homology, and a most useful review of this work is included in the recent book by Koonin and Galperin (*4*).

Families of Protein Domains

The evolutionary relationships of domains in proteins of known structure are described in the Structural Classification of Proteins (SCOP) database (*5*). This information can be used to infer the family relations of the domains in the genome sequences that are homologous to proteins of known structure (*3*, *6*, *7*). In vertebrates, these domains belong to one of ~750 different families in each genome, and the average number of domains in a family is close to 50; those in invertebrates come from one of ~670 families in each genome, and the average family size is close to 20 (*3*, *7*). Plants have a similar range of values. Domains in yeast and bacteria with large genomes come from ~550 families, and those in small parasitic bacteria come from ~220 families. The average size of the known protein families in these two groups is about eight and two, respectively.

In individual genomes, the number of members in the different families fits a power-law or Pareto distribution (*8*) fairly well: A few families have many members and many families have a few members. Stochastic birth, death, and innovation models have been proposed to account for this distribution (*9*–*11*). However, in our opinion, the distribution of family sizes must be mainly the result of selection for useful functions rather than a process that is largely or purely stochastic. There are clearly some families that have properties that lend themselves to a wide variety of molecular functions: for example, the large P-loop nucleotide triphosphate (NTP) hydrolase family, whose members can function as kinases with very different specificities, as different kinds of motor proteins, and as batteries to drive reactions through conformational change.

Some of the large families can have members with very diverse sequences and different functions. Proteins with sequence identities of 40% or more usually have the same function; those with 25 to 40% identity conserve broadly similar functions; and at lower identities, functions can be very different (*12*, *13*).

The larger domain families make up the bulk of the protein repertoire in each genome and are widely distributed across genomes (*3*, *7*). At present, we find 429 families whose members occur in all of the 14 known eukaryote genomes, and the members of these families form between about 80% of the domains in animals and about 90% of those in fungi

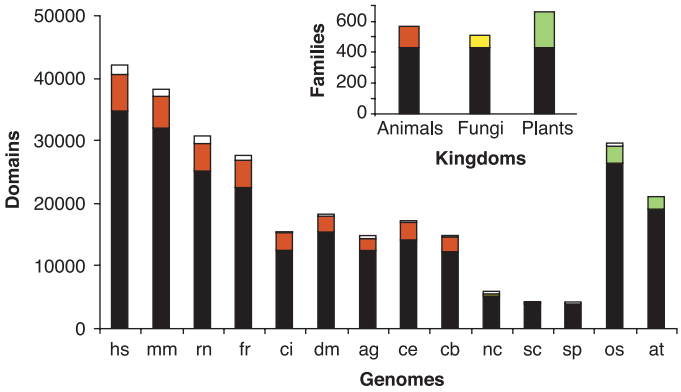

Fig. 1. Contribution of common families to the protein repertoire (*3*). All or part of about 50% of eukaryote sequences are homologous to domains in proteins of known structure. The numbers of domains that belong to the 429 families common to all 14 eukaryotes studied are shown in black. Additional contributions of families common to the genomes in only one kingdom are shown in red for animals, in yellow for fungi, and in green for plants. For the animal genomes—human (hs), mouse (mm), rat (rn), puffer fish (fr), sea squirt (ci), fruit fly (dm), mosquito (ag), and nematodes (ce and cb)—there are 136 additional common families. For the three fungi—bread mold (nc), budding yeast (sc), and fission yeast (sp)—there are 75 additional common families. For the two plants—rice (os) and cress (at)—there are 229 additional common families.

[1]Structural Studies Division, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. [2]Department of Structural Biology, School of Medicine, Stanford University, Stanford, CA 94305–5126, USA.

and plants (*3*). The nine animal genomes, from humans to *C. elegans*, have an additional 136 common families, and their members, together with those of the 429 families, form 96 to 98% of all their known domains (Fig. 1). This implies that all but a small proportion of the protein repertoire is formed by members of families that go back to the origin of eukaryotes or the origin of the different kingdoms. The remainder comes from families that have a spasmodic distribution because of gene loss, gene invention, or just because the sequence-matching procedures are not powerful enough to detect all homologs.

Domain Combinations

Thirty years ago, examination of protein structures showed that many are formed by combinations of two or more domains, and that domains from some families can combine with domains from several different families (*14*). The advent of complete genome sequences made it possible to study the extent to which this occurs. Rough estimates indicate that two-thirds of prokaryote proteins have two or more domains (*15*). In eukaryotes, in which recombination is even more common, about four-fifths of proteins are multidomain. The tendency of eukaryote proteins to have more domains than their prokaryote homologs (that is, more complex architectures and properties) has been termed "domain accretion" (*16*).

The genome sequences matched by proteins of known structure have provided the basis for detailed studies of the nature of domain combinations. Given that there are about 1100 protein families known on the basis of structure, there are potentially 1100^2 = 1,210,000 different pairwise combinations. The combinations that are actually present in genomes will be those that natural selection finds useful. An examination of proteins in 85 genomes that contain two or more domains shows that the actual number is only a tiny fraction of the potential number: A total of only 2500 different pairwise combinations were found (*17*).

A few families have members that take part in many different combinations, but most families combine with just one or two other families. The distribution of the number of combinations made by the different families is again that of a power law (*18*, *19*). There are few nodes with many links, in the case of families that have many different partners, and they include families that are useful in many different contexts, such as the P-loop NTP hydrolase domains mentioned above and the Rossmann domains. Most nodes have just one or two links. As before, we would expect this to be largely the result of selection for function.

In the large majority of cases, combinations of particular pairs of domains are found in only one sequential order: If the domains are A and B, they might occur in the order AB or BA, but very rarely in both (*18*). Case studies of domain combinations showed, first, that the sequential order of domains has little influence on their relative geometric positions, and second, that in the proteins in which the domain pairs come from the same families, the connections between the domains have similar structures (*20*). This suggests that conservation of sequential order in domain combinations is usually found because the combinations descend from a common ancestor.

So far we have only discussed pairs of domains. However, regularities of the same kind are found at the level of complete multidomain proteins. For instance, there are about 600 pairs or triplets of domains that occur in about one-third of all the currently known multidomain proteins either by themselves or in combination with other domains. We call these recurring domain combinations supradomains (*21*). An example is shown in Fig. 2.

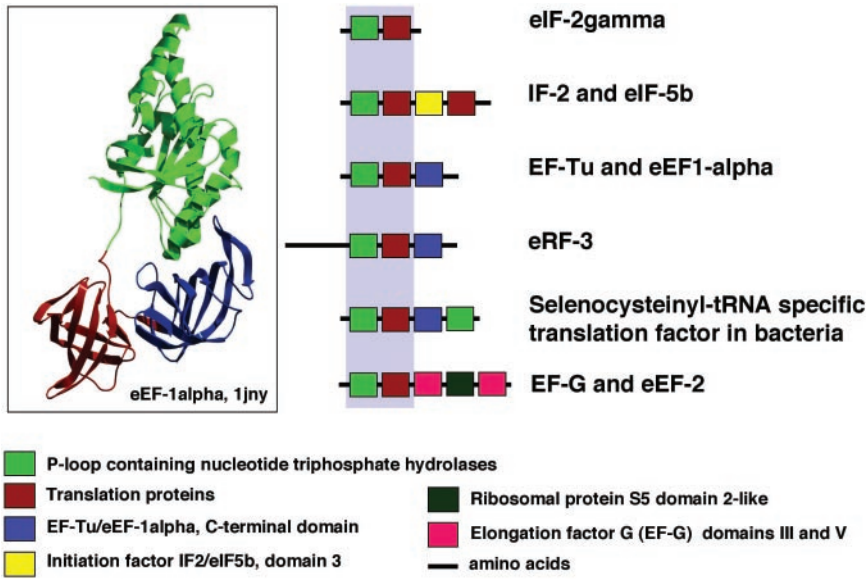

Fig. 2. An example of a supradomain. The P-loop–containing NTP hydrolase domain and the Translation Proteins domain (*5*) occur in prokaryotic and eukaryotic translation factors that hydrolyze guanosine triphosphate (GTP). GTP hydrolysis in the P-loop domain drives the conformational change in the Translation Proteins domain, which is then transmitted onto the ribosome. The supradomain occurs in 35 different domain architectures, and 6 of these are given here. The inset at left shows a protein of known structure, which contains the supradomain. IF, initiation factor; EF, elongation factor; RF, release factor; tRNA, transfer RNA.

The Formation of Metabolic Pathways

Proteins do not function by themselves but as part of an intricate network of physical complexes and pathways. How does the duplication, divergence, and recombination process fit into the formation or extension of pathways? In pathways of the same kind, such as those of small-molecule metabolism, many proteins have similar requirements in terms of functional mechanisms or substrate recognition. Thus, we might expect some fraction of these proteins to be related to one another. Two models for the evolution of pathways were developed from this idea. The first proposes that, because substrates in a pathway retain some similarities in structure, the enzymes within a pathway could evolve by gene duplications and a divergence in which their catalytic mechanisms were changed and some aspects of their recognition properties retained (*22*). The second proposed that enzymes are recruited across pathways, with the duplicated enzymes conserving their catalytic function but evolving different substrate specificities (*23*).

The pathways of small-molecule metabolism in *Escherichia coli* are the most comprehensively characterized so far (*24*), and they are a good starting point for a more detailed investigation of these models. They involve close to 600 proteins that form about 100 pathways. One-quarter of the enzymes are active in more than one pathway, so that small-molecule metabolism can, to a large extent, be viewed as a single network.

Close to 90% of the 600 proteins are homologous to a known structure (*25*). These assignments show that half of the proteins are formed by a single domain, whereas the other half contains between two and six domains. Altogether, 722 domains from one of 213 families were identified. Thus, even this basic set of ancient enzymes has evolved by means of extensive duplication and recombination.

An examination of the functions of the members of different families of domains shows that,

nearly always, it is the catalytic mechanism or cofactor-binding properties that are conserved or slightly modified and the substrate specificity that is changed (*25*). This suggests that it is much easier to evolve new binding sites than new catalytic mechanisms. Most of the members of these families are distributed across different pathways (*25–27*). There are only a tiny number of cases in which domains conserve their substrate binding properties and occur in the same pathway. An inspection of where homologs are found in the network of pathways shows that recruitment primarily occurred on the basis of catalytic mechanism or cofactor binding. This has led to a mosaic pattern of protein families with little or no coherence in the evolutionary relationships in different parts of the network.

To what extent are pathways conserved over a range of different organisms? The same pathway in different organisms can contain species-specific sets of isozymes (*28*, *29*). The comparison of enzymes in the same pathway in different organisms also shows that proteins responsible for the particular functions can belong to unrelated protein families. This phenomenon is called "nonorthologous displacement" (*30*). Variations come not just from changes in specific enzymes. In some organisms, sections of the standard pathway are not found and the gaps are bypassed through the use of alternative pathways (*28*). Together, these variations produce widespread plasticity in the pathways that are found in different organisms; much of this is described in the Clusters of Orthologous Groups (COGs) database (*4*, *31*).

For other sets of pathways, we expect duplications of the type described here, though possibly with more duplications within pathways that have arisen late in evolution, such as those of signal transduction and the immune system.

Causes and Consequences

The earliest evolution of the protein repertoire must have involved the ab initio invention of new proteins. At a very low level, this may still take place. But it is clear that the dominant mechanisms for expansion of the protein repertoire, in biology as we now know it, are gene duplication, divergence, and recombination. Why have these mechanisms replaced ab initio invention? Two plausible causes, which complement each other, can be put forward. First, once a set of domains whose functions are varied enough to support a basic form of life had been created, it was much faster to produce new proteins with different functions by duplication, divergence, and recombination. Second, once the error-correction procedures now present in DNA replication and protein synthesis were developed, they made the ab initio invention of proteins a process that is too difficult to be useful.

Consequently, even the simplest genomes of extant bacteria are the product of extensive gene duplication and recombination (*3*, *15*). An organism's complexity is not directly related to its number of genes; flies have fewer genes than nematodes, and humans have fewer than rice. However, complexity does seem to be related to expansions in particular families that underlie the more complex forms of life. This means that we will be able to trace much of the evolution of complexity by examining the duplications and recombinations of these families in different genomes (*32*, *33*).

References and Notes

1. M. F. Perutz, J. C. Kendrew, H. C. Watson, *J. Mol. Biol.* **13**, 669 (1965).
2. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
3. J. Gough *et al.*, *J. Mol. Biol.* **313**, 903 (2001). Data used here includes updated results that can be found at http://supfam.org.
4. E. V. Koonin, M. Y. Galperin, *Sequence-Evolution-Function* (Kluwer Academic, Boston, MA, 2003).
5. A. G. Murzin *et al.*, *J. Mol. Biol.* **247**, 536 (1995). Names of domain families are taken from the SCOP database at http://scop.mrc-lmb.cam.ac.uk/scop/.
6. Y. I. Wolf *et al.*, *Genome Res.* **9**, 17 (1999).
7. A. Muller *et al.*, *Genome Res.* **12**, 1625 (2002).
8. V. A. Kuznetsov, *J. Biol. Syst.* **10**, 381 (2002).
9. M. A. Huynen, E. van Nimwegen, *Mol. Biol. Evol.* **15**, 583 (1998).
10. J. Qian *et al.*, *J. Mol. Biol.* **313**, 673 (2001).
11. E. V. Koonin, Y. I. Wolf, G. P. Karev, *Nature* **420**, 218 (2002).
12. C. A. Wilson, J. Kreychman, M. Gerstein, *J. Mol. Biol.* **297**, 233 (2000).
13. A. E. Todd, C. A. Orengo, J. M. Thornton, *J. Mol. Biol.* **307**, 1113 (2001).
14. M. G. Rossmann *et al.*, *Nature* **259**, 194 (1974).
15. S. A. Teichmann, J. Park, C. Chothia, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14658 (1998).
16. E. V. Koonin, L. Aravind, A. S. Kondrashov, *Cell* **101**, 573 (2000).
17. G. Apic, W. Huber, S. A. Teichmann, *J. Struct. Funct. Genomics*, in press.
18. G. Apic, J. Gough, S. A. Teichmann, *J. Mol. Biol.* **310**, 311 (2001).
19. S. Wuchty, *Mol. Biol. Evol.* **18**, 1694 (2001).
20. M. Bashton, C. Chothia, *J. Mol. Biol.* **315**, 927 (2002).
21. C. Vogel, C. Berzuini, S. A. Teichmann, unpublished data.
22. N. H. Horowitz in *Evolving Genes and Proteins*, V. Bryson, H. J. Vogel, Eds. (Academic Press, New York, 1965), pp. 15–23.
23. R. A. Jensen, *Annu. Rev. Microbiol.* **30**, 409 (1976).
24. M. Riley, M. H. Serres, *Annu. Rev. Microbiol.* **54**, 341 (2000).
25. S. A. Teichmann *et al.*, *J. Mol. Biol.* **311**, 693 (2001).
26. S. C. G. Rison, S. A. Teichmann, J. M. Thornton, *J. Mol. Biol.* **318**, 911 (2002).
27. R. Alves, R. A. Chaleil, M. J. Sternberg, *J. Mol. Biol.* **320**, 751 (2002).
28. T. Dandekar *et al.*, *Biochem. J.* **343**, 115 (1999).
29. O. Jardine *et al.*, *Genome Res.* **12**, 916 (2002).
30. E. V. Koonin, A. R. Mushegian, P. Bork, *Trends Genet.* **12**, 334 (1996).
31. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
32. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
33. C. Vogel, S. A. Teichmann, C. Chothia, unpublished data.

34. We thank M. Madera, E. Koonin, and A. Finkelstein for comments on the manuscript. J.G. has a Burroughs-Welcome Fellowship from the Program in Mathematics and Molecular Biology, and C.V. has a Boehringer Ingelheim Predoctoral Fellowship.

VIEWPOINT

The Deep Roots of Eukaryotes

S. L. Baldauf

Most cultivated and characterized eukaryotes can be confidently assigned to one of eight major groups. After a few false starts, we are beginning to resolve relationships among these major groups as well. However, recent developments are radically revising this picture again, particularly (i) the discovery of the likely antiquity and taxonomic diversity of ultrasmall eukaryotes, and (ii) a fundamental rethinking of the position of the root. Together these data suggest major gaps in our understanding simply of what eukaryotes are or, when it comes to the tree, even which end is up.

Introduction

Molecular phylogenetic trees have gradually assigned most of the cultivated and characterized eukaryotes to one of eight major groups (Fig. 1). Although these data have largely failed to resolve relationships among these major groups, with the benefit of hindsight it was perhaps somewhat naïve that we ever thought they would. While similarities among gene sequences may indicate the relatedness of the organisms that harbor them, this relationship is far from straightforward, particularly for ancient "deep" branches. Only a fraction of sites in any gene are free to mutate, and these have only so many states (nucleotides or amino acids) to toggle through before they start repeating themselves, and their true history becomes obscured.

With more data, improved methods, and just a better idea of what we're doing, an outline of the tree seems to be emerging. This

Department of Biology, University of York, Box 373, Heslington, York YO10 5YW, UK. E-mail:slb14@york.ac.uk