More than the sum of their parts: on the evolution of proteins from peptides

Johannes Söding and Andrei N. Lupas*

Summary

Despite their seemingly endless diversity, proteins adopt a limited number of structural forms. It has been estimated that 80% of proteins will be found to adopt one of only about 400 folds, most of which are already known. These folds are largely formed by a limited 'vocabulary' of recurring supersecondary structure elements, often by repetition of the same element and, increasingly, elements similar in both structure and sequence are discovered. This suggests that modern proteins evolved by fusion and recombination from a more ancient peptide world and that many of the core folds observed today may contain homologous building blocks. The peptides forming these building blocks would not in themselves have had the ability to fold, but would have emerged as cofactors supporting RNA-based replication and catalysis (the 'RNA world'). Their association into larger structures and eventual fusion into polypeptide chains would have allowed them to become independent of their RNA scaffold, leading to the evolution of a novel type of macromolecule: the folded protein. ***BioEssays*** **25:837–846, 2003. © 2003 Wiley Periodicals, Inc.**

Introduction

Proteins are the central agents of life and their evolution is the object of intense study. An important reason for this interest lies in the extent to which we use inference from homology to explore life, based on the study of model systems. Particularly in molecular biology, searches for homologous relationships based on sequence similarity have become a routine step to gain clues about the function of a new gene.[1]

Proteins are enormously diverse. Estimates of the number of species on earth run into the millions and each species contains thousands of protein-coding genes. Though superficially different, these proteins often display substantial similarity in sequence and three-dimensional structure, since many are derived from a basic complement of autonomously folding units (domains). This allows us to group proteins into a hierarchy of families, superfamilies, and folds. The basic complement of domains was already established to a large extent at the time of the 'last common ancestor',[2] but some very successful domains arose later within the bacteria, archaea, or eukaryotes and radiated into the other kingdoms by endosymbiosis or lateral transfer.

Proteins change in the succession of generations through random drift and natural selection (the molecular clock).[3] Most frequently, changes result from point mutations (which very rarely change the overall structure of the protein significantly, but see Refs. 4 and 5 for exceptions), insertions and deletions. By these processes, proteins may become so dissimilar that their common origin cannot be detected from their sequences, even though they may still fulfill fundamentally the same function. However, their structures diverge much more slowly, providing evidence of common ancestry long after their sequence similarity has decayed.

The protein complement of an organism is the result of parental inheritance, acquisition (through lateral transfer, viruses, or mobile elements) and duplication. Duplication is central to the diversification of proteins. At the level of full genomes, duplication is an effective path to increased complexity, which has been taken repeatedly in the course of evolution.[6–8] At the level of operons, duplication may lead to the efficient evolution of novel pathways. At the level of single genes, duplication allows the emergence of systems with complex functionality, such as the vertebrate olfactory system, which is built on thousands of homologous G-protein-coupled receptors. In each of these cases, the duplicated copies are freed from the selective pressure to maintain function and in fact come under pressure to assume a novel selectable function in order to avoid extinction through mutational inactivation.[9]

Duplication, accompanied by gene fusion, is also essential for a variety of other processes that result in the generation of novel proteins, such as unequal recombination,[10] circular permutation,[12,13] and domain shuffling.[15] Unequal recombination is the primary mechanism that gives rise to repetitive proteins;[10,11] an extreme case is the giant muscle protein, titin, which consists of hundreds of immunoglobulin domains. Circular permutation is the process by which N- and C-terminal deletions in a duplicated protein can result in a structure that appears to have its C-terminal part permuted to the N

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Tübingen, Germany.
*Correspondence to: Andrei Lupas, Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstr. 35, D-72076 Tübingen, Germany. E-mail: andrei.lupas@tuebingen.mpg.de

DOI 10.1002/bies.10321
Published online in Wiley InterScience (www.interscience.wiley.com).

terminus. The importance of circular permutation for protein evolution can be appreciated from the fact that at least 412 out of 3035 domains in proteins of known structure arose by circular permutation.[14] Finally, domain shuffling[15] is the main mechanism for the rapid generation of novel domain combinations. In eukaryotes, this mechanism enabled the burst of creativity in protein evolution, which accompanied metazoan radiation during the Cambrian and yielded many novel proteins specific for multicellular organisms by combining a limited set of modular domains.[16] For example, the vertebrate immune system uses a handful of domain types in nearly endless variations in order to satisfy the extremely complex requirements of self–nonself recognition. Ironically, prokaryotes use the same mechanism to produce the variability in their surface proteins required to evade the immune system. An important effect of domain shuffling is that proteins that are not homologous globally may well contain homologous domains. For this reason, protein classification schemes build on domains, not on entire proteins.

Domain classification

The sequences and structures of domains reflect the evolutionary events that shaped them and retain the traces of their common ancestry. This is the basis of their classification into families and superfamilies[17,18] in a way analogous to the classification of organisms into genera and orders. Superfamilies are further grouped into folds according to the similarity between the arrangement of their main, consecutive elements of secondary structure in space. Whereas superfamilies are made of homologous families, i.e., they represent the result of divergent evolution, folds typically group together analogous, i.e., convergently evolved superfamilies. However, advances in genomics and bioinformatics have made it possible to detect more and more ancient evolutionary events and, as a result, many superfamilies originally considered to be analogous are now recognized as homologous. Indeed, this is occasionally even the case for superfamilies from different folds (for example due to circular permutation). Recognition of these distant evolutionary events has gradually made it possible to describe the basic complement of protein domains that was present in the last common ancestor.[19,20]

In the early 1950s, the discovery of the basic principles of protein structure by Pauling, Crick and Ramachandran raised hopes that the diversity of proteins could be explained by a few universal rules in the same way that the structure of DNA had unified genetic diversity. The first crystal structures of proteins disappointed these hopes, showing seemingly chaotic architectures but, with time, it became clear that proteins prefer certain folds over others. Indeed, some folds turned out to be quite popular. A quarter of all domains of known structure assume one of ten folds, termed 'superfolds' (Table 1),[21] and estimates suggest that 80% of all protein domains will be found to fold into one of 400 'mesofolds'.[22] This is surprising, given the estimated large number of 10 000 folds that are probably populated by existing proteins, most of which occur only in a single family ('unifolds').

Table 1. Superfolds and the fraction of their residues contained in the supersecondary structure elements αα, ββ, βαβ[21]

	Internal symmetry			
Fold	**Sequence***	**Structure**	**Number of superfamilies (%)†**	**% Supersecondary structure content**
β-trefoil	+	+	2 (0.1)	83
Jelly roll	–	+	17 (1.2)	47
Immunoglobin-like	–	+	55 (4.0)	67
TIM-barrel	+	+	28 (2.0)	82
Ferredoxin-like	+	+	65 (4.7)	38
Updown bundle	+	+	17 (1.2)	90
OB fold	–	–	16 (1.1)	77
UB-roll	–	–	16 (1.1)	55
Globin-like	–	–	4 (0.3)	88
Doubly wound	–	–	122 (8.8)	68
All superfolds			342 (24.7)	65
All folds			1386 (100)	62

*'+' signifies that at least one member of the superfold displays internal sequence similarity.

†Superfolds were defined by Orengo, Jones and Thornton from a census of the CATH database in 1994.[80] The numbers of superfamilies in each fold, used to determine the most populated folds, were recompiled for this table from the most recent release of the CATH database (version 2.4; http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html). Note that the β-trefoil and globin folds do not qualify as superfolds any more, their place now taken by αα-solenoids (CATH: 1.25.40) with 13 superfamilies, and by nucleic acid-binding 3-helical bundles (CATH: 1.10.10), SH3-type barrels (CATH: 2.30.30), and SAM domain-like orthogonal helical bundles (CATH: 1.10.150), each with 11 superfamilies. In our census, we did not consider the CATH archtectures 'single α-helix' (1.20.5) and 'helix hairpin' (1.20.15), which do not represent folds in a stricter sense.

Why are some folds so common? Certainly stability and folding efficiency have contributed.[21,23] Some folds may also have yielded better scaffolds for the establishment of active sites, displacing less-suitable folds (fold competition).[24] A further reason may lie in the fact that most folds seem to consist of a limited number of compact, recurring, subdomain-sized fragments, termed supersecondary structures (Fig. 1a). These have been found in many studies, theoretical as well as experimental.[17,19,21–25] The most frequent supersecondary structures are ββ-hairpins, αα-hairpins, and βαβ-elements, and the average fold consists to more than 60% of these elements (Table 1). It is attractive to consider that the limited number of observed domain topologies may be due to the evolution of protein domains from a limited 'vocabulary' of such supersecondary structure elements. In the following sections, we will review the evidence in favor of this hypothesis.

Repetition in the evolution of domains

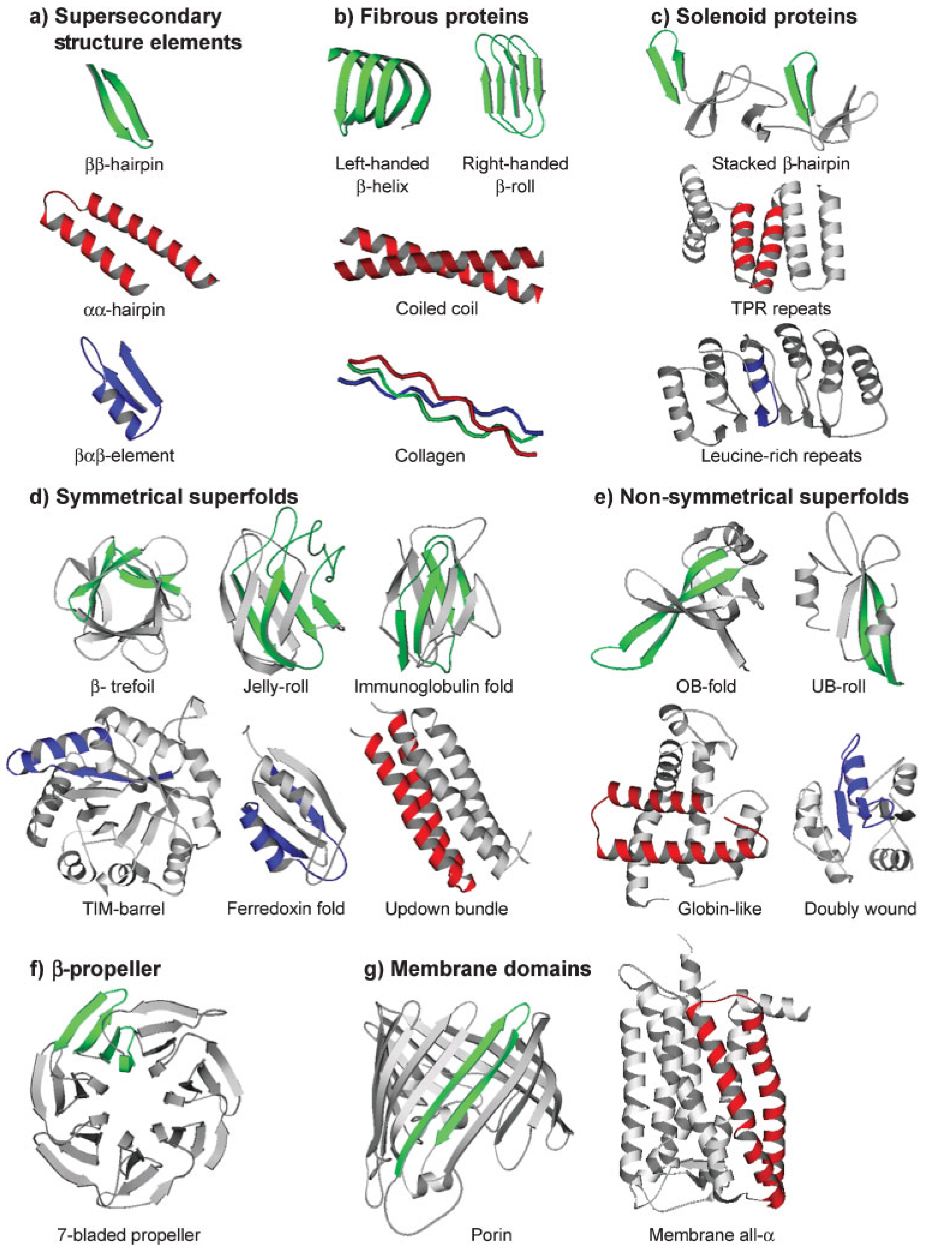
The basic processes of mutation, duplication, and shuffling have led from a set of ancestral domains to the complex proteins observed today. In what ways did these ancestral domains arise? An example of recent de novo evolution of a protein may yield clues to this question.

Arctic and Antarctic fish have evolved a variety of anti-freeze proteins that enable them to live below 0°C. One of these, the serum anti-freeze glycoprotein of Antarctic notothenioid fish, could be shown to have evolved recently (about 10 million years ago) from a pancreatic trypsinogen gene.[28] The 5′ and 3′ regions of the gene contributed the secretory signal and the 3′ untranslated region, respectively, and the coding sequence was generated by the amplification of a nine nucleotide fragment of the trypsinogen precursor. The new protein consists of Thr-Ala-Ala repeats, in which the threonines are glycosylated, thus positioning the sugars into the proper spacing for interaction with the ice lattice and preventing its growth. Astonishingly, a very similar tripeptide-based anti-freeze protein with threonines modified by the same gycosyl moieties has evolved convergently in arctic cod fish.[29] This protein shows no connections with trypsinogen. There are only few examples as yet where the de novo emergence of new proteins has been documented, but this case highlights the ability of repetition to generate novel structures and the frequency of this process.

In the example above, a novel protein arose from the repetition of a simple tripeptide fragment. Many of the most important fibrous proteins today show a similar construction based on short repeated peptides (Fig. 1b): (1) Collagen is composed of three parallel polyproline type II helices, each formed by hundreds of proline- and hydroxyproline-rich Gly-X-Y repeats. The regular spatial arrangement results in a tensile strength that is higher than that of steel. (2) Coiled coils are formed by amphipathic α-helices that associate via their hydrophobic faces in a 'knobs-into-holes' side-chain arrangement.[30] They are primarily built on a repetitive pattern of seven residues in which the first and fourth residues are hydrophobic and the others are hydrophilic. Variations on this pattern, which result from the insertion of three or four residues, are also common. The ubiquity and diversity of coiled-coil domains in today's proteins suggests that coiled coils have been invented on multiple occasions in evolution. Indeed, some coiled-coil sequences show near-identity in their repeats, pointing to a recent de novo amplification from a single repeat unit. (3) Finally, parallel β-helices are formed by stacked coils of two or three β-strands. Many are built of short repeats. For example, the β-roll, a right-handed β-helix composed of two β-strands per turn, has a 9-residue sequence periodicity and the left-handed β-helix, composed of three β-strands per turn, has a 6-residue periodicity. In all these cases, the repeating units are short and have only loose sequence requirements so that analogous evolution on multiple occasions seems probable. The repeating units that gave rise to fibrous proteins may have originated from the coding regions of previously existing proteins, as in the case of the anti-freeze protein, but may also have come from non-coding regions or out-of-frame translation. The evolutionary success of repetitive proteins results from the fact that repetition intrinsically promotes stability through the periodic recurrence of favorable interactions.[31]

Fibrous proteins are generated by the repetition of short peptide segments. Greater structural variability can be obtained by the repetition of larger units with defined secondary structure. Here, the most common units correspond to the same three supersecondary structures described earlier (Fig. 1a). At the simplest level, the monotonous repetition of one supersecondary structure element generally gives rise to open-ended, solenoid structures (Fig. 1c): TPR-, HEAT-, Armadillo-, and Ankyrin-repeat proteins are formed of stacked αα-hairpins, leucine-rich repeat proteins are formed of βαβ elements, and bacterial choline-binding domains are formed of ββ-hairpins.

In some cases, repetition of these elements may lead to closed, globular structures (Fig. 1d–g): for example TIM barrels (βαβ), and β-propellers (ββ), both of which have yielded useful scaffolds for the emergence of catalytic activity and thus help trace a path of increased complexity from repetitive proteins to fully differentiated enzymes. In β-propellers, individual repeats span the range from near sequence identity to complete dissimilarity, providing clear evidence for their origin from identical repeats.[32] In TIM barrels, the repeat units are only recognizable at the structural level and the sequences are so different that evolution by repetition cannot be assumed as proven. Recently, however, Sterner and coworkers discovered TIM barrels with internal sequence symmetry and succeeded in constructing homo- and heterodimeric barrels, supporting the notion of evolution from repetitive, oligomeric


a) Supersecondary structure elements
ββ-hairpin
αα-hairpin
βαβ-element
b) Fibrous proteins
Left-handed β-helix
Right-handed β-roll
Coiled coil
Collagen
c) Solenoid proteins
Stacked β-hairpin
TPR repeats
Leucine-rich repeats
d) Symmetrical superfolds
β- trefoil
Jelly-roll
Immunoglobulin fold
TIM-barrel
Ferredoxin fold
Updown bundle
e) Non-symmetrical superfolds
OB-fold
UB-roll
Globin-like
Doubly wound
f) β-propeller
7-bladed propeller
g) Membrane domains
Porin
Membrane all-α

precursors.[33] Homo-oligomeric precursors pose substantial problems for open-ended repetitive proteins, since they lead to fibrous structures of indeterminate length (although such structures can become very useful when their growth can be regulated, as in actin and tubulin filaments), but in globular domains that retain internal structural symmetry, such precursors are not only possible, but indeed likely. Conversely, the existence of some domains in both oligomeric and fused form is clear evidence for such an origin.

The proposal that internal structural symmetry in a modern protein may be the result of evolution through duplication and gene fusion from an ancestral homodimer dates back to the 1970s and the work of Andrew McLachlan. In a string of articles spanning over ten years, he described protein evolution by repetition at all levels of complexity[34,35] from fibrous proteins (collagen and coiled coils, Refs. 36–39), to individual domains (β-trefoil, Ref. 40 four-helix bundles, Ref. 41 immunoglobulin-like, Ref. 42), to multidomain proteins (serine and aspartic proteases, Refs. 43,44 hexokinase, Ref. 45). In several instances, he showed how a modern multidomain protein could have arisen by successive gene duplications from a subdomain-sized fragment. In more general terms, his work and that of many others has shown that repetition is not only an important mechanism in the evolution of multidomain proteins, but also in the evolution of the domains themselves. Of the ten superfolds defined by Thornton and coworkers,[21] six have internal structural symmetry (Fig. 1d). Four-helix bundles with up-and-down topology are still found in homotetrameric and homodimeric forms, providing the strongest evidence for evolution by duplication. Clear internal sequence symmetry is also seen in proteins with ferredoxin, β-trefoil and TIM barrel folds, although homo-oligomeric forms of these folds are not currently known. Finally, internal symmetry in proteins with immunoglobulin and jelly-roll folds can only be seen on the structural level (but until recently this was also the case for TIM barrels and trefoil folds). The same repetitiveness is detectable in a large class of membrane-embedded domains, the porin β-barrels of bacteria and organelles (Fig. 1g). These are formed of between four and eleven ββ-hairpins in a circular arrangement. The variability in the number of repeat units, the existence of homo-oligomeric structures, and subtle but significant sequence similarity between hairpins all point to an evolution of β-barrels from a basic ββ-hairpin motif.

Proteins from pieces

In the previous section, we have discussed repetition, i.e., the assembly of domains from identical components. Assembly from non-identical components can be viewed as an equally powerful factor in the evolution of protein domains. The four superfolds that do not show internal symmetry are composed to over 70% of the same three supersecondary structure elements[21] (Fig. 1e), that play such a dominant role in repetitive proteins. All-α membrane-embedded domains, which do not show the internal symmetry visible in ββ-hairpin membrane domains, are nonetheless almost exclusively composed of one supersecondary structure motif, the αα-hairpin (Fig. 1g). Other studies give substantial support to the relevance of local, recurring units for protein structure: short protein fragments clustered by sequence similarity occur to a significant extent in one defined conformation;[46] searches for independently folding units ('foldons') in non-homologous proteins yield a limited set of recurring structures;[26] attempts to form folded and functional hybrid proteins by recombination of homologous proteins showed that successfully recombined fragments ('schemas') corresponded essentially to known supersecondary structure elements.[27]

The idea that the first protein domains evolved by recombination from a limited 'vocabulary' of such structural units is attractive for several reasons: (1) The combinatorial complexity of evolving a whole domain in one piece is forbidding; there are 20^{100} possible sequences for a domain of 100 residues, and only a negligible fraction of these will be able to fold, let alone display a biological activity. By comparison, the sequence optimization of the approximately 20 residues required to make a supersecondary structure unit is well within the reach of biological systems. The difference in combinatorial complexity is of the same order as that between the mass of an electron and that of the entire universe. (2) No non-biological mechanisms are known that could produce polypeptide chains of sufficient length to form a domain at all, much less in the quantities required to explore the available sequence space to any significant extent, but the prebiotic synthesis of short peptides seems a realistic assumption.[47] Thus, it seems much more likely that the evolutionary process started from short, rather than long peptide chains. (3) Even if de novo evolution of entire domains were possible, it would be highly inefficient relative to the assembly from modules, which

Figure 1. Proteins from pieces. The panels show **a:** the three most important supersecondary structures; **b:** the main fibrous proteins (left-handed β-helix, 1L0S; right-handed β-roll, 1SAT; coiled coil, 1ZIK; collagen, 1BKV); **c:** solenoid proteins formed by repetition of supersecondary structure elements (stacked β-hairpin, 1HCX; TPR repeat, 1ELR; leucine-rich repeat, 1A4Y); **d:** superfolds with recognizable internal symmetry; the repeat unit is colored (β-trefoil, 4FGF; jelly-roll, 1GOH; immunoglobulin-like, 1JP5; TIM barrel, 1HTI; ferredoxin-like, 1APS; up-and-down four-helix bundle, 1RPR); **e:** superfolds without recognizable internal symmetry; some supersecondary structures are colored for illustration (OB-fold, 1QVC; UB-roll, 1LKK; globin, 1EBC; doubly wound, 5CHY); **f:** a β-propeller (1TBG); and **g:** the two types of membrane proteins: all-β (porin, 2POR) and all-α (rhodopsin, 1L9H).

allows the utilization of individually optimized characteristics at a higher level of complexity. Natural systems use modularity at many levels[48,49] ranging from the molecular level to the construction of multicellular organisms. At the protein level, theoretical studies have shown that shuffling of subdomain-sized modules can vastly accelerate the evolution of folded structures.[50,51] An experimental study designed to test the efficiency with which folded proteins could be obtained by recombination of heterologous fragments found that the success rate may be as high as one in 10^7 events.[52] This would mean, very roughly, given the enormous variations in population size, generation time and levels of intracellular recombination, that a bacterial species would be able to create hundreds to thousands of new folding proteins a year through the recombination of heterologous genes (illegitimate recombination). Even though only a very small fraction of these will become established in the population, the cumulative effect over evolutionary time scales could be substantial. (As an aside, this may be one source for the large number of singletons, i.e., proteins with no known homologs even in the most closely related species that are found consistently in genome sequencing studies.) (4) A rising number of fragments from non-homologous proteins are found to be similar in sequence, structure and sometimes even function, pointing to a common origin[53] (for example two types of $\beta\beta$ element with nucleotide binding loops, an Asp-box-containing $\beta\beta$-hairpin,[54] an RNA-binding $\beta\alpha\beta$ element,[55] a Zn-binding 'treble clef finger' composed of two short $\beta\beta$-hairpins and an α -helix, Ref. 56). Continuously improving methods in the detection of distant genetic events lead us to expect that many more such fragments will be discovered. Systematic studies should allow a description of this ancient peptide set in the same way in which ancient vocabularies (indo-european, altaic, uralic) have been reconstructed from the comparative study of modern languages.

Protein evolution from fragments was first discussed broadly in the context of exon shuffling.[57–59] The discovery that the coding parts of eukaryotic genes (exons) are generally interrupted by non-coding regions (introns), which have to be spliced out of the messenger RNA, suggested an efficient mechanism, by which fragments of non-homologous genes could be recombined. The exon shuffling hypothesis proposed that primitive genomes consisted of minigenes (exons), which primarily coded for supersecondary structures, interrupted by introns, which served as recombination hotspots.[60] 20 years on, it has become clear that most introns are of comparatively recent origin[61–63] and that there is no statistical correlation between exon boundaries and supersecondary structures.[63] Exon shuffling has been found to play a major role in the diversification of multidomain proteins during the metazoan radiation, but its involvment in early protein evolution is now heavily disputed. However, the basic concept of recombining a limited set of fragments has reemerged based on new evidence and on the recognition that other genetic mechanisms, such as illegitimate recombination, are efficient enough to produce the raw material for this process.

RNA as a template for protein evolution

How did this original vocabulary of peptides arise? Because proteins have no coding properties and therefore selected phenotypes cannot be passed on, it seems necessary that they would have evolved in the context of a primitive system capable of replication and coding.[64] Of the macromolecules known to us, only RNA displays the ability to store information, replicate and catalyze chemical reactions.[65] RNA provides the key components that allow proteins to be synthesized from genetic information; these components are highly conserved in all living beings, pointing to their ancient origin and fundamental role.[66] Indeed, first speculations about an ancient living world built on RNA date back to the elucidation of translation and gained substantial momentum with the discovery of catalytic RNAs (ribozymes), culminating in the recent structure determination of the ribosome.[67] 'At present, there are no serious alternatives to an RNA world being one essential intermediate stage in the origin of life'.[68]

It is difficult to reconcile the narrow catalytic spectrum of present-day ribozymes, which essentially only perform phosphodiester chemistry,[69] with the broad catalytic requirements of a whole organism. Certainly it has been possible to develop catalytic RNAs with a broader repertoire in vitro through directed evolution, but some reactions such as redox reactions involving free radicals may well remain outside their capabilities.[70] For this reason, peptides may have provided useful cofactors in an RNA world, as they are good chelators of small molecules, can assist redox reactions via their side chains, and have a natural affinity for nucleic acids. An instance in which a single amino acid, histidine, acts as a cofactor for a catalytic DNA with RNA-cleaving activity has been discovered through in vitro selection,[71] illustrating the potential usefulness of amino acid side-chains for ribozyme catalysis.

It seems reasonable to assume that short peptides were available through abiotic synthesis[49] and that non-specific peptide–RNA complexes would have formed fairly readily. Such peptide–RNA complexes would have provided the starting point for the selection of ribozymes with improved or novel catalytic activity and may have yielded other benefits to ribozymes as well, such as increased thermal stability, improved folding, or a better ability to form oligomers. The effect of peptides and proteins on RNA folding has been studied in some detail and there is substantial evidence that non-specific interactions prevent unproductive, kinetically trapped secondary structures and specific interactions allow the efficient selection of a single tertiary structure over thermodynamically nearly equivalent competing structures.[72,73]

The emergence of useful peptide–RNA complexes would have led to a depletion of the abiotic peptide pool and to a selection for tighter and more specific peptide–RNA interactions. This in turn would have provided the selective pressure in the evolution of ribozymes capable of catalyzing peptide synthesis and the eventual emergence of a primitive code, which would have allowed the synthesis of useful peptides in a more directed way. One hypothesis on the emergence of the genetic code proceeds specifically from the use of amino acids as ribozyme cofactors; it proposes that the amino acids were delivered to their cognate ribozymes by nucleic acid adaptors ('coding coenzyme handles'), which represented the first form of tRNA (reviewed in Ref. 74).

The initial peptide world would not have required folding for its activity. However, peptides have a natural propensity towards forming secondary structures and nucleic acids have been shown to enhance this propensity. A survey of peptide–RNA complexes has shown that the variability of major and minor grooves in RNA molecules promotes the formation of α-helices, 3_{10}-helices, and $\beta\beta$ hairpins.[75] Secondary structure would have led to greater specificity in binding and catalysis by allowing the peptides to assume the same optimized conformations reproducibly. It therefore seems reasonable to expect that the ability to form secondary structures was under positive selection.

The next step in complexity may have been reached when the number of peptide-coding mini-genes grew to the point where efficient reproduction required their fusion into chromosomes. This process must have been largely aleatoric, resulting in the formation of many different mini-gene combinations.

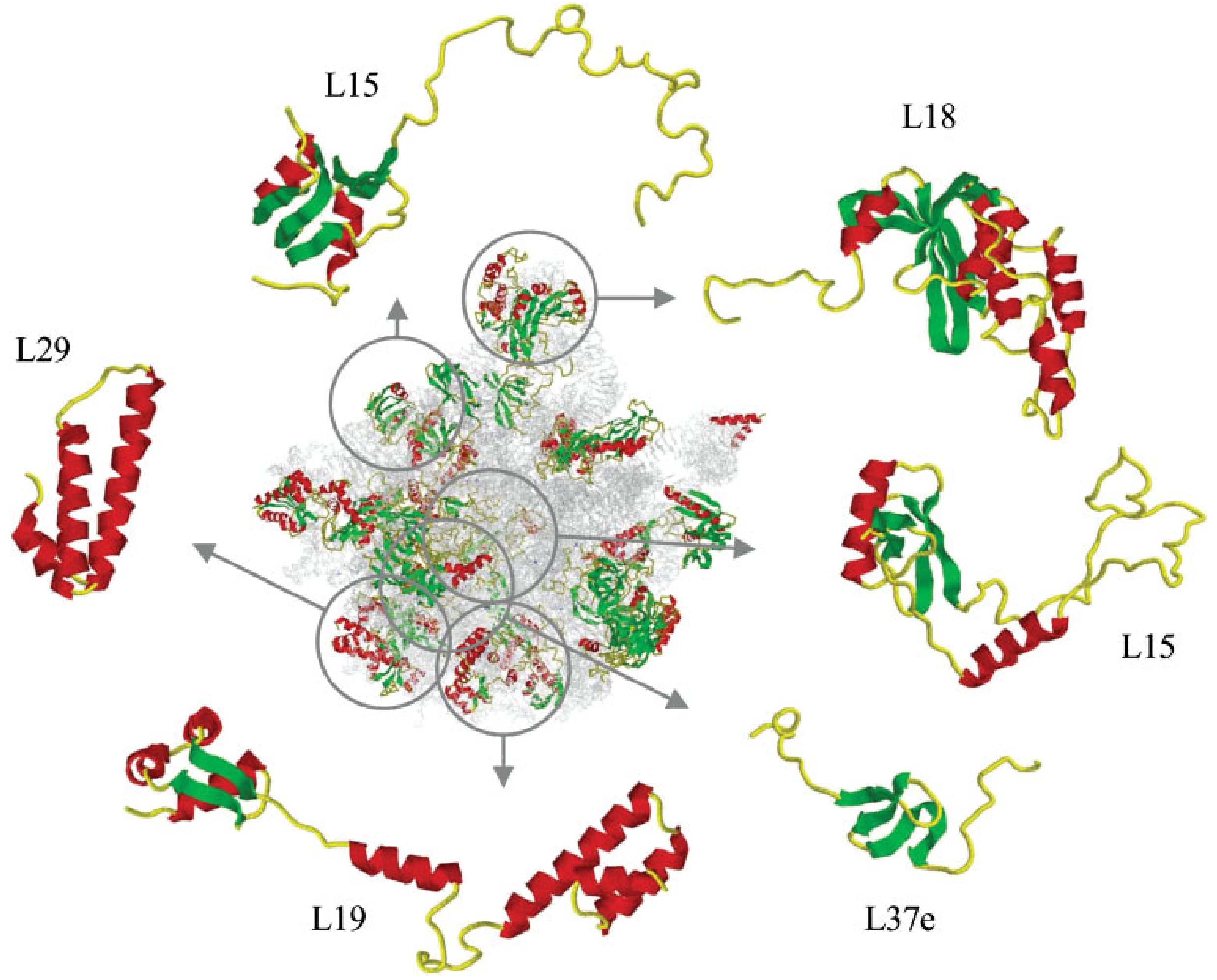


Figure 2. Structure of ribosomal proteins. Protein components of the large ribosomal subunit of the archaeon *Haloarcula marismortui* (1JJ2; center) are folded to various degrees. The long natively unfolded regions present in most ribosomal proteins are in tight association with the ribosomal RNA and extend for considerable distance into the core of the particle.

Because of the inaccuracy of replication, RNA-based organisms would have needed to maintain genes in large copy numbers[76] and would therefore most likely have retained many largely redundant chromosomal variants. Frequent read-through due to the inaccuracy of the primitive code would then have yielded peptides long and diverse enough to allow the selection of the first proteins, i.e., polypeptide chains capable of assuming a fold independently of their RNA template. These first proteins contained peptides that had already been optimized for binding and catalysis in the context of RNA and therefore probably often displayed some catalytic activity on their own. Eventually, the higher catalytic efficiency and broader repertoire possible with polypeptide chains would have led to the gradual emancipation of proteins from ribozymes and the displacement of the latter from most biological reactions.[68–70] An example of the displacement of an RNA domain in a ribozyme to yield a catalytic ribonucleoprotein was described in the *Tetrahymena* ribozyme.[77]

Evidence for this emancipation process comes from the crystal structure of a living fossil—the ribosome.[67] In the protein world, living fossils are molecules whose role in cellular processes is so central that further modification has become nearly impossible. They are essentially frozen in time (a well-known example is that of ubiquitin, which is 97% identical in slime molds and humans). The ribosome is the central component in protein synthesis; it emerged early in the evolution of life and was essentialy fixed at the time of the last common ancestor. Correspondingly, the core complement of ribosomal proteins is more than 40% identical between all living organisms. The crystal structure shows that only few ribosomal proteins are fully folded; many have folded domains ‘sprouting’ from a part of the polypeptide chain that lacks secondary structure and is tightly associated with the RNA, and some proteins have no folded structure at all and practically no secondary structure either (Fig. 2). Cumulatively, the picture is one of progressive structural emancipation, from complete dependence on the RNA template to nearly full independence—a snapshot of the time when proteins learned to fold.

Conclusions

In this article, we have attempted to retrace the path by which proteins may have evolved. Clearly, protein evolution is still very much an ongoing process and new proteins are constantly formed. Sometimes (rarely) this occurs de novo from parts of another gene or from noncoding regions; usually it happens by duplication, mutation and shuffling from existing protein domains. Occasionally, the domains themselves are the result of de novo invention but, for the most part, they belong to an established set that evolved between the emergence of protein-based life and the divergence of the major kingdoms of organisms. Since that time, domain folds have diversified in many ways through accretion (‘piecemeal growth’ (McLachlan 1972)), helix-strand transitions, strand invasions, hairpin flips, circular permutation, and recombination with heterologous domains, sometimes to the point where only a tenuous similarity to the original fold has remained.[78,79] Domains also repeatedly converged towards particular folds, which probably represent particularly foldable and versatile structural solutions (the superfolds). But overall, more and more domains can be seen to have arisen divergently from a basic set, which has not been enriched significantly with new forms in the last two billion years.

There is a growing body of evidence suggesting that this basic set arose by duplication, mutation and shuffling of shorter fragments. Just as proteins today are often a mosaic of homologous and nonhomologous domains, so domains themselves may be mosaics of homologous and nonhomologous fragments. The fragments presumably evolved in the context of RNA-based replication and catalysis and would also have converged towards particularly foldable and versatile structural solutions (the supersecondary structures). Their combination, often through repetition of the same element, was the driving force behind the evolution of folded domains. Overall, the picture is one that is very familiar to anyone studying natural processes: the emergence of entities with progressively higher complexity based on the recombination of simpler elements.

References

1. Bork P, Koonin EV. Predicting function from protein sequences—where are the bottlenecks? Nat Genet 1998;18:313–318.
2. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124–2128.
3. Feng DF, Cho G, Doolittle RF. Determining divergence times with a protein clock: Update and reevaluation. Proc Nat Acad Sci USA 1997; 94:13028–13033.
4. Cordes MH, Burton RE, Walsh NP, McKnight CJ, Sauer RT. An evolutionary bridge to a new protein fold. Nat Struct Biol 2000;7:1129–1132.
5. Glykos NM, Cesareni G, Kokkinidis M. Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. Structure 1999;7:597–603.
6. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 1997;387:708–713.
7. Vision TJ, Brown DG, Tanksley SD. The origins of genome duplications in *Arabidopsis*. Science 2000;290:2114–2117.
8. McLysaght A, Hokamp K, Wolfe KH. Extensive genomic duplication during early chordate evolution. Nature Genet 2002;31:200–204.
9. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science 2000;290:1151–1155.
10. Marcotte EM, Pelligrini M, Yeates TO. A census of protein repeats. J Mol Biol 1998;293:151–160.
11. Kajava AV. Proteins with repeated sequence—structural prediction and modeling. J Struct Biol 2001;134:132–144.
12. Ponting CP, Russell RB. Swaposins: circular permutations within genes encoding saposin homologues. Trends Biochem Sci 1995;20: 179–180.
13. Lindqvist Y, Schneider G. Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 1997;7:422–427.
14. Jung J, Lee B. Circularly permuted proteins in the protein structure database. Prot Sci 2001;10:1881–1886.
15. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem 1995;64:287–314.

16. Patthy L. Genome evolution and the evolution of exon-shuffling—a review. Gene 1999;238:103–114.
17. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
18. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
19. Aravind L, Mazumder R, Vasudevan S, Koonin E. Trends in protein evolution inferred from sequence and structure analysis. Curr Opin Struct Biol 2002;12:392–399.
20. Poole A, Jeffares D, Penny D. Early evolution: prokaryotes, the new kids on the block. BioEssays 1999;21:880–889.
21. Salem GM, Hutchinson EG, Orengo CA. Correlation of observed fold frequency with the occurrence of local structural motifs. J Mol Biol 1999; 287:969–981.
22. Coulson AFW, Moult J. A unifold, mesofold and superfold model of protein fold use. Proteins 2002;46:61–71.
23. Govindarajan S, Goldstein R. Why are some protein structures so common? Proc Natl Acad Sci USA 1996;93:3341–3345.
24. Ponting CP, Russell RB. The natural history of protein domains. Annu Rev Biophys Biomol Struct 2002;31:45–71.
25. Holm L, Sander C. Dictionary of recurrent domains in proteins. Proteins 1998;33:88–96.
26. Panchenko A, Luthey-Schulten Z, Cole R, Wolynes PG. The foldon universe: a survey of structural similarity and self-recognition of independently folding units. J Mol Biol 1997;272:95–105.
27. Voigt C, Martinez C, Wang Z-G, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. Nat Struct Biol 2002;9:553–558.
28. Chen L, DeVries AL, Cheng CH. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci USA 1997;94:3811–3816.
29. Chen L, DeVries AL, Cheng CH. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc Natl Acad Sci USA 1997;94:3817–3822.
30. Lupas A. Coiled coils: new structures and new functions. Trends Biochem Sci 1996;21:375–382.
31. Kohl A, Binz HK, Forrer PF, Stumpp MT, Plückthun A, Grütter MG. Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. Proc Natl Acad Sci USA 2003;100:1700–1705.
32. Fülöp V, Jones DT. β Propellers: structural rigidity and functional diversity. Curr Opin Struct Biol 1999;9:715–721.
33. Höcker B, Beismann-Driemeyer S, Hettwer S, Lustig A, Sterner R. Dissection of a $(\alpha\beta)_8$-barrel enzyme into two folded halves. Nature Struct Biol 2001;8:32–36.
34. McLachlan AD. Repeating sequences and gene duplication in proteins. J Mol Biol 1972;64:417–437.
35. McLachlan AD. Gene duplication and the origin of repetitive protein structures. Cold Spring Harbor Symp Quant Biol 1987;7:411–420.
36. McLachlan AD. Evidence for gene duplication in collagen. J Mol Biol 1976;107:159–174.
37. McLachlan AD, Stewart M, Smillie LB. Sequence repeats in alpha-tropomyosin. J Mol Biol 1975;98:281–291.
38. McLachlan AD. Repeated helical pattern in apolipoprotein-A-I. Nature 1977;267:465–466.
39. McLachlan AD. Analysis of gene duplication repeats in the myosin rod. J Mol Biol 1983;169:15–30.
40. McLachlan AD. Three-fold structural pattern in the soybean trypsin inhibitor (Kunitz). J Mol Biol 1979;133:557–563.
41. McLachlan AD, Bloomer AC, Butler PJ. Structural repeats and evolution of tobacco mosaic virus coat protein and RNA. J Mol Biol 1980;136:203–224.
42. McLachlan AD. Repeated folding pattern in copper-zinc superoxide dismutase. Nature 1980;285:267–268.
43. McLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
44. Blundell TL, Sewell BT, McLachlan AD. Four-fold structural repeat in the acid proteases. Biochim Biophys Acta 1979;580:24–31.
45. McLachlan AD. Gene duplication in the evolution of the yeast hexokinase active site. Eur J Biochem 1979;100:181–187.
46. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
47. Keller M, Blöchl E, Wächtershauser G, Stetter KO. Formation of amide bonds without a condensation agent and implications for origin of life. Nature 1994;368:836–838.
48. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47–C52.
49. Wagner GP, Altenberg L. Complex adaptions and the evolution of evolvability. Evolution 1996;50:967–976.
50. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscape. Proc Natl Acad Sci USA 2002;99:809–814.
51. Bogarad LD, Deem MW. A hierarchical approach to protein molecular evolution. Proc Natl Acad Sci USA 1999;96:2591–2595.
52. Riechmann L, Winter G. Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. Proc Natl Acad Sci USA 2000;97:10068–10073.
53. 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.
54. Copley RR, Russell RB, Ponting CP. Sialidase-like Asp-boxes: sequence similar structures within different protein folds. Prot Sci 2001;10:285–292.
55. Grishin NV. KH domain: one motif, two folds. Nucleic Acids Res 2001; 29:638–643.
56. Grishin NV. Treble clef finger—a functionally diverse zinc-binding structural motif. Nucleic Acids Res 2001;29:1703–1714.
57. Gilbert W. Why genes in pieces? Nature 1978;271:501.
58. Doolittle WF. Genes in pieces: were they ever together? Nature 1978; 272:581–582.
59. Blake CCF. Do genes-in-pieces imply proteins in pieces? Nature 1978;273:26.
60. Gilbert W. The exon theory of genes. Cold Spring Harbor Symp Quant Biol 1987;7:901–905.
61. de Souza SJ, Long M, Klein RB, Roy S, Lin S, Gilbert W. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 1998;95:5094–5099.
62. Cho G, Doolittle RF. Intron Distribution in ancient paralogs supports random insertion and not random loss. J Mol Evol 1997;44:573–584.
63. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF. Testing the exon theory of genes: The evidence from protein structure. Science 1994;265:202–207.
64. Orgel L. The origin of life—a review of facts and speculations. Trends Biol Sci 1998;23:491–495.
65. Yarus M. Boundaries for an RNA world. Curr Opin Chem Biol 1999;3: 260–267.
66. Anantharaman V, Koonin E, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 2002;30:1427–1464.
67. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 2000;289:905–920.
68. Jeffares D, Poole A, Penny D. Relics from the RNA world. J Mol Evol 1998;46:18–36.
69. Doudna JA, Cech TR. The chemical repertoire of natural ribozymes. Nature 2002;418:222–228.
70. Joyce GF. The antiquity of RNA-based evolution. Nature 2002;418:214–221.
71. Roth A, Breaker RR. An amino acid as a cofactor for a catalytic polynucleotide. Proc Natl Acad Sci USA 1998;95:6027–6031.
72. Herschlag D. RNA chaperones and the RNA folding problem. J Biol Chem 1995;270:20871–20874.
73. Clodi E, Semrad K, Schroeder R. Assaying RNA chaperone activity *in vivo* using a novel RNA folding trap. EMBO J 1999;18:3776–3782.
74. Szathmáry E. The origin of the genetic code—amino acids as cofactors in an RNA world. Trends Genet 1999;15:223–229.

75. Draper DE. Themes in RNA-protein recognition. J Mol Biol 1999;293: 255–270.
76. Koch AL. Evolution vs the number of gene copies per primitive cell. J Mol Evol 1984;20:71–76.
77. Mohr G, Caprara MG, Guo Q, Lambowitz AM. A tyrosyl-tRNA synthetase can function similarly to an RNA structure in the Tetrahymena ribozyme. Nature 1994;370:147–150.
78. Grishin N. Fold change in evolution of protein structures. J Struct Biol 2001;134:167–185.
79. Kinch LN, Grishin N. Evolution of protein structures and functions. Curr Opin Struct Biol 2002;12:400–408.
80. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.