The Relationship Between Domain Duplication and Recombination

Christine Vogel*, Sarah A. Teichmann and Jose Pereira-Leal

MRC Laboratory of Molecular Biology, Hills Rd, Cambridge CB2 2QH, UK

Introduction

Protein domains represent the basic evolutionary units that form proteins. Domain duplication and shuffling by recombination are probably the most important forces driving protein evolution and hence the complexity of the proteome. While the duplication of whole genes as well as domain-encoding exons increases the abundance of domains in the proteome, domain shuffling increases versatility, i.e. the number of distinct contexts in which a domain can occur. Here, we describe a comprehensive, genome-wide analysis of the relationship between these two processes.

We observe a strong and robust correlation between domain versatility and abundance: domains that occur more often also have many different combination partners. This supports the view that domain recombination occurs in a random way. However, we do not observe all the different combinations that are expected from a simple random recombination scenario, and this is due to frequent duplication of specific domain combinations. When we simulate the evolution of the protein repertoire considering stochastic recombination of domains followed by extensive duplication of the combinations, we approximate the observed data well.

Our analyses are consistent with a stochastic process that governs domain recombination and thus protein divergence with respect to domains within a polypeptide chain. At the same time, they support a scenario in which domain combinations are formed only once during the evolution of the protein repertoire, and are then duplicated to various extents. The extent of duplication of different combinations varies widely and, in nature, will depend on selection for the domain combination based on its function. Some of the pair-wise domain combinations that are highly duplicated also recur frequently with other partner domains, and thus represent evolutionary units larger than single protein domains, which we term “supra-domains”.

Abbreviation used: SCOP, structural classification of proteins.

E-mail address of the corresponding author: cvogel@mrc-lmb.cam.ac.uk

*Corresponding author

Keywords: domain recombination; neutral evolution; domain versatility; domain shuffling

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
networks are known to be scale-free, which means that a few superfamilies are connected to many different superfamilies, while most superfamilies are adjacent to only one or two types of neighbours.5,6 Duplication, transposition and horizontal gene transfer result in the expansion of domain superfamilies in terms of their abundance; that is, the number of occurrences of a domain in the proteome of one organism (Figure 1(c)). Protein and domain duplication have been studied extensively and have been connected with an increase in an organism’s complexity in evolution, for example expansions of particular families in higher eukaryotes.7–15 Domains can recombine to form multi-domain proteins, and proteins with two or more domains constitute the majority of proteins in all organisms studied.16,17 Thus, the recombination of existing domains may be a major mechanism that modifies protein function18 and increases proteome complexity.16,19 The combination or shuffling of domains increases what we term the versatility of a domain superfamily; that is, the number of different partner domains that domains of a particular superfamily are adjacent to (Figure 1(d)).

The exact interplay of the mechanisms that underlie the evolution of the protein repertoire is still not completely resolved. In particular, it is unclear to what extent the recombination of protein domains, or domain versatility, is under selection, or if it is the result of a purely random shuffling process.20 Next, we discuss the arguments for both mechanisms.

The frequency distribution of domain family sizes, of protein length measured as number of domains per protein and the distribution of domain versatility are all best approximated by a power-law function.5,6,21 These mathematical relationships are often taken as support of a neutral or random mode of evolution.5,20–23 However, it is still unclear to what extent a random model holds true for detailed examples.

The observation that the number of domain combinations in nature is only a small fraction of the possible number of combinations suggests that domain recombination is under strong selection.5,24 In fact, several lines of evidence support this view.

First, eukaryotic proteins have more different domains per protein than prokaryotic ones.11,16,25–27 They also have more different domain combinations than simpler organisms.5,11,16,27,28 This means that individual domain superfamilies are observed with a greater variety of combination partners in higher organisms. However, eukaryotes also have larger genomes and higher rates of retention of duplicated genes, which questions the extent to which these observations are evidence for selection. Furthermore, a detailed analysis of the number of co-occurring domains in human and other eukaryotes showed no change in versatility in terms of co-occurring domains, but just a change in the repertoire.15

Figure 1. The duplication and recombination of domains form the protein repertoire. The Figure illustrates, from a domain perspective, the main processes that formed the protein repertoire and which are discussed here: the duplication and recombination of domains. (a) Proteome of an organism, consisting of eight different proteins (black lines) with domains of different superfamilies represented as boxes of different colours. (b) An alternative representation of the proteome: the colour of a circle represents a particular superfamily, the size of the circle corresponds to the abundance of the domains of that superfamily in the genome. The connections between the circles denote domain combinations. (c) Abundance is the result of a variety of duplication processes. The abundance of each domain superfamily is the number of domains in the proteins of the proteome, disregarding tandem repeats. (d) Versatility is the number of different N and C-terminal partner domains of a particular superfamily that are observed. The example shows the combinations of the blue superfamily: it occurs with one N-terminal and two different C-terminal partner superfamilies.
Next, domains with “generic” or especially “useful” functions clearly are the most versatile superfamilies in many organisms. These are, for example, cofactor or co-substrate-binding domains, like the P-loop nucleotide triphosphate hydrolase or the NAD(P)-binding Rossmann domains. Protein–protein interaction domains are also found in multi-domain proteins adjacent to a variety of other domains, and these different combinations are used to regulate distinct aspects of cellular organization. These domain superfamilies are, however, also very abundant in genomes, thus it is unclear if their great versatility is suggestive of selection or just a consequence of their high abundance.

Furthermore, domains that co-occur in proteins are more likely to display similar function or localisation than domains in separate proteins, and may support selection acting on domain combinations. However, this could be due to a bias in the function classification schemes towards annotation of whole proteins rather than domains. Finally, since some highly abundant folds tend to have particular structures, it is possible that the three-dimensional structure of a domain may impose constraints on its ability to combine with other domains and hence reduce its versatility. However, to the best of our knowledge no such constraints have been observed.

In summary, although there is support for selective forces acting on the recombination of protein domains, there are also observations that suggest a random process. Here, we present a comprehensive, genome-wide study of domain recombination, testing the hypothesis that it is a random process. We use a data-driven approach combined with simulations to show that the domain combinations observed are consistent with stochastic recombination together with differential duplication of domain combinations. Duplication of domain combinations is much more common than invention of new combinations. Selection at this level has resulted in a few highly duplicated, very abundant domain combinations in genomes, while the bulk of domain combinations are rare, occurring in only a few proteins.

Results and Discussion

The abundance of domain superfamilies correlates with their number of combination partners

In a random scenario, more trials result in higher success rates. For domains, the domain abundance represents the number of trials for recombination. This means that the more abundant domains would be expected to be “luckier” and have more combination partners than the less abundant domains. This is illustrated in Figure 1(b) where the number of edges pointing to and from a node (connectivity) would be expected to correlate with the size of the node.

A random scenario would also imply that the abundance of a domain alone should determine how often it is found with a different combination partner. Thus, we compared abundance and versatility of each superfamily in various genomes, and the data for the human proteome is plotted in Figure 2. We observe that the absolute versatility ($V$) of a particular superfamily. The plot shows the data for human, and the distribution is very similar for all other eukaryotes. In bacteria and archaea, the abundance and versatility is much smaller, thus a relationship between the two measures is much less obvious. The plotted data fit a power-law function with $V \sim A^q$ and $0.42 \pm 0.04 (r^2=0.75)$ best, compared to an exponential, linear or logarithmic function.

Figure 2. Domain versatility and abundance. The abundance and versatility of all superfamilies is plotted on a log-log-scale. Each point represents the abundance and versatility of a particular superfamily. The plot shows the data for human, and the distribution is very similar for all other eukaryotes. In bacteria and archaea, the abundance and versatility is much smaller, thus a relationship between the two measures is much less obvious. The plotted data fit a power-law function with $V \sim A^q$ and $0.42 \pm 0.04 (r^2=0.75)$ best, compared to an exponential, linear or logarithmic function.
Figure 3. The relationship between versatility and abundance across different domain properties. Versatility is plotted as a function of abundance similar to Figure 2. In each panel, the superfamily data on 42 different genomes is grouped according to four different criteria: structural class and phylogeny (affiliation to kingdom) domain molecular function and biological process. See Methods and Data Sets for a detailed description of the groupings. Left: each panel shows the lines fitted through the superfamily groupings. Right: a graph displays the slopes $\theta$ of the fitted lines on the left together with their standard deviations.
eukaryote genomes in our data set; similar to the human data shown in Figure 2.

The versatility relative to abundance is constant

Next, we consider the factors that can affect domain versatility, as discussed in the Introduction. We ask whether distinct groups of domains, defined according to their function, structure and evolutionary history, display different relative versatilities. Such differences would reveal selection, as some combinations would be more duplicated, thus having higher abundances without the corresponding increase in versatility expected from a random scenario.

In order to address this hypothesis, we test whether the relationship \( V \sim A^q \) holds true in particular subsets of domains and, whether the exponent \( q \) is constant. If the relative rates of domain duplication and combination are dictated by characteristics of their component domains, one would expect that: (i) the power-law relationship does not hold for one or more of the domain groups; or that (ii) the slope \( q \) for domains of group \( x \) is different from the slope \( q \) for domains of group \( y \). This will tell us whether the relative rate is different for distinct groups of domains.

We divide protein domains according to structural classes and folds, phylogenetic affiliation, molecular function and biological process as described in Methods and Data Sets. We examine the relationship between versatility and abundance in each subset (Figure 3). For each grouping, we observe a remarkable consistency of (i) the power-law function being the best fit to the data and (ii) the slope \( q \), the relative versatility across all subgroups (Figure 3). Thus, the relationship between versatility and abundance is very similar in all subsets of domains.

This confirms that the versatility of a domain in a particular genome can be predicted relatively accurately purely from its abundance in that genome, independently of other properties of the domain. The combination of domains into novel contexts is constant relative to its duplication. With respect to the illustration in Figure 1(b), this means that the connectivity of the node is correlated with its size.

Given the vast variety of domain functions, structures and other characteristics considered here, this is a surprising result. It is consistent with the hypothesis that the versatility of protein domains is the result of a stochastic process in which more abundant domains naturally have a higher chance of combining with different partner domains than less abundant ones. In other words, our findings can be explained by a random model of domain shuffling; and the null hypothesis of random recombination cannot be rejected.

**S0: a random model of domain combination**

The dependence of domain versatility on abundance suggests that new combinations of domains occur in a random or neutral fashion, independently of any other particular circumstances. If this is the case, random shuffling of domains given their distribution of abundances should result in a similar absolute versatility for individual superfamilies and thus a similar relationship between abundance and versatility as that observed. We term this model S0.

We performed 10,000 random shuffling experiments where the abundance of each domain, the number and their size in terms of domains are maintained for each genome considered. The expected maximal versatility we obtained in this random shuffling is also related to abundance by a power-law function (Figure 4, blue), which supports a random model. Highly abundant domains are more versatile, and conversely, less abundant domains are less versatile.

However, the expected maximal relative versatility of each domain according to this random scenario \( \theta_{\text{rand}} = 0.72 \pm 0.01 \) is considerably higher than what we observe in the genomes \( \theta_{\text{obs}} = 0.42 \pm 0.04 \). This means that overall the observed relative versatility of protein domains is lower than that expected by simple random shuffling of domains. This represents a marked discrepancy between the model S0 and the observed data (Figure 4). Inspection of this discrepancy reveals that a small proportion of the domain combinations are more frequent than expected by random shuffling of domains, whereas a larger proportion of combinations are under-represented; the random shuffling of domains results in a more even distribution of domain combinations and abundances. This means that in the genomes, many domain superfamilies have fewer combination partners than expected by chance, which is consistent with previously reported results.20

For example, in the human genome, these under-represented combinations correspond to 7% (61/856) of all observed pair-wise combinations and under-represented combinations correspond to 41% (352/856). One example is the SH3 domain, which is a highly abundant domain (226 copies in the human genome). This domain combines with the SH2 and protein kinase-like domains much more frequently than would be expected given the abundance of the two domain superfamilies: 31 times with an SH2 domain and 27 times with a protein kinase domain given a single expected occurrence of each type. This suggests that the two domain combinations have been duplicated repeatedly, and as a consequence other domain combinations involving the SH3 domain are under-represented.

Thus, when the combination is duplicated, the abundance of each component domain is increased but the versatility is kept constant. This explains why the versatility of protein domains is not at the maximal level possible according to their abundance in a random shuffling model. The discrepancy between the predictions of the S0
Domain Duplication and Recombination

model and the observed data is because the model does not take into account the duplication of domain combinations.

**S1: a random model of domain combination with duplication of combinations**

Given the observation that domain combinations in genomes are often more duplicated than simulated combinations in the random model S0, we implemented a different model, which we termed S1. In this model we explicitly consider the duplication of domain combinations. As in model S0, the domain abundance, number and size of proteins are maintained. Thus, we examine recombination given selection processes that may have produced the observed number of domains per superfamily. All domains are randomly shuffled. As soon as one combination is formed, we regard this combination as fixed and it is duplicated until all the instances of at least one of the component domains are used up. This reduces the abundance of the component domains that are available for further re-shuffling. The recombination and subsequent duplication is repeated until all domains are placed into combinations. This simulation was repeated 10,000 times. The relationship between the average versatility and abundance predicted by this model is shown in Figure 4. It is a substantially better approximation of the real data than model S0.

Model S1 reproduces the observed data four times better than model S0: model S1 correctly predicts the number of different combination partners for 39% of all superfamilies in human, whereas model S0 predicts only 10% (within 20% error). A correct prediction of domain versatility is most meaningful for large domain superfamilies, and this is where model S1 improves enormously compared to model S0: for superfamilies that occur in ten or more proteins, model S1 predicts 54% correctly, and for superfamilies that occur in 30 or more proteins, model S1 even predicts two-thirds correctly (66%). Model S0 correctly predicts only 8% and 3% of the abundant superfamilies, respectively (Figure 4). In contrast to model S0, model S1 is best approximated by a linear relationship between domain versatility and abundance ($r^2 = 0.991$) but is also well approximated by a power-law ($r^2 = 0.948$). Since we lack data especially for domains of high abundance and versatility, it remains inconclusive when the approximations from model S1 would break down.

Given that the model S1, emphasizing duplication of combinations, is much more accurate than the first random model, we conclude that duplication and retention of domain combinations due to selection plays an important role in the evolution of the protein repertoire. We show that recombination and duplication of domain combinations can be approximated by our stochastic simulation, and this suggests that selection favours duplicates of an existing combination much more often than novel combinations. This means that it is easier or more likely that a functional protein retained by selection evolves by sequence divergence and mutation of an existing combination, rather than combining existing domains in a new way.

**Prevalent domain combinations**

Model S1 is consistent with the observed data, predicting versatilities that are lower than S0, and approximating better the overall relationship between versatility and abundance. It also predicts that for a given domain, a few combinations will be highly duplicated whereas most combinations will be infrequent. To what extent do we observe this prediction in the protein repertoire? In order to answer this question, we analysed the abundance distribution of combinations of particular superfamilies: we wanted to know whether for each particular superfamily, there are one or a few prevailing combinations.

The abundance distribution for the pair-wise combinations of large superfamilies in the human genome is shown in Figure 5. For example, the SH3 domain occurs 31 times in combination with the SH2 domain, as mentioned above. The next most abundant combination of the SH3 domain is with the PDZ peptide-binding domain and the P-loop nucleotide triphosphate hydrolase domains: each combination occurs 12 times in the human genome. The other 24 combinations of the SH3 domain occur much less frequently. The distribution in Figure 5 is in accordance with our prediction that large superfamilies have only one or few combinations with many duplicates.

These results and the striking fact that model S1 accounts for about two-thirds of the large superfamilies support the possibility that the mechanisms simulated in this model are an approximation...
of the evolution of protein repertoires. We do not claim that the model resolves the chronological or evolutionary sequence of events, but simply that the concerted action of random re-shuffling and selection and duplication of specific combinations can have formed the protein repertoire. Furthermore, while model S1 clearly does not reproduce the exact versatility of each individual superfamily, we demonstrate how the observed relationship can result from an underlying stochastic process of evolution, involving random domain shuffling. The model also includes selection at the level of duplication of domain combinations: a few combinations are highly duplicated, while most have been duplicated and retained to a much lesser extent. As is obvious from Figure 4, the observed domain combinations have a broad distribution of abundance and versatility, suggesting that nature employs some combination of model S0, S1 and other processes.

Conclusions

The interplay of neutral processes and selection has been debated for a long time. Upon gene duplication, the redundant copy is under selection, and the “usefulness” and ability to sub-functionalise, i.e. modify function, is thought to increase a duplicate’s chance of retention. In order to modify function, the duplicate diverges, for example, with respect to the sequence or the spatial and temporal expression. Another mode of divergence that can change protein function has been discussed here: the recombination with other protein domains.

We described an analysis of the general processes that may have governed the recombination of protein domains. We compared the known domain superfamilies and their arrangements in proteins in a variety of genomes to stochastic models of domain recombination and duplication. We defined domain versatility as the number of domain superfamilies that a particular domain can combine with, and we observed a robust correlation between domain abundance and versatility by a power-law. This correlation is independent of domain function, structure or evolutionary history. It suggests that, similar to transcriptome evolution, domain reshuffling represents a form of random protein divergence.

Despite this evidence for random evolution of domain combinations, the observed domain versatility in genomes is very different to that resulting from a simple random model (S0). This is due to highly duplicated domain combinations, and many of these qualify as supra-domains, which are evolutionary units larger than a single domain and which recur in many different proteins. One such supra-domain is the combination of the P-loop nucleotide triphosphate hydrolase domain with the translation proteins domain. Proteins with this domain combination hydrolyse GTP and interact with the ribosome, and these are functions that are needed by all translation factors. Thus, the combination recurs in 23 proteins in human. When incorporating the duplication of particular domain combinations into a more elaborate model, we were able to approximate the observed versatility fairly accurately. Thus, selection appears to occur at the level of retaining particular domains and domain combinations after duplication, while the footprint of selection is less clear at the level of domain recombination.

The emphasis on duplication of domain combinations supports the view that most combinations formed once during evolution, and all examples of a combination are related duplicates rather than emerging from independent recombination events. This is consistent with: (i) the prevalence of most combinations in only one N to C-terminal order,
Despite no obvious structural or functional constraints on the sequential order; (ii) structural evidence of an evolutionary relationship between all instances of the same domain combination; and (iii) the lack of functional conservation if the domain order is modified.

Such a view implies that duplication of combinations is more parsimonious than several independent recombination events, and predicts the prevalence of one or two combination partners for a particular domain superfamily in a genome. We observe such a distribution of combination partners in Figure 5. Usually, modification of an existing domain occurs by incremental mutation of the coding sequence. The recombination with a novel partner domain, however, is a different and more drastic type of modification as it can require the evolution of a new domain interface, for example. Once the domain combination is established as a functional unit, it is “easier” for nature to retain duplicates of this combination and re-use them in different proteins.

**Methods and Data Sets**

**Genomic and domain assignment data**

The 14 eukaryote, 14 bacterial and 14 archaeal genomes used in our analysis are listed in Table 1. The gene predictions and domain assignments to the gene predictions were taken from the SUPERFAMILY database version 1.63. On average, about 45% of the sequences had at least one domain predicted.
Domain properties

Abundance and versatility

The duplication or abundance of a domain superfamily in each genome was measured as the number of domains belonging to the particular superfamily (Figure 1(c)). The recombination or absolute versatility of a particular superfamily was measured as the sum of different N and C-terminal domains that are immediately adjacent to the domain (Figure 1(d)). Tandem duplications of the same domain within one protein are ignored. The power-law function was fitted using regression. The minimal versatility of a domain and thus the intercept of the graph is known (no combination partners), so it seemed meaningful to restrict the analysis to the slope or exponent $q$.

Alternative definitions of abundance and versatility

We tested our results on alternative definitions of abundance and versatility, and found the same general relationships. When we considered the number of proteins instead of the number of domains as a measure of abundance, or when we considered the number of co-occurring domains instead of immediately adjacent domains as a measure of versatility, the exponent $q$ increased slightly. In both cases, the versatility relative to the abundance is slightly higher, but their power-law relationship is the same. The same holds true if we do account for tandem duplications of domains within one protein.

Structural classification

Our analysis is based on the four main structural classes in the SCOP database: alpha, beta, alpha/beta and alpha + beta. Each class contains several folds, and each fold contains several superfamilies. For the hierarchical classification of protein domains, please refer to the SCOP database.2,3

Annotation of domain function and process

Manual procedure

All domain superfamilies were manually annotated with respect to their dominant molecular function, i.e. the function within the protein, and their dominant biological process, i.e. their role in the cell. Using a manual scheme, we ensured a high-quality annotation and a domain-centred annotation instead of the widely used whole-gene or whole-protein centred annotation. The different schemes are described below.

Scheme for biological processes of domains. For an analysis of the biological process in which domain superfamilies act, we annotated all superfamilies with respect to their affiliation to one of 43 different process categories (Table 2). All 43 functional categories were then mapped onto one of eight process classes which are defined as follows.

(i) Information: storage, maintenance of the genetic code.
(ii) Regulation: regulation of gene expression, protein activity; information processing in response to environmental input.
(iii) Metabolism: all anabolic and catabolic processes; cell maintenance/homeostasis.
(iv) Energy: energy production/storage.
(v) Processes_IC: intra-cellular processes; cell motility/division; transport.
(vi) Processes_EC: inter or extra-cellular processes; organismal processes (such as blood clotting).
(vii) General: general enzymatic reactions and multiple functions.
(viii) Unknown/Other: unknown or unclassifiable superfamilies, viral functions.

The functional distribution of all SCOP superfamilies and those in the respective genomes is available upon request.

Scheme for molecular functions of domains. All domain superfamilies were assigned to one of 27 functional

---

Table 2. Classification scheme for biological processes of domains

<table>
<thead>
<tr>
<th>Biological process class</th>
<th>Biological process category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>Chromatin structure and dynamics; translation, ribosomes, ribosome biogenesis; tRNA metabolism; transcription, DNA replication, recombination, repair; RNA processing, dynamics; nuclear structure</td>
</tr>
<tr>
<td>Regulation</td>
<td>RNA processing and modification; DNA-binding (transcription factors); kinases and phosphatases and inhibitors; signal transduction</td>
</tr>
<tr>
<td>Metabolism</td>
<td>Electron transfer/transport; amino acid transport and metabolism; nitrogen metabolism; nucleotide transport and metabolism; carbohydrate transport and metabolism; polysaccharide metabolism; lipid/poly saccharide storage; coenzyme metabolism; lipid transport and metabolism; cell envelope biogenesis, outer membrane; secondary metabolites biosynthesis, transport and catabolism; oxidation/reduction; transferases; other enzymes</td>
</tr>
<tr>
<td>Energy</td>
<td>Energy production and conversion; photosynthesis</td>
</tr>
<tr>
<td>Intra-cellular processes, Processes_IC</td>
<td>Cell division and chromosome partitioning, cell cycle; phospholipid metabolism; cell motility, cytoskeleton; intracellular trafficking and secretion; post-translational modification, protein turnover, chaperones; proteases, peptidases and their inhibitors; inorganic ion transport and metabolism; transport</td>
</tr>
<tr>
<td>Inter-cellular processes, Processes_EC</td>
<td>Cell adhesion; immune response; blood clotting; toxins and defense enzymes</td>
</tr>
<tr>
<td>General</td>
<td>Small molecule binding; general or several functions; protein–protein interaction</td>
</tr>
<tr>
<td>Unknown/Other</td>
<td>Function unknown; viral proteins</td>
</tr>
</tbody>
</table>
Table 3. Classification scheme for the molecular functions of domains

<table>
<thead>
<tr>
<th>Functional class</th>
<th>Functional category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme/related function</td>
<td>Catalytic; substrate-binding; regulatory only; other enzyme function</td>
</tr>
<tr>
<td>Cofactor binding</td>
<td>Cofactor-binding; ion-binding; ion-cluster-binding; other molecule-binding function</td>
</tr>
<tr>
<td>Nucleic acid binding</td>
<td>DNA-binding; RNA-binding; other nucleic acid binding function</td>
</tr>
<tr>
<td>Protein binding</td>
<td>Permanent protein–protein interaction; transient protein–protein interaction; part of complex; other protein binding function</td>
</tr>
<tr>
<td>Lipid binding</td>
<td>Lipid binding; other lipid binding function</td>
</tr>
<tr>
<td>Carbohydrate binding</td>
<td>Carbohydrate binding; other carbohydrate binding function</td>
</tr>
<tr>
<td>Transport</td>
<td>Channel protein; electron transfer; other transport function</td>
</tr>
<tr>
<td>Other function</td>
<td>Structural role; general/multiple functions; viral proteins; toxin; other/ambiguous/unknown function</td>
</tr>
</tbody>
</table>

categories. Each category maps to one of the eight rough functional classes (Table 3). The annotation was based on information in SCOP\(^3\) and the literature.

**Controls for our functional classification**

**Automated domain annotation**

As a control, we used the automated annotation of GO process, function and location\(^4\) to Pfam\(^5\) domains in InterPro.\(^6\) Pfam domains were mapped onto SCOP domain superfamilies based on sequence similarity. This provided annotation for 647, 667 and 343 domain superfamilies, respectively. The manual domain annotation was largely consistent with the Gene-Ontology annotation for Pfam domains and their mappings to the domains described in SUPERFAMILY.\(^30\)

**Random annotation of function**

As a further control, all domain superfamilies were randomly assigned to 15 categories and the relationship between versatility and abundance tested as described above (data not shown). The relationship between abundance and versatility and the slope \(\theta\) were the same for all subsets.

**Random models**

For a simulation of the versatility expected from random shuffling, we considered each genome separately. The abundance of a particular domain superfamiliy and the number of proteins and their lengths in terms of domains were kept constant. The resulting versatility was recorded for each superfamily; and the average was taken after 10,000 repeats of the simulation.

S0. All domains were considered at their observed abundance. The domains were then picked and assigned to random positions in the proteins until each position was filled.

S1. All domains were considered at their observed abundance. Two different domain superfamilies were chosen at random to combine with each other. The combination was then duplicated until one of the component domains was “used up”. The process was repeated taking the new abundances for both superfamilies into account. Eventually, all domains were assigned to combinations.

**Acknowledgements**

We are grateful to Cyrus Chothia, M. Madan Babu, Shinhan Shiu and Christine Orengo for helpful discussion. C.V. acknowledges financial support by the Boehringer Ingelheim Fonds.

**References**


*Edited by G. von Heijne*