BJ www.biochemj.org

BJ Structure

Biochem. J. (2009) **419**, 15–28 (Printed in Great Britain) doi:10.1042/BJ20090122

REVIEW ARTICLE

Genomic and structural aspects of protein evolution

Cyrus CHOTHIA* and Julian GOUGH†

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, U.K., and †Computer Science Department, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, U.K.

It has been known for more than 35 years that, during evolution, new proteins are formed by gene duplications, sequence and structural divergence and, in many cases, gene combinations. The genome projects have produced complete, or almost complete, descriptions of the protein repertoires of over 600 distinct organisms. Analyses of these data have dramatically increased our understanding of the formation of new proteins. At the present time, we can accurately trace the evolutionary relationships of about half the proteins found in most genomes, and it is these proteins that we discuss in the present review. Usually, the units of evolution are protein domains that are duplicated, diverge and form combinations. Small proteins contain one domain, and large proteins contain combinations of two or more domains. Domains descended from a common ancestor are clustered into superfamilies. In most genomes, the net growth of superfamily members means that more than 90 % of domains are duplicates. In a section on domain duplications, we discuss the number of currently known superfamilies, their size and distribution, and superfamily expansions related to biological complexity and to specific lineages. In a section on divergence, we describe how sequences and structures diverge, the changes in stability produced by acceptable mutations, and the nature of functional divergence and selection. In a section on domain combinations, we discuss their general nature, the sequential order of domains, how combinations modify function, and the extraordinary variety of the domain combinations found in different genomes. We conclude with a brief note on other forms of protein evolution and speculations of the origins of the duplication, divergence and combination processes.

Key words: biological complexity, domain duplication, domain superfamily, evolution, genome, protein structure.

INTRODUCTION

The proteins encoded in a genome, and how these genes are expressed, determine the material basis of an organism's anatomy and physiology. During the course of evolution, life forms of increasing complexity have arisen, and these increases have involved the invention of new proteins. The dominant mechanisms that produce these new proteins are: (i) the duplications of the genes of old proteins; (ii) the divergence of these sequences to produce modified structures whose usefulness leads to their selection, and, in many cases, (iii) their combination with other genes to modify further their properties.

Gene duplication and divergence was first clearly demonstrated by the sequences and structures of myoglobin and haemoglobin [1], and subsequently the various dehydrogenase structures clearly demonstrated gene combination [2]. The presence in proteins of adjacent regions that have very similar structures led to the discovery of tandem gene duplications [3,4]. Later work showed how intronic recombination could facilitate the formation of large complex proteins [5].

Proteins are made of domains. A domain, as the term is used here, is both a structural unit and an evolutionary unit. In the large majority of cases, gene duplications and combinations involve DNA sequences that code for one or more whole domain(s) [6]. Small proteins contain just one domain with a particular function. Combinations of two or more domains form larger proteins with more sophisticated functions. Domains typically have 50–200 residues, although smaller and larger domains do occur.

Domains can be clustered into families or superfamilies (see below) whose members are descended from a common ancestor. The ability to detect the evolutionary relationships of domains by sequence similarity is limited because they frequently diverge beyond the point where true relationships can be recognized by this means. In the absence of other data, the failure to find a significant match between two protein sequences is an 'agnostic' result. The two proteins may not be related or they may be related, but have sequences that have diverged beyond the point where they can be recognized by sequence comparison tools. If the three-dimensional structure of a protein is known, its domain structure and the evolutionary relationships of the domains can usually be recognized [7].

This means we can define two levels of homologous relationships for domains. Domains whose sequences have significant similarities form domain families. Those whose common evolutionary origin are indicated by their structure, function and aspects of sequence form domain superfamilies.

In the SCOP (Structural Classification Of Proteins) database, the related domains that occur in the proteins of known structure are clustered into families and superfamilies [6]. In the SUPERFAMILY database, HMMs (hidden Markov models) constructed for members of each SCOP superfamily are matched with all sequences predicted to be present in the sequenced genomes [8]. Good matches are made to domains in between half and two-thirds of the sequences in animal, plant, fungal and bacterial genomes. In some multidomain proteins, all domains are matched, and, in other cases, only some of the domains are matched. It is

Abbreviations used: AFGP, antifreeze glycoprotein; CspA, cold-shock protein A; HMM, hidden Markov model; SCOP, Structural Classification Of Proteins; SMBD, small-molecule-binding domain.

[1] Correspondence may be addressed to either author (email chc1@mrc-lmb.cam.ac.uk or gough@cs.bris.ac.uk).

Table 1 The size of the nine largest domain superfamilies in humans, the number of sequences in which they occur and the same data for these superfamilies in *Fugu*, *Drosophila* and *C. elegans*

In *Drosophila*, the superfamilies in rank positions 5, 6 and 9 are trypsin-like serine proteases (267 domains; 259 sequences), invertebrate chitin-binding proteins (260 domains; 87 sequences) and glucocorticoid receptor-like DNA-binding domains (206 domains; 145 sequences) respectively. In *C. elegans*, the superfamilies at rank positions 5, 6, 7, 8 and 9 are c-type lectins (356 domains; 280 sequences), glucocorticoid receptor-like DNA-binding domains (348 domains; 319 sequences), nuclear receptor ligand-binding domains (280 domains; 280 sequences); L domains (222 domains; 129 sequences) and MF (major facilitator) general transporters (216 domains; 214 sequences) respectively. The data in this Table are taken from the SUPERFAMILY database release 1.73. The total number of genes for each species are 2300 (human), 18500 (*Fugu*), 14100 (*Drosophila*) and 20100 (*C. elegans*). Abbreviations: EGF, epidermal growth factor; PH, pleckstrin homology.

	Human			*Fugu*			*Drosophila*			*C. elegans*		
Domain superfamily	Domains	Sequences	Rank	Domains	Sequences	Rank	Domains	Sequences	Rank	Domains	Sequences	Rank
C2H2 and C2HC zinc fingers	3693	742	1	949	319	4	519	223	2	134	91	23
Immunoglobulins	1778	796	2	1379	494	1	490	127	3	360	70	4
P-loop nucleoside triphosphate hydrolases	1024	861	3	1112	958	2	618	499	1	621	487	2
G-protein-coupled receptors: family A	824	824	4	470	470	9	84	84	38	970	970	1
Fibronectin type III	802	189	5	1042	225	3	199	57	10	173	40	12
EGF/laminin	697	183	6	695	182	6	209	48	8	145	57	20
Cadherins	686	100	7	892	113	5	211	19	7	128	16	24
Protein kinases	539	526	8	666	651	7	295	290	4	486	478	3
PH domains	491	410	9	587	501	8	152	135	18	124	117	27

the nature and properties of the domains matched by the HMMs that we discuss throughout much of the present review.

Analyses of the sequences predicted from the first genome projects showed that the process of forming new proteins through duplication, divergence and combination is extensive and pervasive [9–11]. The current genome projects have produced complete, or almost complete, descriptions of the protein repertoires in more than 600 distinct organisms. The analyses of their sequences has extended and quantified our understanding of how new proteins evolve, and this is the main subject of the present review. In some of the cases where the discoveries were reported using preliminary or limited data, we have recalculated the results using more recent data.

In the next three major sections we discuss, in turn, duplications of domains, their divergence, and the combinations that they form. In the last section of the review, we briefly discuss alternative processes of protein evolution and speculate on the origins of evolution by gene duplications, divergence and combination.

DUPLICATIONS

The number of known superfamilies in genomes

The current SUPERFAMILY HMMs match, at least in part, the domains in around two-thirds of the protein sequences in animals, around a half of those in fungi and in plants, and a half to three-quarters of those in bacteria [8]. The domains matched in animals come from between 800 and 1000 superfamilies, those in fungi from 650–800 superfamilies, those in plants from 800–900 superfamilies, and those in bacteria from 250–700 superfamilies.

In the human genome, the current SUPERFAMILY HMMs match all or part of 14000 of the 23000 gene loci. Matches are made to 30065 domains, and these belong to one of 1020 different superfamilies. This means that the proportion of domains that are duplicate members of a superfamily are $(30065 - 1020)/30065 = 97\,\%$ (see also [11]). Overall, in animal genomes, the proportion of the matched domains that are duplicates is 93–97 %, in fungi it is 85–90 %, and in bacteria it is 50–90 % [8].

The size and distribution of superfamilies in genomes

The number of domains in different superfamilies varies greatly. In Table 1 we list, for the matched part of the human genome, the nine largest superfamilies. These superfamilies have between 491 and 3693 members. Altogether, their domains comprise nearly 20 % of all those matched by the HMMs. At the lower end of the frequencies, there are some 220 superfamilies that have only one member; together, they form less than 1 % of all of the matched domains.

The frequency distribution of the sizes of domain superfamilies was examined in detail by two groups [12,13]. They both found that the sizes of families compared with their frequencies have power-law distributions, i.e. many families with no or few duplicates and a few families with many duplicates. To explain the distribution, one group [12] proposed a stochastic birth, death and innovation model, and the other group [13] proposed a mechanism based on the preferential attachment principle.

Before this work on domains, the distribution of the sizes of families composed of whole proteins, rather than of their constituent domains, had been examined [14]. As in the case of domain superfamilies, a power law describes the size distribution of these families. This group's analysis of the distributions implies that whole protein groupings behave in a coherent fashion within the genome in that the probabilities of duplications within one family are not independent of other families [14]. This view is strongly supported by subsequent work [15,16]. This showed that, for genomes of increasing size, the number of genes in different functional categories increases at different rates. The rate of increase is described by a power-law equation of the type $y = x^a$ and it was found, for example, that, for proteins involved in regulation of transcription the exponent (a) is 1.9, whereas for those involved in protein biosynthesis, it is 0.13 [15]. Increases in the total number of genes require increases in the number of proteins involved in regulation, whereas the overall number of genes has little effect on the number of proteins involved in protein biosynthesis.

Similarities and differences in superfamilies between genomes

An examination of the superfamilies found in animal, fungal and plant genomes showed that 95 % of domains belong to superfamilies common to eukaryotes or to the kingdom to which the organism belongs [17]. This calculation was originally carried out using the domain matches made to sequences from 38 eukaryote genomes. Since that time, many more genomes have had their sequences determined. We have therefore repeated the calculation using the sequences now available from 148 different eukaryote genomes: those from 65 animals, 17 plants and 66

fungi. [For subsequent calculations that include eukaryotes and prokaryotes, we make comparisons across 677 genomes: 181 eukaryotes (including protists), 447 bacteria and 49 archaea.]

In the previous work [17], for a superfamily to be counted as present in a kingdom (or in eukaryotes), it had to be present in all the genomes. However, the absence of a particular superfamily member from a genome could be for biological reasons, e.g. it had been lost during evolution or was only created in a more recent branch of the evolutionary tree. It could also be absent for technical reasons: it is present, but had diverged beyond the reach of the HMMs or it had been missed in the annotation of the genome. A more realistic calculation is to determine the number of superfamilies common to different proportions of sets of genomes.

Superfamilies in eukaryotes

First, we looked across the animal, fungi and plant kingdoms to determine the extent to which they have common superfamilies and the contribution that these superfamilies make to the domain repertoires in the three kingdoms. Figure 1(A) gives this information: on the *x*-axis, we give the minimum proportion of genomes required in all of the three kingdoms for a superfamily to be considered common; at the top of the Figure, we give the number of superfamilies deemed common by the criterion on the *x*-axis, and on the *y*-axis, we give the percentage of the known domains that belong to the common superfamilies. Examination of Figure 1(A) shows that there are 707 superfamilies common to at least 0.2 of the genomes in the three kingdoms and that their members form 96 % of the domains in fungi; 94 % of those in plants and 75 % of those in animals. If we look at 0.8 of the genomes in the three kingdoms, we find 535 common superfamilies and their domains form 91 % of those in fungi and in plants and 70 % of those in animals. A significant part of the fall in number of common superfamilies that occurs when all genomes are included (extreme right of *x*-axis of Figure 1A) are likely to be the result of errors in sequencing or the inability of the HMMs to match very divergent sequences.

Examination of the superfamilies common to individual kingdoms shows that the common core is shared by most genomes in that kingdom. There are 795 superfamilies that occur in at least 0.8 of the animal genomes, and their domains form on average 98 % of all domains in animals. In fungi, there are 649 such superfamilies that form 97 % of all domains, and in plants, there are 681 such superfamilies that form 96 % of all domains.

From the data described in the two preceding paragraphs, we conclude that, within each eukaryote kingdom, there is a common core of superfamilies that accounts for almost all of the domains, but, when looking across eukaryote kingdoms, we see that, in animals, this core is expanded with respect to plants and fungi.

Superfamilies common between eukaryotes, bacteria and archaea

The number of superfamilies common to 0.8 of the genomes in each of these three sets is 213 and they form 34, 61 and 70 % of the domains respectively (Figure 1B). We also observe that as many as 15 % of the domains in eukaryotes belong to 294 superfamilies not yet observed in bacteria or archaea, whereas less than 0.5 % of the domains in bacteria or archaea (from 124 and 18 superfamilies respectively) are yet to be observed in one of the other two.

Superfamilies within bacteria and archaea

The superfamilies common to 0.8 of bacterial genomes form 80 % of the domains in those genomes. The superfamilies common to 0.8 of archaeal genomes form 79 % of their domains. However, the number of common superfamilies that contain these domains (387 and 310 respectively) is significantly smaller than the equivalent number in 0.8 of eukaryote genomes (586). The total number of superfamilies found at present in archaea is only 770. In bacteria, however, the combined repertoire is of 1107 superfamilies, which is very similar in size to that of the eukaryotes, 1282, so it is the number in the individual bacterium, not the shared repertoire of bacteria, that is significantly smaller.

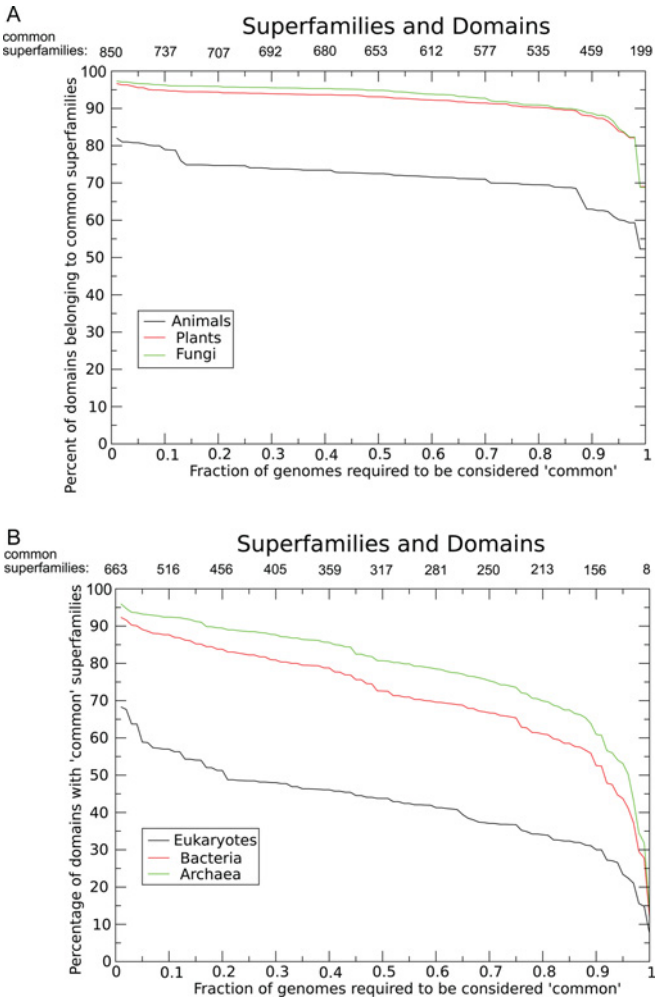


Figure 1 Superfamilies common to eukaryotes, prokaryotes and kingdoms

The graphs show what proportion of domains in a given set belong to superfamilies which are common to all of the sets. The sets are (**A**) animals, plants and fungi, and (**B**) eukaryotes, bacteria and archaea. The *x*-axis denotes the fraction of genomes in each set which are required to contain the superfamily in order to be considered 'common' for the purpose of calculating the values shown on the *x*-axis. Written across the top of the graphs are the numbers of superfamilies considered common under the condition specified by each tickmark on the *x*-axis. Data taken from the SUPERFAMILY database [8].

Superfamily expansions related to biological complexity and specific lineages

Before the genome projects were undertaken, there was a general assumption that the number of genes in an organism would be at least roughly proportional to its biological complexity. When the genome projects found that the number of genes in humans, *Drosophila* (an insect), *Caenorhabditis elegans* (a nematode

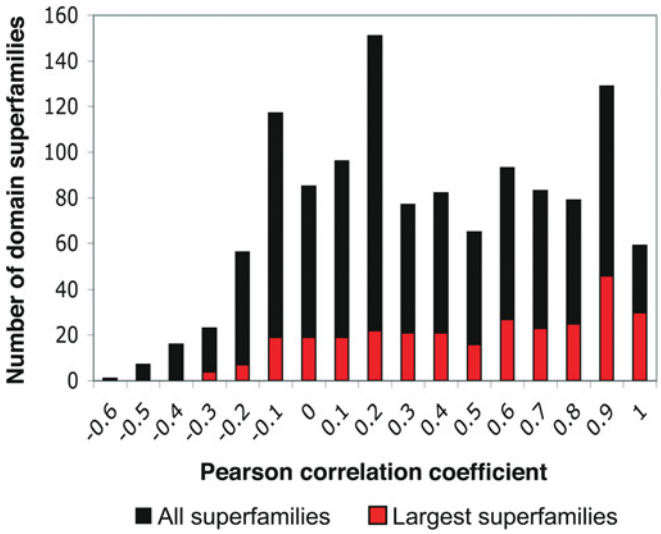


Figure 2 Correlation coefficients for the size of superfamilies against biological complexity

Correlation coefficients were calculated for match between the size of 1219 domain superfamilies in 38 eukaryotes and the number of different cell types found in the eukaryotes. Only 194 superfamilies have a correlation coefficient >0.8. Large superfamilies, those with at least 25 members in at least one genome, are shown in red. Reproduced from [18] with permission.

worm) and *Arabidopsis* (a plant) are close to 23000, 14000, 20000 and 27000 respectively, it was clear that this assumption is not correct.

Superfamilies and biological complexity

To discover the superfamilies that are related to increases in biological complexity, calculations were carried out to determine which superfamilies have increases in size that correlate with increases in complexity [18]. The measure of the complexity of the 38 eukaryotes (17 animals, ten fungi, three plants and eight protists) was taken to be the number of cell types of which they are composed: between three (some fungi) and 170 (humans).

Altogether, the 38 organisms have domains from 1219 different superfamilies. Of this total, there are close to 200 superfamilies whose sizes in the 38 organisms are strongly correlated with the number of cell types in the organisms: they have a Pearson correlation coefficient of 0.8 or greater. More than half of the superfamilies have no significant correlation (Figure 2).

Although the members of large superfamilies can have different specific functions, the types of processes in which they are mostly involved are often similar. The superfamilies were assigned to the process in which most members, if not all, participate. Of the 55 superfamilies involved in extracellular processes, there are 39 whose membership is strongly correlated with the number of cell types (Figure 3). Of the 163 superfamilies involved in regulation, there are 57 such superfamilies. On the other hand, only 26 of the 448 superfamilies involved in metabolism have a high correlation.

These calculations show that only a relatively small number of domain superfamilies have memberships that correlate with increases in biological complexity and are likely to have played the major role in its creation. However, this still leaves unexplained why there is no correlation with gene number. Inspection of superfamily sizes in genomes reveals that there are not only expansions that correlate with complexity, but also expansions that are specific to particular lineages.

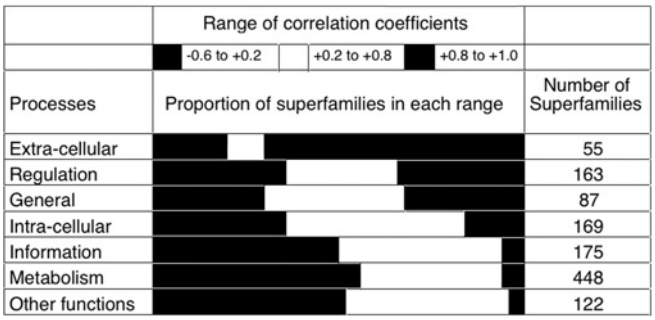


Figure 3 Contributions of different processes to complexity

For proteins involved in different cellular processes, we give the proportion of superfamilies whose correlations with biological complexity (number of cell types) in 38 eukaryotes are in the ranges −0.6–+0.2, +0.2–+0.8 and +0.8–+1.0. The number of superfamilies assigned to each process is given on the right of the Figure. Data taken from [18].

Lineage-specific expansions

Detailed investigations of the expansions of lineage-specific protein families were described in two papers [19,20]. There is also a list of proteins that could be unique to humans, fruitflies, nematode worms, yeast and bacteria [11]. They defined lineage-specific expansions as proliferations of protein families in a particular lineage, relative to the sister lineage(s) with which it is compared. In eukaryotes, the functional categories most prone to these sorts of expansion are structural proteins, enzymes that respond to pathogens and stress, components of signalling pathways and transcription factors [20].

Variations in the size of superfamiles and genomes

We have described above how almost 95% of domains come from superfamilies common to eukaryotes or to the kingdom to which the organism belongs. However, the variable expansions that occur in lineage-specific expansions, and the systematic changes that occur in superfamilies whose sizes correlate with complexity, mean that the size of superfamilies in distantly related organisms can be very different. In Table 1, we list the nine largest domain superfamilies known in humans. All but one of these is a superfamily whose membership has a high correlation with biological complexity. Table 1 also lists the rank order, number of domains, the number of sequences found for the same domain superfamilies in *Fugu*, *Drosophila* and *C. elegans*. The nine largest domain superfamilies in *Fugu* are the same as those in humans, although their rank order differs somewhat (Table 1). In *Drosophila*, five of the human superfamilies are found among the top eight positions. The superfamily at position 5 is that of the trypsin-like proteases and that at position 6 is that of the invertebrate chitin-binding proteins. The high rank of these families is the result of lineage-specific expansions. In *C. elegans*, only four of the nine human families have high positions, and superfamilies subject to lineage specific expansions occur at positions 5, 7, 8 and 9 (see Table 1).

Genomes are composed of different sets of genes: those that belong to superfamilies whose size correlates with the organism's complexity, those that belong to superfamilies that have lineage-specific expansions and others that belong to neither of these groups. It is the combination and variations of these different sets of genes that give total gene numbers that are not simply related to complexity. It has been suggested that expansions that correlate with complexity can be called 'progressive' in that they lead to new physiological features in an organism, whereas those that are lineage-specific can be called 'conservative' in that they allow an

organism to adapt better to its environment, but do not change its physiology [21].

DIVERGENCE

Sequence divergence

The nature of the selection processes that decides whether or not a mutation is acceptable has been a subject of argument and controversy. The early views argued that it is the effects of a mutation on function and stability that are the major determinants [22]. Subsequently, the role of expression was shown to be significant [23], and some now argue that this is the single major determinant [24]. It has been shown that, on average, the sequences of orthologous proteins that form stable complexes diverge somewhat more slowly than those that form transient complexes and that these in turn diverge somewhat more slowly than proteins not known to be involved in interactions [25]. An earlier review, with the somewhat ironic title 'An integrated view of protein evolution' [26], discusses the roles proposed for these factors and of several others.

In this section, we describe the nature of the divergence process that is actually observed in proteins that maintain the same or very similar functions and structures.

To determine the general nature of the mutations that can be accepted by orthologous proteins, three large sets of sequences were examined [27]. The three sets are the orthologous proteins found in humans and mice (h_m), which diverged 90 million years ago, humans and chickens (h_c), which diverged 310 million years ago, and *Escherichia coli* and *Salmonella enterica* (e_s), which diverged 100 million years ago. The total number of orthologous pairs in the three sets is 21738 and they contain nearly 2 million mutations.

The first two sets involve eukaryotes whose time of divergence differs by a factor of 3.5. The third dataset involves prokaryotes. Although biologically very different, the three datasets gave very similar results.

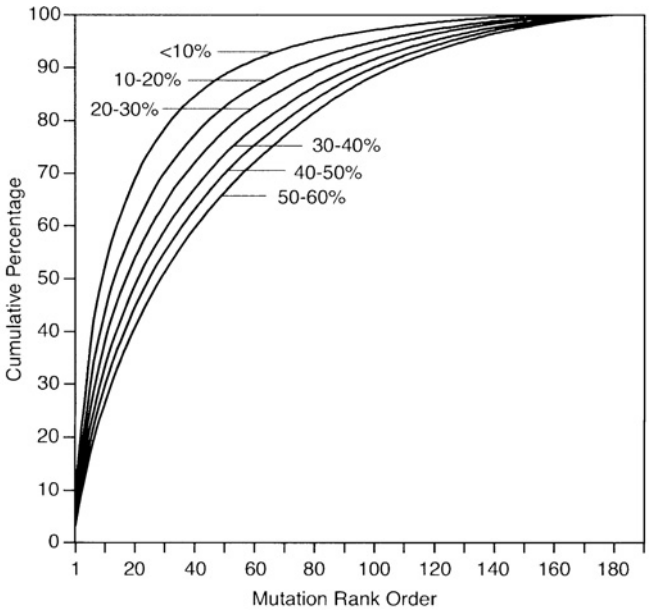
The direction of protein mutations was not taken into account, so this gives $[20 \times (20-1)]/2 = 190$ possible types of mutations that were considered. All, or almost all, of the 190 are seen to occur, but with frequencies that have an exponential density distribution.

For sequences that have diverged by a similar amount, the exponential distributions of the mutation frequencies are very similar in all three sets of orthologues. For example, in the h_m, h_c and e_s sets with up to 10 % divergence, we find that 75 % of all mutations are formed by the most frequent 28 or 29 types of mutations (Figure 4). As divergence increases, the exponent of the distribution becomes smaller. Thus, for sequences that have diverged by 50–60 %, the most frequent 64–67 mutations form 75 % of all mutations (Figure 4).

The rank order of the common mutations in the h_m, h_c and e_s categories was examined. For categories with low divergence, the rank orders of the most common mutations in the three datasets are very similar, whereas at high divergence, they are roughly similar (Figure 5).

In a few cases, there are large differences in rank order of particular types of mutations. For example, the serine ↔ proline mutation has a high rank in h_m and h_c categories and a much lower rank in the e_s categories: see Figure 5. This can be explained, at least in part, by the frequencies of the codons for serine and proline in humans, mice and chickens being different from those in *E. coli* and *Salmonella* [27].

The very similar results obtained from the three sets of orthologues imply that there is a common selection process for



The number of frequent mutation types that form 75% of all mutations in sets of orthologs whose divergences differ

Divergence of ortholog datasets:	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%
human-mouse	29	37	44	52	59	65
human-chicken	28	38	46	54	61	67
E.coli-S.enterica	29	37	45	50	57	64

Figure 4 The number of frequent mutation types that form 75 % of all mutations in sets of orthologues whose divergences differ

Cumulative contributions made by each type of mutation when placed in descending rank order of their frequencies. For each range of divergence: < 10 %, 10–20 %, 20–30 % etc., the mutations were placed in descending rank order and summed. Their cumulative contributions were calculated as a percentage of all mutations. For the three sets of orthologues, the table lists the number of mutation types that form 75 % of all mutations in each divergence category. Reproduced from [27] with permission.

sequence divergence. The selection process accepts mutations whose composition and distribution (see below) is a characteristic of the extent of the divergence. The average composition is not affected by the time taken to diverge. For example, the frequencies and rank order of the h_m and h_c mutation in the >10 % divergence category are very similar, although the time taken to produce these mutations is 90 million years for the h_m set and 310 million years for the h_c set.

Changes in stability produced by acceptable mutations

Many protein-engineering experiments have been carried out to measure the effects of engineered mutations on protein stability and function: the current version of ProTherm, a database that lists the thermodynamic effect of mutants, has data on more than 23000 such mutations [28]. What are of interest for the present review, however, are the changes in stability that are produced by the natural mutations that have survived selection: i.e. those that are seen in native homologous proteins.

Barnase and binase are members of the microbial RNase family. Their sequences have 110 and 109 residues respectively, and, at 17 sites, the identity of the residues differ. Each of the 17 sites in barnase was mutated, independently, to the residue that occurs at the equivalent sites in binase [29]. The 17 individual changes altered the stability of barnase by between −1.0

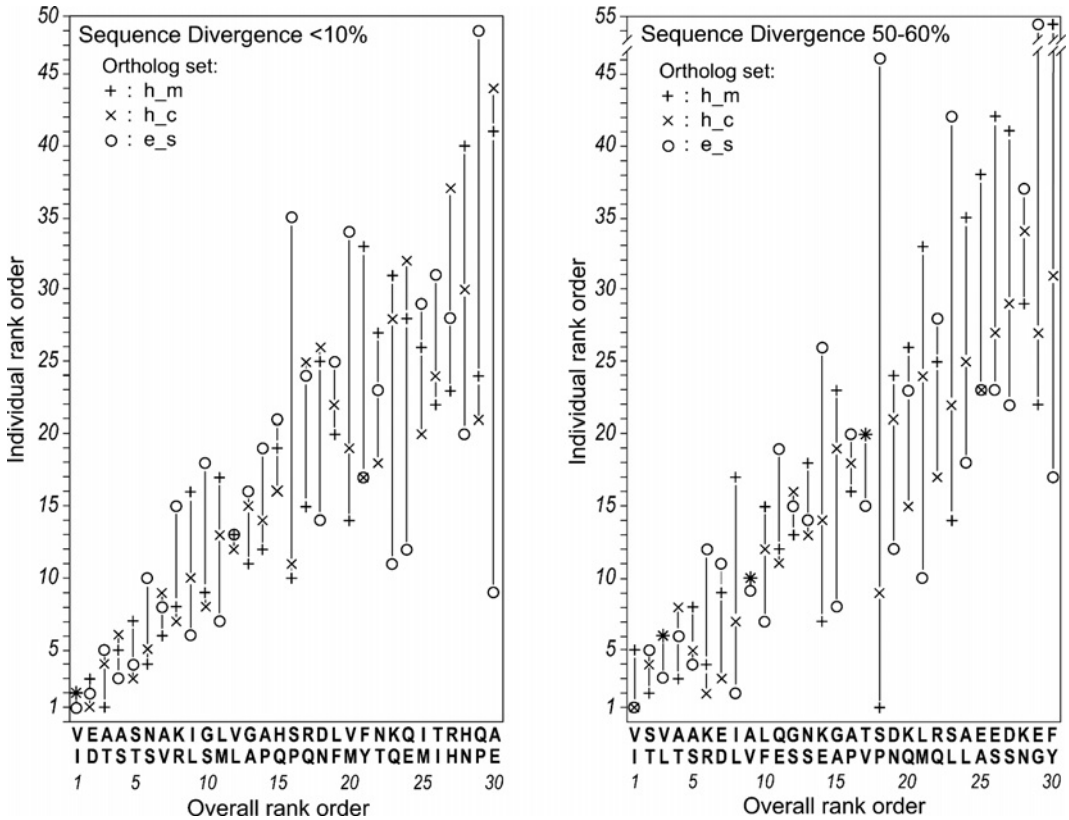


Figure 5 Rank order of the frequent mutations in sequences that have diverged by up to 10 % and between 50 and 60%

Overall rank position of the 30 most frequent mutation types in the < 10 % and 50–60 % divergence categories. The overall rank position was determined by summing the individual rank positions (see *y*-axis) in the three datasets. Note that that there are differences in the rank order of mutations in the < 10 % and 50–60 % divergence categories. h_m, humans and mice; h_c, humans and chickens; e_s, *Escherichia coli* and *Salmonella enterica*. Reproduced from [27] with permission. ©2007 the National Academy of Sciences.

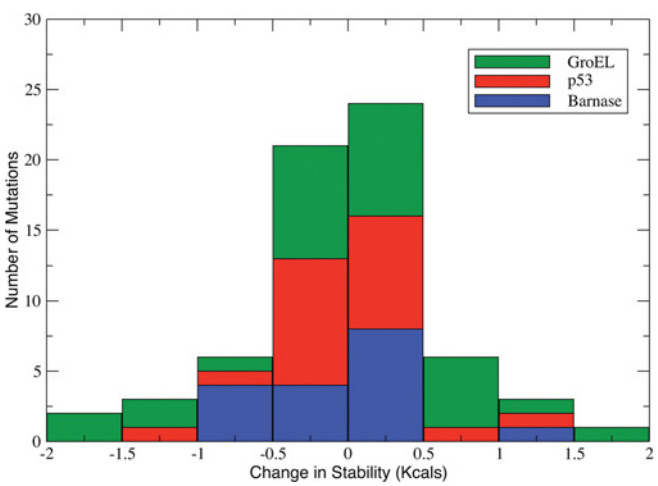


Figure 6 Changes in protein stability produced by observed mutations

Changes in protein stability produced by engineered mutations that mimic native mutations seen in barnase [29], p53 [31] and GroEL [32].

and $+1.2\ \text{kcal}\cdot\text{mol}^{-1}$ (1 kcal = 4.184 kJ); 12 changes altered stability by -0.5–$+0.5\ \text{kcal}\cdot\text{mol}^{-1}$. Seven of the mutations increase stability, three have very small effects and seven reduce stability (Figure 6).

It was shown that if, in a family of related proteins, a residue at a site was different from that of the consensus for that site, its mutation back to the consensus residue usually improved stability [30]. Using this approach, mutations were made to p53 [31] and to GroEL minichaperone [32]. These experiments gave results that were similar to each other and to those given by the barnase–binase mutations (Figure 6). For p53, 21 residues were mutated and, for GroEL, 32 residues. The overall range of the changes in stability is a little larger: -1.5–$+1.3\ \text{kcal}\cdot\text{mol}^{-1}$ and -1.8–$+1.6\ \text{kcal}\cdot\text{mol}^{-1}$ respectively, but, as in the case of the barnase–binase mutations, most mutations change stability by -0.5–$+0.5\ \text{kcal}\cdot\text{mol}^{-1}$ (Figure 6).

The consistency of the results produced by these three experiments suggests that divergence in proteins largely occurs through a rough alternation of mutations which produce small positive or negative changes in stability: if several of the mutations that reduce stability in barnase had occurred in succession, it would not be stable.

Another striking result of all three of these experiments is that, whereas the folding of a protein is highly co-operative, no co-operativity is found between mutated sites: the change in stability produced by combinations of mutations is very close to the sum of the individual mutations. Thus compensation for changes in stability does not occur through direct interactions between mutating residues, but through the net contribution they make to stability.

Acceptable radical mutations

Examination of the first pair of homologous proteins to have their structure determined clearly showed that sites buried within

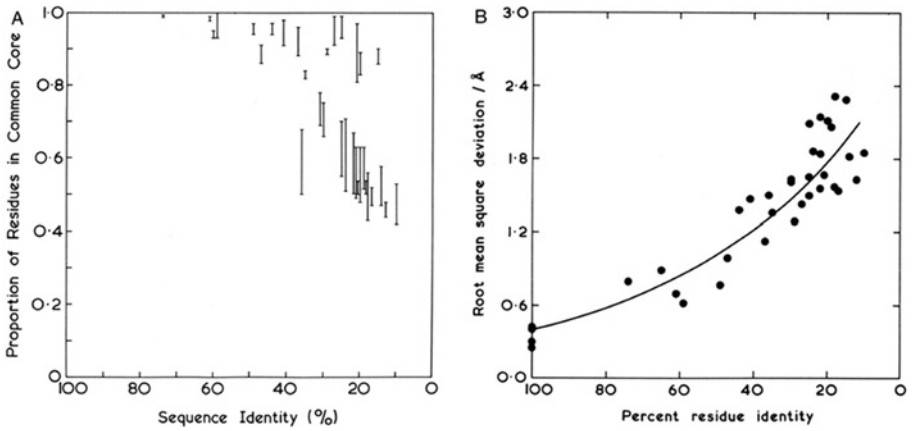


Figure 7 Divergence of structure and sequence in homologous proteins

Left-hand panel: the proportion of residues in the core regions of 32 homologous proteins (the regions that maintain the same local conformation) plotted against the sequence identities of the residues in the core regions. If two homologous proteins with n_1 and n_2 residues have c residues with the same fold, the proportion of each sequence in the core regions are c/n_1 and c/n_2. For 32 pairs of homologues from eight protein families, we plot the values of c/n_1 and c/n_2 linked by a vertical bar. Right-hand panel: the root mean square deviation in the position of the main-chain atoms of the residues in regions that maintain their local conformation homologous pairs of proteins is plotted against their residue identities. Reproduced from [40] with permission.

Table 2 Proportion of residues that are mutated in the buried, intermediate and surface sites of proteins whose sequences have diverged by < 10 % to 50–60%

For six divergence categories, < 10 % to 50–60 %, we give the average proportion of residues in each ASA (accessible surface area) range that have mutations. Thus, for example, for orthologues whose sequences have diverged by 20–30 %, 12.4 % of the buried residues (ASA = 0–20 Å^2), 23.7 % of the intermediate residues (ASA = 20–60 Å^2) and 38.2 % of the surface residues (ASA = 60–~140 Å^2) have mutations.

ASA of the three regions (Å^2)	Proportion (%) of mutated residues in each ASA region					
	< 10 %	10–20 %	20–30 %	30–40 %	40–50 %	50–60 %
0–20	2.4	6.7	12.4	17.7	29.3	39.6
20–60	5.0	13.5	23.7	33.2	44.0	55.7
60–~140	9.8	24.3	38.2	49.8	57.5	67.8

the protein are less susceptible to mutations than those on the surface [33]. The interior of proteins is largely formed by close-packed secondary structures, and, in most cases, this places strong constraints on acceptance of radical mutations [34].

Radical mutations in the interior are, however, seen on rare occasions. The variable domains of antibodies have buried at the centre of their structure a disulfide. This is almost always conserved in variable domains [35]. In the V_H domain of the natural antibody ABPC48, a cysteine residue that is part of the conserved disulfide is replaced by tyrosine [36]. An examination of the stability of the V_LV_H dimer found that it is significantly less stable than the average V_LV_H dimer and that mutation back to cysteine created a V_LV_H dimer whose stability is significantly greater than that of the average V_LV_H dimer [37].

These results imply that the extent of the stability of a protein will play a role in determining whether or not a mutation is acceptable. If stability is very high and function is not affected, a mutation making a large reduction in stability may well be tolerated. If stability is low, a mutation that produces only a small reduction in stability may not be acceptable. Thus, in ABPC48, the natural mutation of cysteine to tyrosine at site 92 is made tolerable by the fortuitous evolution of a variable domain whose structure has an exceptionally high stability. Such radical mutations are likely to do permanent damage to the proteins in which they occur [38].

Structural divergence

The general effects of mutations on the structure of proteins were described some time ago [39–41]. There are two kinds of structural changes: those that occur in peripheral regions (surface loops, small surface helices and strands on the edges of β-sheets) and those that occur in the core secondary structures [40]. The peripheral regions of the structure have fewer restraints than the regions in the buried core, and this allows mutations, insertions and deletions that can change the local conformation. In sequences that differ in identity by up to 50 % of their residues, the extent of the local changes are small: usually 10 % or less of the residues are affected by conformation changes. With divergence above 50 %, the extent of the local changes rapidly increases, and it is common for pairs of homologous proteins with identities of ~20 % to have local changes in the conformation of peripheral regions that, taken together, comprise half the structure (Figure 7).

The mutations of the residues that make contacts between secondary structures rarely change local conformation. Their usual effect is to change the relative positions of the secondary structures. In distantly related proteins, mutations change the relative position of close packed helices by shifts of several angstroms and rotations of up to 30° [39]. Data from the pairs of homologous proteins showed that, to a good approximation, Δ, the root mean square difference in the position of their main chain atoms, is related to the extent of the sequence differences by the equation:

$$\Delta = 0.4e^{1.87H}$$

where Δ is measured in Å (1 Å = 0.1 nm) and H is the fraction of the residues in the common core that are not identical (see Figure 7).

The reason for the exponential relationship between sequence divergence and structural divergence has been explained only recently [27]. A detailed examination was made of the distribution of natural mutations in the structures of 995 pairs of orthologous

proteins. For each of the 995 structures, the total number of buried, intermediate and exposed residues was counted, as was also the number of mutations. For the structures in different divergence categories, >10 %, 10–20 %, etc., the proportion of residues that are mutated in the buried, intermediate and exposed regions was determined (Table 2).

Examination of Table 2 shows that mutations occur at different rates in different regions of the structures. On going from an overall divergence of <10 % to 50–60 % the mutations in the buried region increase from 2.4 % to 39.6 %, i.e. by a factor of ~16.5, those in the intermediate region increase from 5 % to 55.7 %, i.e. by a factor of ~11.1, and in the exposed regions, they increase from 9.8 % to 67.8 %, i.e. by a factor of ~7.

This behaviour arises as a result of two factors. The mutations in the exposed regions are some four times more acceptable than mutations in the buried regions (see the <10 % category in Table 2). This means, of course, that mutations accumulate more rapidly in the exposed region. But, as the divergence increases, new surface mutations will increasingly occur in residues that have already been subject to mutation. In the buried region, which has more residues but a smaller proportion with mutations, this will occur more slowly and a higher proportion of mutations will occur at sites that have not had mutations. The large relative increase in the proportion of mutations in the buried region is the reason for the exponential relation between structural and sequential divergence described previously.

Radical divergence of structures

The previous section has described the general process of structural divergence. For some related proteins, there are cases where more radical changes in structure occur.

Permutations of protein sequences

There are proteins for which there is good evidence of homology, but which have segments of their sequences in different sequential order(s). The Circular Permutation Database [42] lists some 120 protein clusters in which this is seen. The mechanisms that could have produced the change in the order of protein segments have been reviewed in [43].

In the case of concanavalin A and flavin, two proteins that are homologous, but which have circularly permuted sequences, it has been shown to arise from post-translational formation of a new peptide bond between the original N- and C-termini of concanavalin and cleavage at another site to give new N- and C-termini [44]. A second mechanism is a whole-gene duplication that produces a fused tandem repeat of the gene that is subsequently truncated at both termini to remove redundant segments. Recent experiments have given strong support to this mechanism being the origin of the different sequential order of segments in the DNA methyltransferase families [45].

Related proteins whose structures have parts that have different conformations

For at least some proteins that diverged over long periods of time and whose functions do not place strong constraints on their evolution, it is likely that evidence of their common origin has been lost. However, three papers [7,46,47] have discussed criteria for detecting homology in proteins that have little sequence similarity and whose structures have significant differences. They show that close inspection of groups of proteins in which only a small region has the same substructure, whose functional residues are quite different, and/or whose chains form different topological isomers, can, at least in some cases, produce clear evidence that indicates their homology [7,46].

The formation and evolution of oligomers can facilitate domain swapping, duplication, deletion of redundant active sites and decoration with additional structures. Again, taken together, these processes can produce structures that are very different. But again, if sufficient intermediate structures are known, their evolution can be traced, as has been shown for the PGDH (prostaglandin dehydrogenase)-like oxidoreductases, which have large regions whose conformations are different [47].

Divergence of function

Enzymes

An examination of metabolic proteins, specifically the 510 enzymes that form a large part of the *E. coli* metabolic pathways [48], showed that those which are homologous with each other usually conserve their catalytic and/or cofactor properties; as was proposed in 1976 [49] and more recently [50]. Twice as many homologues are distributed across different pathways than within pathways. Homologues that conserve substrate binding and change catalytic mechanism are rare: only a few are found in five of 106 pathways examined [48].

Duplication of proteins in general can, however, lead to homologues that have modified or very different functions. An initial survey showed that change in function is usually restricted to homologues whose sequence identities are less than 40 % [51]. A detailed comparison of homologous enzymes showed that, for homologues with identities of 30–40 %, the first three EC numbers are conserved in 90 % of cases [52]. For homologues with identities of less than 20 %, it is common for only the first or no EC number to be conserved. Examination of 167 enzyme superfamilies showed that 70 % have members with different functions, 43 have absolutely no conservation in EC number, and 59 superfamilies have non-enzyme members. A detailed examination of 31 of these superfamilies showed that duplicates, which, in some cases, are combined with other domains, carry out close to 400 different functions [52].

Although homologues usually conserve catalytic mechanisms, the position of their catalytic residues can differ. An examination of 31 enzyme superfamilies showed that this is found in 12 of them. It occurs when divergence has produced structural changes in the catalytic region and the position of the catalytic residues is reconfigured to maintain function [52].

Examination of 27 pairs of homologous enzymes that have totally different functions showed that, in spite of their very different functions, they do tend to retain certain common features, namely the position of the active site and residues that bind catalytic metal ions and cofactors and/or particular steps of their reactions [53].

Enzymes and non-enzyme homologues

As mentioned above, some superfamilies have members that are enzymes and non-enzymes. The sequence identities of enzyme and non-enzyme homologues are usually less that 20 % [54]. Phylogenetic data indicate that, in two-thirds of the superfamilies, non-enzymes had evolved from enzymes, and, in one-third, that their evolution had been in the opposite direction. In half of the superfamilies, the enzymes and non-enzymes have similar binding properties; the other half have no similarity in their functions [54]. Examination of the members of 47 enzyme families in seven genomes that range between *C. elegans* and humans showed that 25 have three or more sequences that have lost catalytic residues and 13 have no such sequences. Functions formed by the non-catalytic homologues include regulation of other enzymes and modulation of signalling pathways [55].

Proteins that carry out eukaryote-specific functions

Eukaryotes carry out many complex processes that are not found in prokaryotes. These included chromosome organization and dynamics, RNA processing, vesicular transport, signalling systems and apoptosis. Two papers [56,57] examined these and other processes to determine, as far as is possible at present, the extent to which the proteins involved in these processes (i) have prokaryote precursors with related functions, (ii) have prokaryote precursors with different functions, or (iii) are eukaryote inventions. They found that about half of the proteins are in the first category and a quarter are in each of the other two categories. Although, as they mention in a footnote, new discoveries of prokaryote homologues are reducing the proportion in the third category.

Divergence and selection

Protein domains within a superfamily may diverge from each other to the point where they can be grouped into subfamilies. New families can come about in two possible ways: either there is speciation and the domain evolves differently in the two organisms (orthologous) or the domain is duplicated within one organism and the duplicates subsequently diverge from each other (paralogous). Work comparing events across all genomes and superfamilies [58] showed that 80 % of the events leading to a new family are paralagous and 20 % are orthologous; once internally duplicated within a genome, a domain is under less selective pressure to remain the same. The distribution of orthologous events fits a multinomial distribution, which is what would be expected from a model of random divergence followed by opportunistic selection. Work examining the three-dimensional structures of different domains in combination with a Rossmann domain [59] showed that the relative order of domains in a sequence was not related to the orientation in three-dimensions, supporting further the independence between natural divergence and the process of selection.

DOMAIN COMBINATIONS

In parts of this section, we discuss domain architecture(s). This term is applied to multidomain proteins. Proteins have the same domain architecture if they are formed by domains from the same superfamilies arranged in the same sequential order. Differences in domain architecture can be small, e.g. domains in the same order with one terminal exception, or total, i.e. no homologous domains.

The extent of domain combinations became apparent when the results of the early genome projects became available. An examination of the domain structure of the genome sequences from the small bacterium *Mycoplasma genitalium* suggests that some two-thirds of its proteins contained two or more domains [10] and, in the genomes of more complex organisms, the proportions are higher [60,61].

A survey of two-domain proteins established certain basic features of the combination process [62]. First, the members of most superfamilies make combinations with the members of only the one or two other superfamilies, and a few superfamilies have members that make combinations with members of many other superfamilies; this pattern can be described as a scale-free network. Secondly, in the large majority of cases, domain combinations have a unique sequential order: if, for a two-domain combination, BA is found, then it is rare that AB will also be found, and vice versa. Thirdly, examining the two-domain proteins formed by superfamilies common to eukaryotes, bacteria and archaea showed that only 15 % of the combinations occur in all three of these, and 70 % are unique to one of the three. Fourthly, the number of different combinations found in the matched regions of genomes increases on going from archaea to bacteria to eukaryotes. They also showed that many more multidomain proteins are formed by combinations than by tandem duplications [62].

Subsequently, a number of papers used the more extensive genome data that became available to examine these features in more detail.

Domain proliferation

Abundance, versatility and alternative splicing

As mentioned in the section on domain duplications, a small proportion of superfamilies have many members, i.e. a high abundance in genomes. Their abundance gives them the possibility of also having high versatility, i.e. to be able to combine with domains from many different superfamilies. Examination of the domain combinations described in the SUPERFAMILY database showed that there is indeed a high correlation between domain abundance and domain versatility ($r^2 = 0.75$) [63].

Alternative splicing can affect protein structure to a greater extent than gene duplication and divergence, and they might be thought of as independent processes. However, an examination of the extent to which protein families undergo gene duplications and alternative splicing found that the two processes are inversely correlated [64,65]. The reason for this is unclear. One possibility is that the two processes are routes to different types of versatility and, given the functional nature of the family, one process is more appropriate than the other.

Supra-domains

Some combinations, mostly of two or three domains, occur with high frequency and form combinations with several or many different partner domains. These combinations have been given the name ‘supra-domains’ in a study that found them to be present in over one-third of matched multidomain proteins [66].

The individual domains in a supra-domain have functions that act co-operatively and properties that can be used in different contexts. The domains that combine with supra-domains determine its specific function. To give two examples, the two-component signal transduction proteins have different SMBDs (small-molecule-binding domains) that determine their specificity; a supra-domain formed by an FAD-binding protein and a NADPH-binding protein form an electron-transfer pathway in many enzymes.

We determined the extent of supra-domains in the much larger number of genomes that are now available (677). These genomes have a total of 62 938 different domain architectures between them in combination. We found 7648 supra-domains, defined as having more than one domain and found with more than one partner.

There are 258 supra-domains in the set which have clear evidence for having been selected for in that they are seen in as many or more contexts (different domain architecture of the whole protein) as one of their component domains. The combinations in this group are found in 6974 different domain architectures. The majority of these supra-domains, 205, contain two domains, 36 are combinations containing three domains, and 17 have four or more domains. There are 12 supra-domains containing tandem repeats and these account for most of those consisting of four or more domains.

Domain insertions/deletions, repeats and exchanges

To measure the roles of these events in the formation of multidomain proteins that share one or more common domains, the

measure 'domain distance' was used [67]. Given an alignment of related, but not identical, domain architectures, their domain distance is the number of domains that are not matched in the alignment. Using this measure on a variety of sequence datasets, it was shown that, of the events producing multidomain proteins, close to two-thirds involve insertion/deletion events and one-third involve internal repeats, and only a very small number are exchanges of domains.

The insertions/deletions referred to in the previous paragraph involved domain combinations that are sequential, i.e. have C-terminal to N-terminal links. There are, however, cases where one (or more) domain(s) are inserted into an internal region of a 'parent' domain [68]. The large majority of known cases have a single domain inserted in a parent domain, but there are cases of two or three sequentially linked domains being inserted into a parent domain and of an inserted domain itself having an inserted domain to form a nested set of three domains. Almost all proteins having inserted domains are enzymes and domains with α/β- and $\alpha+\beta$-folds are the most common constituents [68].

Domain fusion/fission

An examination of the relative rates of fusion and fission in multidomain proteins found that fusion events are approximately four times more common than fission events [69]. Fission events can be relatively benign if, for example, they involve bifunctional proteins whose functions are on separate domains. Deletions largely occur at the chain termini and they usually involve the loss of whole domains [70]. This suggests that they arise from mutations that produce new start or stop codons.

Domain repeats

A detailed examination of the tandem duplication of domains in genomes has been carried out [71]. In the genomes of bacteria, *Drosophila* and humans, domain repeats are found in 5, 11 and 17 % of their respective sequences. The sequence similarities of neighbouring domains showed that repeats can involve duplication of one or, more frequently, several domains. They often occur in the middle of the repeat region (in contrast with combination, which is usually at one of the termini).

Conservation of the sequential order of domain combinations

The observation that, in the large majority of cases, combinations occur in only one sequential order [62] was the subject of two investigations. One examined in detail the structures of sets of paralogous two-domain proteins [59]. It showed that, except in rare cases, domain order was not a functional requirement, concluding that, in the large majority of cases, the domains have the same order because the domain pairs in current proteins each came from a unique combination event. The other used data from phylogenetic groupings, sequence alignments and mutation rates to trace the history of multidomain protein architectures [72] and estimates that between 96 and 99.6 % of the architectures have only been created once in evolution. Of those few architectures which are shared by proteins from multiple origins, most are caused by varying the number of tandem repeats.

Domain combinations common to different genomes

We discussed above the remarkably little variation seen in eukaryotes in the repertoire of the core superfamilies that form most of the known proteins. Most eukaryotic proteins are formed by superfamilies common to eukaryotes in general or to their kingdoms. In contrast with this, we see a great deal of variation in the repertoire of domain architectures. Eukaryote proteins tend to have more domains than their prokaryote homologues, i.e. more complex architectures and properties, and this has been termed 'domain accretion' [73]. Examination of the domain architectures of the immunoglobulin and cadherin proteins in *C. elegans* and *Drosophila* [21,74] showed that less than half the sequences in the two organisms have common architectures.

Within animals, the human genome has no unique superfamilies and almost its entire protein content is made from a repertoire of superfamilies common to the other animals, but the human genome has 165 unique architectures (out of 5212) not seen in any other genome, and only ~60 % of proteins have architectures common to 90 % of other animals. Much of the uniqueness in humans is shared with the other five primates whose genome sequences are known. Together, their sequences have one of 8786 different domain architectures. Individually, the have between 3467 (*Galago*) and 5212 (humans) domain architectures. Of the 8786, there are 1530 not found in non-primate genomes.

To get an overview of diversity of domain architectures, we carried out calculations equivalent to those carried out for superfamilies and domains (see Figure 1 and insets in Figure 8), but using domain architectures in place of domain superfamilies (Figure 8).

Architectures common to the animal, plant and fungi kingdoms

Together, the current genomes for these three kingdoms contain 45776 different domain architectures. In Figure 8(A), plots show, for different proportions of the genomes in the three kingdoms, the number of common architectures and the percentage of sequences that have these architectures. There are 1311 architectures common to at least 0.2 of the genomes in the three kingdoms and their members form 79 % of the domains in fungi, 43 % of those in plants and 59 % of those in animals. If we look at 0.8 of the genomes in the three kingdoms, we find 434 common architectures and their domains form 59 % of the sequences in fungi, 54 % of those in plants and 40 % of those in animals.

The number of common architectures, 434 for 0.8 of the genomes in each kingdom, is a small fraction of the total number of different architectures (45 776). They do, however, cover about half of the sequences in the kingdoms. This implies that at least some of the common architectures have many paralogues.

Architectures common to eukaryotes, bacteria and archaea

The sequences in these three sets of genomes have one of 62 894 different domain architectures. If we look at Figure 8(B), we see that 0.8 of the genomes have 94 common architectures and their domains form 36 % of the sequences in fungi, 30 % of those in plants and 15 % of those in animals. Again, although on a smaller scale, this implies that at least some of the common architectures have many paralogues.

The most radical difference in diversity of superfamilies and architectures can be seen in the absolute numbers written across the top of Figures 1 and 8. There are naturally more architectures than superfamilies that the constituent domains belong to, so the numbers of core architectures are ~2–4-fold greater than the numbers of core superfamilies. However, we can imagine subtracting these numbers in Figures 1 and 8 of common superfamilies/architectures from the total available: 1254 superfamilies in Figure 1(A), 1454 superfamilies in Figure 1(B), 45 776 architectures in Figure 8(A) and 62 894 architectures in Figure 8(B). If we were to do these subtractions, we would see that there are more like 20–100 times more non-core architectures than non-core superfamilies. What this tells us is that Nature has achieved very little innovation since the last common ancestor of life by creating new

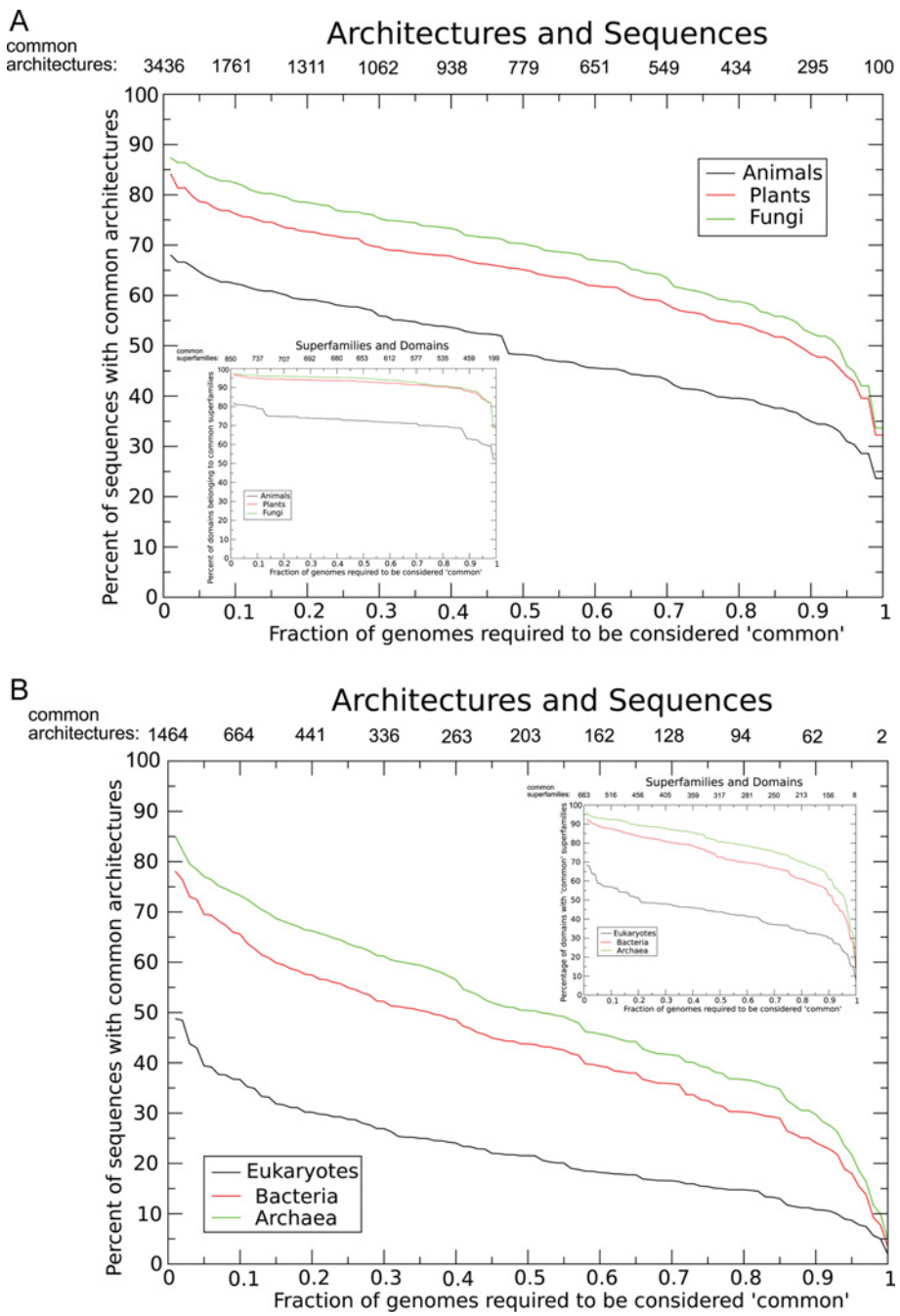


Figure 8 Domain architectures common to eukaryotes, prokaryotes and kingdoms

These two graphs show what proportion of sequences in a given set belong to architectures which are common to all of the sets. These graphs are similar to those in Figure 1 for domains and superfamilies; the corresponding graph in Figure 1 is shown in the inset for comparison. The sets are (**A**) animals, plants and fungi, and (**B**) eukaryotes, bacteria and archaea. The *x*-axis denotes the fraction of genomes in each set which are required to contain the superfamily in order to be considered 'common' for the purpose of calculating the values shown on the *y*-axis. Written across the top of the graphs are the numbers of architectures considered common under the condition specified by each tickmark on the *x*-axis. Data taken from the SUPERFAMILY database version 1.69 [8].

domains and has exploited duplication and divergence of existing domains to some degree, but the main way in which organisms adapt and evolve at the protein level is by the recombination of domains in its existing repertoire to produce novel architectures.

Functional modifications produced by domain combinations

A comparison was made of the functions of one-domain proteins and their homologues that are found in multidomain proteins [75]. It showed how the functions of individual domains in the multidomain protein combine to produce their overall functions and also the extent to which these functions are similar to those in the one-domain protein.

The types of functional changes produced by domain combinations fall into one of a number of different categories, as follows.

1. Proteins whose own functions are modified by another domain. The additional domain can: (a) be a non-enzyme that directly modifies substrate binding of an enzyme and vice versa;

(b) modify substrate binding through the formation of oligomers; (c) not have a function itself, but link together domains that are functional; (d) regulate enzyme function; (e) regulate DNA/RNA binding.

2. Formation of bifunctional enzymes.

3. Transfer of a part of the function found in a one-domain protein to additional domains, e.g. the placement of catalysis and substrate recognition on separate domains.

4. Combinations that allow a domain to function in new contexts, e.g. a one-domain electron-transport protein has a homologue that forms part of an electron-transport pathway in a multidomain protein.

5. Change in function in a one-domain homologue of a subunit of an oligomer, e.g. the homologue of a superoxide dismutase subunit that transports copper to this enzyme.

6. Gain of a catalytic activity, not present in a one-domain protein, by a homologue in a domain combination, e.g. naphthalene 1,2-dioxygenase: the catalytic domain is homologous with a transport protein. The second domain, a homologue of an electron-transport protein, provides a pathway to the active site.

7. The substrate, product and reaction of a one-domain protein are quite different from that of a homologue in a domain combination.

This list was derived from the data available at the time when the work was carried out, and it is likely that an analysis of structures determined subsequently will give new examples of functional changes produced by domain combination.

In categories 1–4, the function of the domains in the multi-domain protein and that one-domain homologue has often been conserved, in large part, but the overall function of the multi-domain protein has been modified or made more specific than that found in the one-domain homologue [75]. This is achieved by placing the homologous domain into a new domain combination context with an additional domain that serves to expand, alter or modulate its functionality.

The regularity roles of a set of 21 families of intracellular SMBDs have been described [76]. Members of the 21 families are commonly found in bacteria and archaea, and members of 16 of the families are also commonly found in eukaryotes. The families bind a wide variety of small organic molecules or metal ions. Their members are widely distributed among different multidomain proteins, which they regulate through their ligand-binding properties. The activities of the multidoman proteins include small-molecule transport, metabolic enzymes, regulation of transcription and signal transduction. The regularity role of the intracellular SMBD families is greater in prokaryotes than in eukaryotes that have their own specific SMBDs and extracellular domains that recognize small molecules [76].

ALTERNATIVE PROCESSES FOR THE EVOLUTION OF PROTEINS

Up to now, we have discussed how, during evolution, protein repertoires have greatly increased in their extent and complexity through the processes of gene duplications, sequence divergence and domain combinations. The retention of sequence similarities and/or particular structural features has made it possible to follow these processes in many large protein families over millions of years. There must, however, be cases where the divergence of sequence and structure has been so great, and the loss of intermediate structures so extensive, that the common origin of current members cannot be detected. There is at least one case where two proteins have structures that are entirely different, but whose gene sequence indicates their common origin.

A partial gene duplication that produces a new protein with entirely different sequence and structure

Fish in polar seas, where the temperature can be as low as $-1.9\,^{\circ}C$, are protected from freezing by AFGPs (antifreeze glycoproteins) that bind to ice crystals and prevent their growth. A family of antifreeze proteins in the Antarctic notothenioid fish have evolved through a partial duplication of the trypsinogen gene.

AFGPs have a repetitive sequence [77]. The basic unit is a tripeptide, Thr-Ala/Pro-Ala. These tripeptides form a number of different tandem repeats, and each of the different repeats are linked by three residues that have the conserved sequence Leu-Ile/Asn-Phe.

The repetitive nature of the AFGPs means that their protein sequences and structures are very different from that of the serine protease trypsinogen. Comparison of their gene sequences, however, shows that 5′ and 3′ regions of the trypsinogen gene are $\sim 95\,\%$ identical with regions of the AFGP genes [77]. The nucleotides that code for the Thr-Ala/Pro-Ala motif in AFGPs' protein sequences straddle an intron–exon boundary in the trypsinogen gene. Thus the formation of AFGPs involved the recruitment of terminal regions of the trypsinogen gene and extensive duplications of the nucleotides that code for the Thr-Ala/Pro-Ala motif.

Non-homologous recombination

It was proposed that combinations of non-homologous domain segments [78] or of exons [79] could produce novel proteins. So far, there is little evidence for this being a common process in the evolution of new proteins. However, a series of *in vitro* experiments have shown that novel stable proteins can be produced by combinatorial shuffling of polypeptide segments [80–82].

CspA (cold-shock protein A) has a sequence of 70 residues that form a five-stranded β-barrel. The initial 36 residues form a three-stranded β-sheet in the structure, but not in isolation. This segment was taken and copies fused to $\sim 10^8$ polypeptide segments encoded by randomly fragmented genomic *E. coli* DNA. From this collection of fused proteins, seven protease-resistant chimaeric proteins were isolated. One has the CspA segment fused to the 34 residues that form the C-terminal region of the 30S ribosomal protein [80]. Determination of the structure of this protein shows that the CspA segment retains a conformation very similar to that in the native protein. The structure of the 30S segment is a modified version of that found in the native protein. Together, the two segments form a six-stranded β-barrel with a novel fold [81]. A complication to this description of the monomer is that it rapidly forms dimers, which involves segments being swapped between two momomers, and then dimers combine slowly to form a tetrameric structure [82].

DISCUSSION

In the present review, we have described how the combined investigations of genome sequences and protein structures has greatly increased our understanding of the major evolutionary processes: gene duplication, divergence and combination. Many of the results we describe involve only the proteins and domains in genomes that are related to known structures. This is because it is only for these proteins that we have an accurate view of their evolutionary relationships. This set covers about half of those present in the genomes. It is likely that much of the remaining half will have features similar to those described here.

We conclude with three general comments on some of the results described above.

Nature's exploration of 'combination space'

Only a small proportion of the possible domain combinations are selected in Nature. At the present time, there are 677 genomes in SUPERFAMILY version 1.69 that together have 2.5 million sequences with assignments. These sequences have 63 000 unique architectures composed of domains from 1455 superfamilies. The total possible number of different domain architectures (restricted to only the observed numbers of domains in an architecture) is approx. 10^{11} times greater than the number seen in Nature.

The small size of domains

In most cases, domains have between 50 and 200 residues. This narrow range of sizes was explained by an analysis of the theoretically predicted and observed folding times of one-domain proteins [83,84]. This showed that folding time is related to the number of residues (N) and the free-energy difference between the native and unfolded states, ΔG, by the equation:

$$\text{Time} \approx \exp[(1 \pm 0.5)N2/3 + \Delta G/2RT] \text{ ns}$$

where the coefficient 1 ± 0.5 depends on the structure of the native fold. The experimentally measured term, $\Delta G/2RT$, hardly exceeds 10 for one-domain proteins and it is relatively small. Thus this equation implies that, domains of 300 residues and having a common type of structure, i.e. corresponding to $\exp[N2/3]$ ns, could not fold in any reasonable time.

The origin of duplication divergence and combination

The earliest evolution of the proteins must have involved *ab initio* invention of new proteins. But, in biology as we now know it, there is, at the present time, little evidence for *ab initio* processes. Two reasons have been suggested for it being absent or rare [17]. 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 modified or changed functions by duplication, divergence and combination. Secondly, the error-correction procedures now present in DNA replication and protein synthesis make the *ab initio* invention process too slow to be significantly useful.

ACKNOWLEDGEMENTS

We thank Madan Babu, Sarah Teichmann, Alexey Murzin, Antonia Andreeva and Martin Madera for discussions of this review. We also thank Ralph Pethica for help with the Figures.

FUNDING

This work received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1 Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. and North, A. C. T. (1960) Structure of haemoglobin. Nature **185**, 416–422
2 Rossmann, M. G., Moras, D. and Olsen, K. W. (1974) Chemical and biological evolution of a nucleotide-binding protein. Nature **259**, 194–199
3 Birktoft, J. J., Blow, D. M., Henderson, R. and Steitz, T. A. (1970) I. Serine proteases: the structure of α-chymotrypsin. Philos. Trans. R. Soc. London Ser. B **257**, 67–76
4 Tang, J., James, M. N. G., Hsu, I. N., Jenkins, J. A. and Blundell, T. L. (1978) Structural evidence for gene duplication in the evolution of the acid proteases. Nature **271**, 618–621
5 Patthy, L. (1994) Exons and introns. Curr. Opin. Struct. Biol. **4**, 383–392
6 Murzin, A. G., Brenner, S. E., Hubbard, T. H. and Chothia, C. (1995) SCOP: the structural classification of proteins database. J. Mol. Biol. **247**, 536–540
7 Murzin, A. G. (1998) How far divergent evolution goes in proteins. Curr. Opin. Struct. Biol. **8**, 380–387
8 Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Madera M., Chothia C. and Gough, J. (2009) SUPERFAMILY: sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. **37**, D380–D387
9 Brenner, S. E., Hubbard, T., Murzin, A. and Chothia, C. (1995) Gene duplications in *H. influenzae*. Nature **378**, 140
10 Teichmann, S. A., Park, J. and Chothia, C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. Proc. Natl. Acad. Sci. U.S.A. **95**, 14658–14663
11 Muller, A., MacCallum, R. M. and Sternberg, M. J. E. (2002) Structural characterization of the human proteome. Genome Res. **12**, 1625–1641
12 Qian, J., Luscombe, N. M. and Gerstein, M. (2001) Protein family fold occurrence in genomes. J. Mol. Biol. **313**, 673–681
13 Koonin, E. V., Wolf, Y. I. and Karev, G. P. (2002) The structure of the protein universe and genome evolution. Nature **420**, 218–223
14 Huynen, M. A. and van Nimwegen, E. (1998) The frequency distribution of gene families in complete genomes. Mol. Biol. Evol. **15**, 583–589
15 van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. Trends Genet. **19**, 479–484
16 Ranea, J. A., Buchan, D. W., Thornton, J. M. and Orengo, C. A. (2004) Evolution of protein superfamilies and bacterial genome size. J. Mol. Biol. **336**, 871–887
17 Chothia, C., Gough, J., Vogel, C. and Teichmann, S. A. (2003) Evolution of the protein repertoire. Science **300**, 1701–1703
18 Vogel, C. and Chothia, C. (2006) Protein family expansions and biological complexity. PLoS Comput. Biol. **2**, e48
19 Jordan, I. K., Makarova, K. S., Spouge, J. L. and Koonin, E. V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. **11**, 555–565
20 Lespinet, O., Wolf, Y. I., Koonin, E. V. and Aravind, L. (2002) The role of lineage-specific gene family in the evolution of eukaryotes. Genome Res. **12**, 1048–1059
21 Vogel, C., Teichmann, S. A. and Chothia, C. (2003) The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. Development **130**, 6317–6328
22 Zuckerkandl, E. (1976) Evolutionary processes and evolutionary noise at the molecular level. 1. Functional density in proteins. J. Mol. Evol. **7**, 167–183
23 Rocha, E. P. C. and Danchin, A. (2004) An analysis of the determinants of amino substitution rates in bacterial proteins. Mol. Biol. Evol. **21**, 108–116
24 Drummond, D. A., Raval, A. and Wilke, C. O. (2006) A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. **23**, 327–337
25 Teichmann, S. A. (2002) The constraints protein–protein interactions place on sequence divergence. J. Mol. Biol. **324**, 399–407
26 Pal, C., Papp, B. and Lercher, M. J. (2006) An integrated view of protein evolution. Nat. Rev. Gene **7**, 337–348
27 Sasidharan, R. and Chothia, C. (2007) The selection of acceptable mutations. Proc. Natl. Acad. Sci. U.S.A. **104**, 10080–10085
28 Kumar, M. D., Bava, K. A., Gromiha, M. M., Parabakaran, P., Kitajima, K., Uedaira, H. and Sarai, A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res. **34**, D204–D206
29 Serrano, L., Day, A. G. and Fersht, A. R. (1993) Step-wise mutation of barnase to binase: as procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. J. Mol. Biol. **233**, 305–312
30 Steipe, B., Schiller B., Pluckthun A. and Steinbacher, S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. J. Mol. Biol. **240**, 188–192
31 Nikolova, P. V., Henckel, J., Lane, D. P. and Fersht, A. R. (1998) Semirational design of active tumor suppressor p53 DNA binding suppressor. Proc. Natl. Acad. Sci. U.S.A. **94**, 14675–14680
32 Wang, Q., Buckle, A. M., Foster, N. W., Johnson, C. M. and Fersht, A. R. (1999) Design of a highly stable functional GroEL minichaperones. Protein Sci. **8**, 2186–2193
33 Perutz, M. F., Kendrew, J. C. and Watson, H. C. (1965) Structure and function of haemoglobin II: some relations between polypeptide chain configuration and amino acid sequence. J. Mol. Biol. **13**, 669–678
34 Gerstein, M., Sonnhammer, E. L. L. and Chothia, C. (1994) Volume changes in protein evolution. J. Mol. Biol. **236**, 1067–1078
35 Chothia, C., Gelfand, I. and Kister, A. (1998) Structural determinants in the sequences of immunoglobulin variable domains. J. Mol. Biol. **278**, 457–479
36 Lieberman, R., Potter, M., Humphrey, Jr, W., Mushinski, E. B. and Vrana, M. (1975) Multiple individual and cross-specific idiotypes of 13 levan-binding myeloma proteins of BALB/c mice. J. Exp. Med. **142**, 106–119
37 Proba, K., Honegger, A. and Plückthun, A. (1997) A natural antibody missing a cysteine in VH: consequences for thermodynamic stability and folding. J. Mol. Biol. **265**, 161–172

38 Hamill, S. J., Cota, E., Chothia, C. and Clarke, J. (2000) Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary *cul-de-sac*. J. Mol. Biol. **295**, 641–649

39 Lesk, A. M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J. Mol. Biol. **136**, 223–268

40 Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. **5**, 823–826

41 Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. and Matthews, B. W. (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science **255**, 178–183

42 Lo, W.-C., Lee, C.-C., Lee, C.-Y. and Lyu, P.-C. (2009) CPDB: a database of circular permutations in proteins. Nucleic Acids Res. **37**, D328–D332

43 Vogel, C. and Morea, V. (2006) Duplication, divergence and the formation of novel protein topologies. BioEssays **28**, 973–978

44 Cunningham, B. A., Heperley, J. J., Hopp, T. P. and Edelman, G. M. (1979) Flavin versus comcanavalin A: circularly-permuted amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. **76**, 3218–3222

45 Peisajovich, S. G., Rockah, L. and Tawfik, D. S. (2006) Evolution of new protein topologies through multistep gene rearrangements. Nat. Genet. **38**, 168–174

46 Grishin, N. V. (2001) Fold change in the evolution of structures. J. Struct. Biol. **134**, 167–185

47 Andreeva, A. and Murzin, A. G. (2006) Evolution of protein fold in the presence of functional constraints. Curr. Opin. Struct. Biol. **16**, 399–408

48 Teichmann, S. A., Rison, S. C. G., Thornton, J. M., Riley, M., Gough, J. and Chothia, C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. J. Mol. Biol. **311**, 693–708

49 Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. **30**, 409–425

50 Babbitt, P. C. and Gerlt, J. A. (1997) Understanding enzyme superfamilies: chemistry as the fundamental determinant in the evolution of new catalytic activities. J. Biol. Chem. **272**, 30591–30594

51 Wilson, C. A., Kreychman, J. and Gerstein, M. (2000) Assessing annotation transfer for genomics. J. Mol. Biol. **297**, 233–249

52 Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001) Evolution of function in protein superfamilies from a structural perspective. J. Mol. Biol. **307**, 1113–1143

53 Bartlett, G. J., Borkakoti, N. and Thornton, J. M. (2003) Catalysing new reactions during evolution: economy of residues and mechanism. J. Mol. Biol. **331**, 829–860

54 Todd, A. E., Orengo, C. A. and Thornton, J. M. (2002) Sequence and structural differences between enzyme and nonenzyme homologues. Structure **10**, 1435–1451

55 Pils, B. and Schultz, J. (2004) Inactive enzyme-homologues find new function in regulatory processes. J. Mol. Biol. **340**, 399–404

56 Aravind, L., Lyer, L. M. and Koonin, E. V. (2006) Comparative genomics and structural biology of molecular innovations of eukaryotes. Curr Opin. Struct. Biol. **16**, 409–419

57 Koonin, E. V. and Aravind, L. (2002) Origin and evolution of eukaryotic apoptosis: the bacterial connection. Cell Death Differ. **9**, 394–404

58 Gough, J. (2006) Genomic scale sub-family assignment of protein domains. Nucleic Acids Res. **34**, 3625–3633

59 Bashton, M. and Chothia, C. (2002) The geometry of domain combination in proteins. J. Mol. Biol. **315**, 927–939

60 Gerstein, M. (1998) How representative are the known structures of proteins in a complete genome? A comprehensive structural census. Fold. Des. **3**, 497–512

61 Basu, M. K., Carmel, L., Rogizin, I. B. and Koonin, E. V. (2008) Evolution of protein domain promiscuity in eukaryotes. Genome Res. **18**, 449–461

62 Apic, G., Gough, J. and Teichmann, S. A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol. **310**, 311–325

63 Vogel, C., Teichmann, S. A. and Pereira-Leal, J. (2005) The relationship between domain duplication and recombination. J. Mol. Biol. **346**, 355–365

64 Kopelman, N. M., Lancet, D. and Yanai, I. (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat. Genet. **37**, 588–589

65 Talavera, D., Vogel, C., Orozco, M., Teichmann, S. A. and de la Cruz, X. (2007) The (in)dependence of alternative splicing and gene duplication. PLoS Comput. Biol. **3**, 375–388

66 Vogel, C., Berzuini, C., Bashton, M., Gough, J. and Teichmann, S. A. (2004) Supra-domains: evolutionary units larger than single protein domains. J. Mol. Biol. **336**, 809–823

67 Björklund, I. K., Ekman, D., Light, S., Frey-Skött, J. and Elofsson, A. (2005) Domain rearrangements in protein evolution. J. Mol. Biol. **353**, 911–923

68 Aroul-Selvam, R., Hubbard, T. and Sasidharan, R. (2004) Domain insertions in protein structures. J. Mol. Biol. **338**, 633–641

69 Kummerfeld, S. K. and Teichmann, S. A. (2004) Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. **17**, 589–596

70 Weiner, J., Beaussart, F. and Bornberg-Bauer, E. (2006) Domain deletions and substitutions in the modular protein evolution. FEBS J. **273**, 2037–2047

71 Björklund, A. K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats. PLoS Comput. Biol. **2**, 959–970

72 Gough, J. (2005) Convergent evolution of domain architectures (is rare). Bioinformatics **21**, 1464–1471

73 Koonin, E. V., Aravind, L. and Kondrashov, A. S. (2000) The impact of comparative genomics on our understanding of evolution. Cell **101**, 573–576

74 Hill, E., Broardbent, I. D., Chothia, C. and Pettitt, J. (2001) Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. J. Mol. Biol. **305**, 1011–1024

75 Bashton, M and Chothia, C. (2007) The generation of new protein functions by the combination of domains. Structure **15**, 85–99

76 Anantharaman, V., Koonin, E. V. and Aravind, L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small molecule binding domains. J. Mol. Biol. **307**, 1271–1292

77 Chen, L., DeVries, A. L. and Cheng, C.-H. (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. U.S.A. **94**, 3811–3816

78 Blake, C. C. F. (1978) Do genes-in-pieces imply proteins-in-pieces? Nature **273**, 267

79 Gilbert, W. (1978) Why genes in pieces? Nature **271**, 501

80 Reichmann, L. and Winter, G. (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. Proc. Natl. Acad. Sci. U.S.A. **97**, 10068–10073

81 de Bono, S., Riechmann, L., Girard, E., Williams, R. L. and Winter, G. (2005) A segment of cold shock protein directs the folding of a combinatorial protein. Proc. Natl. Acad. Sci. U.S.A. **102**, 3811–3816

82 Riechmann, L., Lavenir, I., de Bono, S. and Winter, G. (2005) Folding and stability of a primitive protein. J. Mol. Biol. **348**, 1396–1401

83 Finkelstein, A. V. and Badretdinov, A. Y. (1997) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold. Des. **2**, 115–121

84 Finkelstein, A. V. and Ptitsyn, O. B. (2002) Lecture 21. Protein Physics: a Course of Lectures, pp. 263–277, Academic Press, London

Received 19 January 2009; accepted 23 January 2009
Published on the Internet 13 March 2009, doi:10.1042/BJ20090122