

## Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination

Gordana Apic<sup>1</sup>, Wolfgang Huber<sup>2</sup> & Sarah A. Teichmann<sup>1\*</sup>

<sup>1</sup>MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and <sup>2</sup>DKFZ (German Cancer Research Center), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; \* To whom correspondence should be addressed at [sat@mrc-lmb.cam.ac.uk](mailto:sat@mrc-lmb.cam.ac.uk)

Received 26 November 2002; Accepted in final form 11 March 2003

**Key words:** structural genomics, multi-domain proteins, recombination, domain combinations

### Abstract

There is a limited repertoire of domain families in nature that are duplicated and combined in different ways to form the set of proteins in a genome. Most proteins in both prokaryote and eukaryote genomes consist of two or more domains, and we show that the family size distribution of multi-domain protein families follows a power law like that of individual families. Most domain pairs occur in four to six different domain architectures: in isolation and in combinations with different partners. We showed previously that within the set of all pairwise domain combinations, most small and medium-sized families are observed in combination with one or two other families, while a few large families are very versatile and combine with many different partners. Though this may appear to be a stochastic pattern, in which large families have more combination partners by virtue of their size, we establish here that all the domain families with more than three members in genomes are duplicated more frequently than would be expected by chance considering their number of neighbouring domains. This duplication of domain pairs is statistically significant for between one and three quarters of all families with seven or more members. For the majority of pairwise domain combinations, there is no known three-dimensional structure of the two domains together, and we term these novel combinations. Novel domain combinations are interesting and important targets for structural elucidation, as the geometry and interaction between the domains will help understand the function and evolution of multi-domain proteins. Of particular interest are those combinations that occur in the largest number of multi-domain proteins, and several of these frequent novel combinations contain DNA-binding domains.

**Abbreviations:** SCOP: Structural Classification of Proteins database, PDB: Protein DataBank, HMM: hidden Markov model

### Introduction

Multi-domain proteins are the result of duplication and combination of domains. In nature, there is a limited number of domain families (Chothia 1992, Wolf et al. 2000), currently 1110 families of known structure (Murzin et al. 1995, LoConte et al. 2002). Though the domains from these families occur on their own in single-domain proteins, multi-domain proteins represent the majority of proteins in a proteome: two thirds of prokaryote proteins and eighty

percent of eukaryote proteins are multi-domain proteins (Teichmann et al. 1998, Gerstein 1998a).

The domains and evolutionary relationships of multi-domain proteins can be determined through structural annotation of genome sequences. The definitions of domains and protein families for the proteins of known three-dimensional structure are provided by the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995) at the superfamily level. Based on the same sequential domain architecture, we group multi-domain proteins into

families, and explore the characteristics of these families in eighty-five genomes from the three kingdoms.

The domain architectures of multi-domain proteins consist of sets of pairwise domain combinations from N- to C-terminus along a polypeptide chain. Previously, we showed that there is a limited repertoire of domain combinations (Apic et al. 2001) and that this repertoire is mainly formed by the combinatorial versatility of a few large families. This relationship between the size of a family and the number of different types of neighbouring domains in protein sequences might suggest a stochastic basis for domain combinations, rather than selection for particular domain architectures. Here we show that all families with more than three members have fewer domain combinations than expected according to a random model of domain shuffling, suggesting that selection pressure is exerted to maintain the same domain pairs in the evolution of domain combinations.

The function of each multi-domain protein is determined by its domain composition and, in most cases, their interactions. The protease function of chymotrypsin, for example, is carried out by the active site buried at the interface of two domains (Sigler et al. 1966, Blevins et al. 1985). Therefore, elucidating the structure of novel combinations is important in terms of the structural and functional characterization of multi-domain proteins as well as for elucidating the history of protein evolution (Ponting and Russell 2002). We analyze the pairwise combinations of unknown structure and present the most popular combinations as targets for structural genomics projects in the same way that individual new folds are viewed as targets (Brenner 2001; Blundell & Mizuguchi 2000).

### Identifying domains and families

The eighty-five genomes we study here are from the three kingdoms of life: the fifty-five Bacteria, seventeen Archaea and thirteen Eukarya given in Table 1a. The domain assignments were taken from the SUPERFAMILY database (Gough et al. 2001; Gough and Chothia, 2002), which provides matches between the domains of the Structural Classification of Proteins database (SCOP, LoConte et al. 2002) and predicted proteins of completely sequenced genomes. The method used in the SUPERFAMILY database is the iterative hidden Markov model (HMM) method SAM-T99 (Karplus et al. 1998), and the assignments described here are based on SCOP version 1.61. The

procedure to define domain boundaries is described in Gough et al. (2001).

The eighty-five completely sequenced genomes studied here have 449,823 predicted protein sequences, and for more than half (241,831) of these there is at least one structural domain assignment (Table 1b) from 1041 domain families (SCOP superfamilies). In this work, we are interested in direct neighbours in a protein sequence, and so allow linkers between domains not more than thirty residues long, which is about the size of the smallest SCOP domain, but do not consider domains that are inserted into other domains. About one half of the sequences with structural assignments have complete assignments according to this criterion, as shown in Table 1b.

### Families of multi-domain proteins

We consider multi-domain proteins with the same sequential arrangement of assigned structural domains as being members of the same family. In other words, if structural domains A,B,C and D occur in two different sequential arrangements as A-B-C-D and as A-B-D-C, we would form two families containing all the proteins with each respective domain architecture. The assumption that almost all proteins with the same domain architecture have descended from a common ancestor is supported by an analysis of two-domain proteins of known structure (Bashton and Chothia, 2001). Though many of these two-domain proteins had little sequence similarity, conservation of details of the sequence and structure of linker regions showed that all domain pairs of the same type were related by duplications of the whole two-domain module. This is also the idea behind the Conserved Domain Architecture Retrieval Tool (Geer et al. 2002).

Using our definition of families of multi-domain proteins, 241,831 protein sequences with structural assignments cluster into 13,606 families of which 69 are single-domain families. The size distribution of multi-domain protein families follows a power law (Figure 1) with a similar exponent to the power law of the size distribution of individual domain families (Qian et al., 2001). There is an additional pattern within the family size distribution in that longer domain architectures tend to be less duplicated (data not shown). In other words, longer multi-domain protein families tend to have fewer members. This means that shorter domain combinations are re-used more

Table 1a. 85 completely sequenced genomes used.

| Bacteria                                  |   |   | Archaea   | Eukarya                                  |
|---|---|---|---|--|
| <i>Agrobacterium tumefaciens</i> C58      | <i>Haemophilus influenzae</i>           | <i>Rickettsia prowazekii</i>            | <i>Aeropyrum pernix</i>                           | <i>Anopheles gambiae</i>                 |
| <i>Aquifex aeolicus</i>                   | <i>Helicobacter pylori</i> 26695        | <i>Salmonella typhimurium</i> LT2       | <i>Archaeoglobus fulgidus</i>                     | <i>Arabidopsis thaliana</i>              |
| <i>Bacillus halodurans</i>                | <i>Lactococcus lactis</i>               | <i>Shewanella oneidensis</i> MR-1       | <i>Halobacterium</i> sp. NRC-1                    | <i>Caenorhabditis elegans</i> rel. WS93b |
| <i>Bacillus subtilis</i>                  | <i>Listeria innocua</i>                 | <i>Sinorhizobium meliloti</i>           | <i>Methanobacterium thermoautotrophicum</i>       | <i>Ciona intestinalis</i> 1.0            |
| <i>Borrelia burgdorferi</i>               | <i>Listeria monocytogenes</i>           | <i>Staphylococcus aureus</i> Mu50       | <i>Methanococcus jamaashii</i>                    | <i>Drosophila melanogaster</i> rel. 3    |
| <i>Brucella melitensis</i>                | <i>Mesorhizobium loti</i>               | <i>Streptococcus pneumoniae</i> R6      | <i>Methanopyrus kandleri</i> AV19                 | <i>Encephalitozoon cuniculi</i>          |
| <i>Buchnera</i> sp.                       | <i>Mycobacterium leprae</i>             | <i>Streptococcus pyogenes</i>           | <i>Methanosarcina acetivorans</i> C2A             | <i>Fugu rubripes</i> 8.1                 |
| <i>Campylobacter jejuni</i>               | <i>Mycobacterium tuberculosis</i> H37Rv | <i>Streptomyces coelicolor</i>          | <i>Methanosarcina mazei</i> Go1                   | <i>Homo sapiens</i> 8.30                 |
| <i>Caulobacter crescentus</i>             | <i>Mycoplasma genitalium</i>            | <i>Synechocystis</i> sp. PCC 6803       | <i>Methanothermobacter thermautotrophicus</i> DLH | <i>Mus musculus</i>                      |
| <i>Chlamydia muridarum</i>                | <i>Mycoplasma pneumoniae</i>            | <i>Thermoanaerobacter tengcongensis</i> | <i>Pyrobaculum aerophilum</i>                     | <i>Oryza sativa</i> ssp. indica          |
| <i>Chlamydia trachomatis</i>              | <i>Mycoplasma pulmonis</i>              | <i>Thermotoga maritima</i>              | <i>Pyrococcus abyssi</i>                          | <i>Plasmodium falciparum</i>             |
| <i>Chlamydophila pneumoniae</i> J138      | <i>Neisseria meningitidis</i> MC58      | <i>Treponema pallidum</i>               | <i>Pyrococcus furiosus</i> DSM 3638               | <i>Saccharomyces cerevisiae</i>          |
| <i>Chlorobium tepidum</i> TLS             | <i>Nostoc</i> sp. PCC 7120              | <i>Ureaplasma urealyticum</i>           | <i>Pyrococcus horikoshii</i>                      | <i>Schizosaccharomyces pombe</i>         |
| <i>Clostridium acetobutylicum</i>         | <i>Pasteurella multocida</i>            | <i>Vibrio cholerae</i>                  | <i>Sulfolobus solfataricus</i>                    |  |
| <i>Clostridium perfringens</i>            | <i>Pseudomonas aeruginosa</i>           | <i>Xanthomonas campestris</i>           | <i>Sulfolobus tokodaii</i>                        |  |
| <i>Corynebacterium glutamicum</i>         | <i>Ralstonia solanacearum</i>           | <i>Xanthomonas citri</i>                | <i>Thermoplasma acidophilum</i>                   |  |
| <i>Deinococcus radiodurans</i>            | <i>Rhodobacter sphaeroides</i>          | <i>Xylella fastidiosa</i>               | <i>Thermoplasma volcanium</i>                     |  |
| <i>Escherichia coli</i> K12               | <i>Rickettsia conorii</i>               | <i>Yersinia pestis</i>                  |   |  |
| <i>Fusobacterium nucleatum</i> ATCC 25586 |   |   |   |  |

Table 1b. Structural assignments to 85 genomes: proteins and multi-domain protein families

| Phylogenetic group | Number of proteins |                    |                  | Number of multi-domain protein families |                    |                  |
|--------------------|--------------------|--------------------|------------------|---|--------------------|------------------|
|                    | Total              | Incomplete matches | Complete matches | Total                                   | Incomplete matches | Complete matches |
| Archaea            | 39497              | 6432               | 14342            | 1859                                    | 1104               | 749              |
| Bacteria           | 153190             | 27871              | 58002            | 4302                                    | 2960               | 1342             |
| Eukarya            | 257136             | 68838              | 66346            | 10489                                   | 8028               | 2461             |
| All 85 genomes     | 449823             | 103141             | 138690           | 13606                                   | 10407              | 3199             |

The total number of proteins, and then those with structural assignments is given for each phylogenetic group. Structural assignments of domains can completely cover a protein, or partially match a protein and leave thirty residues or more of unassigned region. The multi-domain families are clusters of proteins grouped according to domain architecture, in other words the sequence of domains and, in some cases, unassigned regions.

often by organisms, perhaps because their function can be modified more easily than that of longer domain combinations.

### Combinatorial versatility of domain families in multi-domain proteins

The components of multi-domain families are individual domain families, and we are interested in the combinatorial versatility of these individual families in the context of direct domain neighbours and also entire domain architectures. It has been shown that

the versatility of domain families in terms of neighbouring domains follows a power law distribution: there are a small number of large families with many types of neighbours, and a large number of families with few types of neighbours (Apic et al. 2001; Wuchty, 2001; Kuznetsov, 2002). The same is true for the number of domain architectures in which a particular family occurs (data not shown). The versatility of a domain family in terms of pairwise N-to-C terminal combination partners correlates well with the versatility in terms of the number of multi-domain protein families. Therefore, in general, a family with a large number of pairwise combination partners will

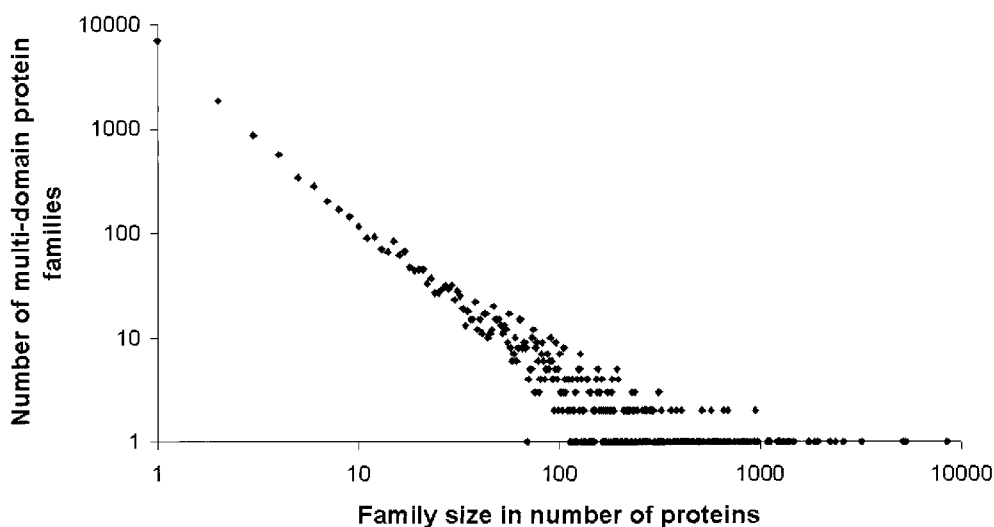


Figure 1. Multi-domain family size distribution

The number of multi-domain protein families of different sizes in terms of numbers of proteins is plotted on double-logarithmic axes. The x-axis is the family size in number of proteins, and the y-axis is the number of multi-domain families of each size. Small multi-domain families, with only one or two protein representatives in the eighty-five genomes, are the majority of families (65%). Only a very small proportion of protein contexts are actually large families with many protein representatives: 193 families (1.4% of families) have between 200 and 8000 family members.

also be part of many multi-domain protein families. On average, each pairwise domain combination is a member of about 4 and 6 different multi-domain families in prokaryotes and eukaryotes respectively.

### Pairwise domain combinations: random shuffling or selection?

Above, we show that multi-domain proteins are composed of pairwise domain combinations that mostly occur in three to six domain architectures. In our set of multi-domain proteins, there are 2545 different pairwise domain combinations in total, composed of 796 different domain families. We define pairwise domain combinations as domain assignments less than thirty residues apart, and increasing or decreasing the allowed gap between domains by 10 or 20 residues changes the number of novel combinations by 10 to 20 percent.

Given the roughly 634 000 ( $\approx 796^2$ ) different combinations that are possible in principle out of all families in the combinations, the number of observed combinations is a small fraction of all possible combinations. This suggests that there may be selection for particular domain combinations. However, the tendency for larger families to have more domain neighbours may point towards a random pattern of domain shuffling with duplication as suggested by Koonin et al. (2002).

To support or refute the suggestion that domain combinations follow the pattern expected stochastically given the set of domains and family sizes, we propose a random model for the formation of pairwise combinations. According to this model, a combination occurs with a probability that only depends on the number of occurrences of the two constituting families in domain pairs, and the total number of domain pairs in the genome. This means that each of the domains belonging to a given family can combine with any of the entire set of domains available in a genome, without a bias of any kind, for instance towards duplication of existing combinations. In particular, the occurrence of a combination is independent of the occurrence of other combinations or of the identity or function of the constituting families. In this respect, our model is analogous to the random graph model by Erdős and Rényi (1961).

### Random shuffling model of domain combinations

In our model, we calculated  $m$ , the total number of pairs, and the relative frequency of occurrence  $p_i$  for each domain family  $i$  for each genome. Under the random combination assumption, the probability that a given pair of domains consists of  $i$  in the first position and  $j$  in the second is  $p_i p_j$ . Thus, the probability for the ordered combination  $(i, j)$  occurring at least once in the whole genome of  $m$  pairs is

$$\pi_{ij} = 1 - (1 - p_i p_j)^m$$

Random domain combination data  $d_{ij}$  was drawn from the numbers 0 and 1 with

$$P(d_{ij}=1) = \pi_{ij}$$

Here,  $d_{ij}$  is an indicator which is 1 if combination  $(i, j)$  is present and 0 if not. The number of combination partners of domain  $i$  is

$$n_i = \sum_j d_{ij} + d_{ji} - d_{ii}$$

with expectation value

$$E(n_i) = 2 \sum_j \pi_{ij} - \pi_{ii}.$$

The distribution of  $n_i$  was calculated from 10,000 samples of the  $d_{ij}$  data matrix. Comparison of the distribution with the actual number of combination partners of domain  $i$  yielded a simulation  $p$ -value for each superfamily in each genome.  $p$ -values per kingdom were calculated as the median over the individual genomes from the kingdom. We do not interpret the  $p$ -values formally in the sense of statistical hypothesis testing, which would require addressing the multiple testing problem. Instead, we simply use them as scores to rank and select the domain families that are furthest from the random model, using a  $p$ -value of 0.01 as a cut-off.

### Observed domain combinations compared to random model

With this model, we were able to compare the observed numbers of combinations of families to their null distribution generated by the random model. Figure 2 shows a plot of the number of expected combinations against the number of observed combinations of each family averaged across the genomes

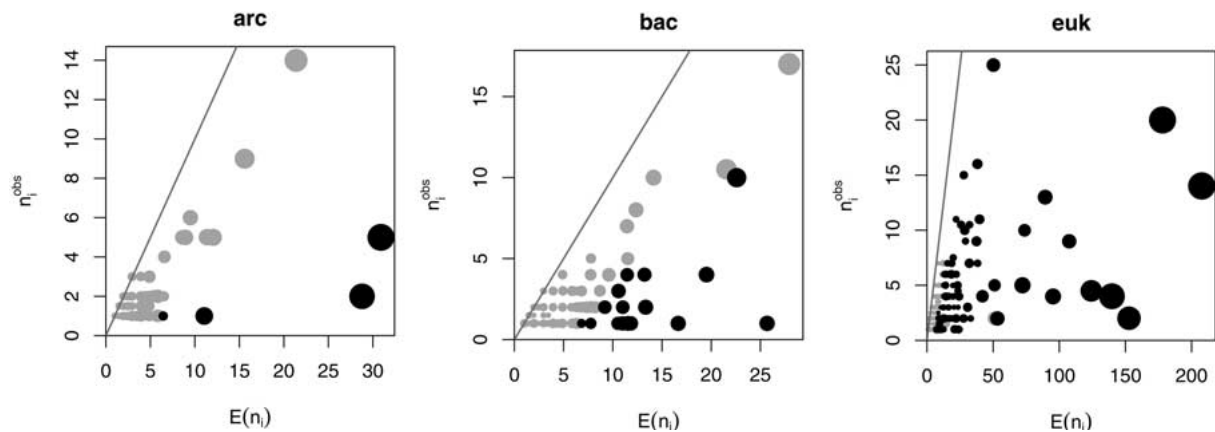


Figure 2. Number of expected combinations according to random model versus number of observed combinations.

Each dot represents one domain family, and the size of the dot is related to the size of the family in average number of domains per genome in that kingdom. Grey dots are families that cannot be distinguished from the random model. Off-diagonal families differ significantly from what would be expected from random combination, and are represented as black dots. The first, second and third panels represent the archaeal, bacterial and eukaryote genomes respectively. In all three phylogenetic groups, the families with more than three members are all below the diagonal, having duplicated more often than expected given the number of types of combination partner. Only families of size seven or larger have duplicated as domain pairs so frequently that they are statistically significant compared to the random model (black dots).

in each of the three kingdoms. Off-diagonal families are those that are different from what would be expected from random combinations. As is obvious from the three graphs in Figure 2, all families have either the same or fewer types of combination partner than expected given the size of the family. As shown in Table 2, between one and two thirds of the families have the same number of combination partners as expected. The families with the same number of combinations as expected are all small: they are all families with just one domain in the prokaryotes, and one to three domains in eukaryotes.

All families with more than one domain in the prokaryotes and more than three domains in the eukaryotes have fewer combinations than expected.

This means that these families have many duplicates of a limited number of domain pairs. This trend is statistically significant only in families with seven or more members, as shown in Table 2. Of these larger families, between one and roughly three quarters of the families deviate from the random model at a statistically significant level across the three kingdoms. These calculations show that a limited repertoire of domain combinations is duplicated and re-used in proteins. Domains are more often duplicated and selected as pairs or series of domains than as new combinations of domains. Furthermore, there is a large scatter of observed combinations for each number of expected combinations. Since the number of expected combinations is a direct reflection of the number of domains in a family, this shows that there is no simple relationship between domain duplication (the size of a family, or the number of expected com-

Table 2. Numbers of families with domain combination types different from the random model.

| Phylogenetic group | No. of families with the same / fewer combinations than expected by chance* | No. of families with significant p-values out of all families | No. families with significant p-values out of all families size 7 or greater | Percentage of families size 7+ with significant p-values |
|--------------------|---|---|--|--|
| Archaea            | 221/135   | 4/356   | 4/15   | 27   |
| Bacteria           | 336/203   | 19/539  | 19/37  | 51   |
| Eukarya            | 245/444   | 118/689   | 118/164  | 72   |

\*All the families with the same number of combinations consist of a single domain in the two prokaryote groups, and of one to three domains in the eukaryote group. This means that all larger families have fewer combination types than expected by chance.



bbinations in Figure 2) and domain recombination (the number of observed combinations in Figure 2).

### Novel domain combinations versus PDB combinations

The limited repertoire of 2545 pairwise domain combinations is not the product of random domain shuffling as proposed in our model above. The domains in these 2545 combinations are building blocks for multi-domain proteins belonging to 796 domain families of known structure. In order to understand the function of multi-domain proteins, knowledge of the geometry, the three-dimensional orientation of the set of domains to each other, is needed as much as of the three-dimensional structure of individual domains.

Recent work on the conservation of geometry in sets of two-domain proteins (Bashton and Chothia, 2002) as well as on families of complexes (Aloy and Russell, 2002) suggests that the geometry of domain combinations is conserved in a manner analogous to the conservation of the structure of individual domains with divergent sequences. The conservation of geometry is probably linked to the conservation of function of domain combinations, as Hegyi and Gerstein (2001) showed that conservation of function is greater in multi-domain proteins than for individual domains. Assuming conservation of geometry holds, the geometry of a domain combination can be inferred, if the domain combination is the same as that in a protein of known structure. This approach has been used in recent efforts in modeling large complexes (Aloy et al. 2002, Spahn et al. 2002).

The pairwise combinations of domains where the three-dimensional structure of individual domains is known, but the relative orientation in three-dimensional space of the domains forming a pair is still unknown, are called *novel* domain combinations here. In order to gain a better understanding of the function of multi-domain proteins, as well as insight into the evolution of domains in the context of the contacts they make to other domains in three dimensions, these novel combinations would be good targets for structure elucidation.

In the eighty-five completely sequenced genomes in this study, there are 2063 novel domain combinations, and we show below that these contain 1949 combinations that are reliably predicted. We term combinations of domains with known relative three-dimensional orientation *PDB* (or *known*) combina-

tions. There are 482 PDB combinations that are present in the genomes as well.

### Classification of novel domain combinations

The 2063 novel combinations can be classified into three categories, as illustrated in Figure 3b: (i) reversals of an N-to-C sequential order of a known combination (for domains A and B, if AB is of known structure, BA would be a reversal); (ii) new internal duplications (if a structural domain is present as a single domain in the structure and/or in combination with a domain from a different family, then a tandem repeat of the same family will be an internal duplication) and (iii) new combinations (a combination of two structural domains, which are alone and/or in combination with other domain families amongst the proteins of known structure, but not combined with each other).

Only about two percent of the novel combinations are cases of reversed N-to-C terminal orientations as compared to the proteins of known structure. In gen-

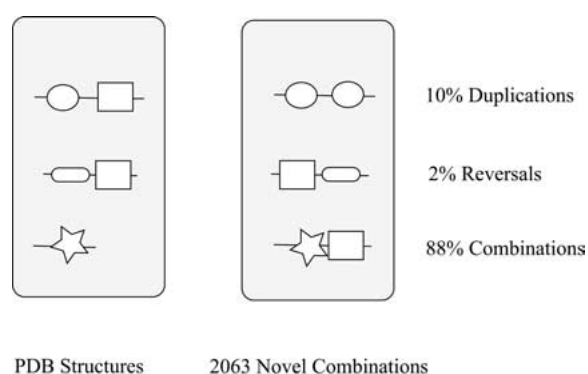


Figure 3. Types of novel combinations

This cartoon represents three different types of novel combinations: duplications, reversals and new combinations. Structural domains from the same family are presented as one shape, e.g. circle and labeled A, B, C. In this simplified example, given on the left is a complete set of combinations in the PDB, the “old” combinations. On the right, there is a complete set of “novel” combinations in the eighty-five completely sequenced genomes. For example, domains A, B and C are combined with each other in pairwise combinations AB and CB in the PDB. Novel combinations can be of different types: **duplications** of domains that are present in the PDB as a single domain or in combination with a domain from a different family, such as AA; **reversals** of PDB combinations – for example CB would be a reversed N-to-C-terminal orientation of the combination BC; or **combinations** where two domain families that are in combination with other partners or in isolation in the PDB combine together.

Table 3a. Novel combinations in genomes, which are reversals of N-to-C terminal orientation of a combination present in the PDB.

| Novel Reversals, YX in genomes    |  | King-<br>dom | Representative PDB<br>chain identifier for<br>XY combination |
|-----------------------------------|--|--------------|--|
| Domain Y                          | Domain X   |              |  |
| “Winged helix” DNA-binding domain | Periplasmic binding protein-like II                  | A,B,E        | 1hsjA  |
| Fibronectin type III              | Immunoglobulin                                       | E            | 1f42A  |
| EGF/Laminin                       | LDL receptor YWTD domain                             | E            | 1ijqA  |
| Transketolase C-terminal domain   | Thiamin diphosphate –binding fold                    | A,B,E        | 1tka   |
| EGF/Laminin                       | C-type lectin  | E            | 1esl   |
| Immunoglobulin                    | Caspase  | E            | 1cvrA  |
| GFP-like domain                   | EGF/Laminin  | E            | 1gl4A  |
| Immunoglobulin                    | Protein kinase                                       | E            | 1koa   |
| “Winged helix” DNA-binding domain | P-loop containing nucleotide triphosphate hydrolases | B            | 1in4A  |
| Metallo-dependent hydrolases      | Composite domain of metallo-dependent hydrolases     | E            | 1ejrC  |

There are 50 novel reversal combinations, and the ten most common are given here, which are present in 8-915 copies in the eighty-five genomes. If a combination in PDB is XY, the novel combination in genomes will be YX. The combinations can be present in proteins from one or more of the three phylogenetic kingdoms: Archaea (A), Bacteria (B) or Eukaryotes (E).

eral, within one genome, combinations that are seen in both sequential orientations represent one to four percent of all combinations, while reversals across genomes are only marginally more frequent (three to ten percent, Apic et al. 2001). In Table 3a, we show the ten most frequent combinations of the fifty that are reversals of combinations of known structure. The structures of these combinations are of interest from an evolutionary point of view. As for the eight domain pairs studied by Bashton & Chothia (2002), we expect these reversed domain pairs to have different geometrical and functional relationships in each N-to-C terminal orientations in almost all cases.

Ten percent of novel combinations are duplications of a family that is seen only in isolation and/or in combination with a family other than itself in the ProteinDataBank (Berman et al. 2002). The vast majority of the novel combinations, 88%, are combinations of domain families which are combined with other domain families or which are present as single domains in the PDB, combined in a new way. The balance between duplications, reversals and new combinations hardly changes when novel combinations with poor scores that are unreliable are removed, as described below.

### Comparison of PDB and novel combinations

In order to estimate the quality of prediction of novel combinations, we compare them to the assignments

of the PDB combinations. The structures of the PDB combinations are known, so we compare the predicted, novel combinations to the known combinations with respect to the following two criteria: (i) score distributions of matched combinations and (ii) the frequency of combinations amongst the assignments to the eighty-five genomes.

### Score distributions

A structural domain, represented by a hidden Markov model in the SUPERFAMILY database (Gough et al. 2002) matches a target protein sequence with a score or expectation value (E-value). The E-value threshold for accepting matches is calibrated to a 1% error rate (Gough et al. 2001). The error rate at an E-value threshold of  $10^{-5}$  is about 0.1%. We consider the matches in the region of scores from below  $10^{-5}$  to be poor scores, while scores better than  $10^{-5}$  are regarded as “good” scores.

In order to compare the known and novel combinations, we have plotted a histogram of two scores for each pairwise domain combination shown in Figure 4. Overall, the distributions for novel and known combinations look very similar. The region of poor scores, where both domains in a combination are matched with a score that is worse than  $10^{-5}$ , accounts for 0.8 percent of all hits for novel combinations and none of the *known* combinations. There are very few hits with one domain matched with a very high score



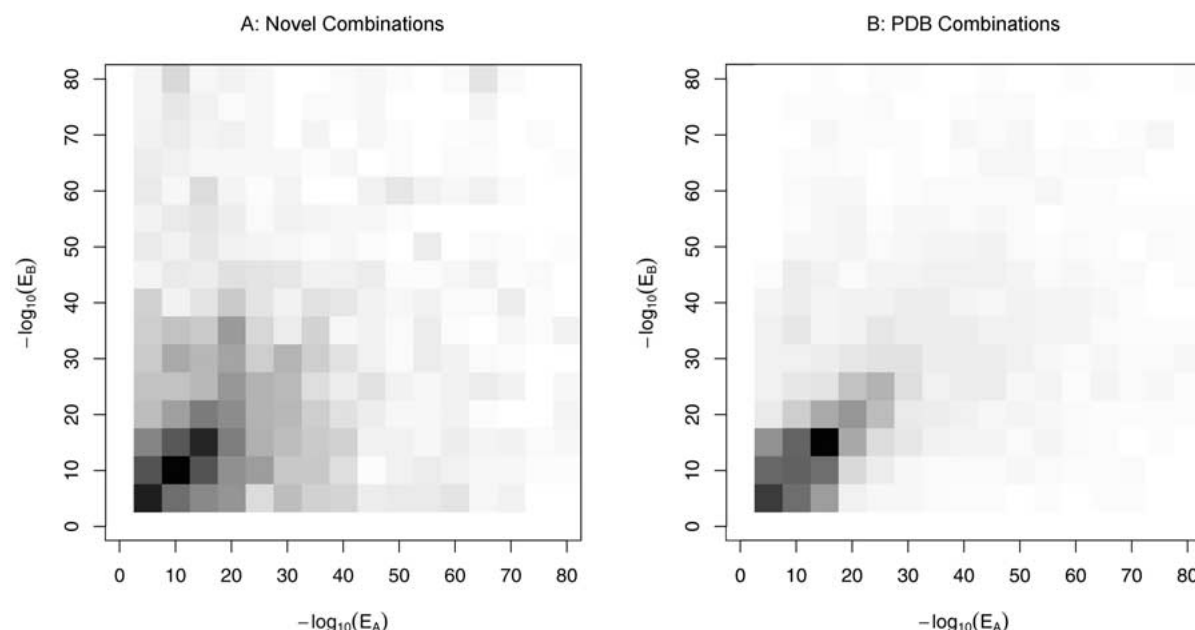


Figure 4. Distribution of scores for matched pairs of domains

The two axes are the negative logarithms of the expectation value scores for the N- and C-terminal (A and B respectively) domains of a pairwise domain combination. The darkness of the shading of a square represents the number of domain combinations with scores within the square. Panel A shows the score distribution for novel combinations, panel B for PDB combinations.

and the other domain matched with a poor score for both known and novel combinations.

### Singletons in the region of poor scores

To inspect the most likely candidates for false positive combinations, we now study the poor matches in more detail. If one domain pair occurs in only one sequence, we call this match a singleton. In cases where there are multiple hits for a particular domain combination, it is more likely that this is a true combination, even if the scores are poor.

The first column next to the origin of the histogram of novel combinations in Figure 4 represents hits with poor scores for both domains. For novel combinations there are 127 poor hits. Of these, 114 combinations are present in only one sequence among the eighty-five genomes studied here, and 9 are present in two copies. Only four combinations with poor hits are present in three to five sequences in the eighty-five genomes. Thus assuming that poor single hits of a domain combination may be false positives, we can only be confident about 1949 novel combinations instead of 2063.

### Frequency of novel combinations

Apart from the scores of SCOP domains matching the target protein sequence for one combination, we can use the frequencies to estimate the quality of predictions for novel combinations. Known combinations are expected to have more matches, because the homologues in the genomes are likely to be direct duplicates of known combinations and thus have better scores. However, the distribution of frequencies for novel and PDB combinations are similar in character: there are few combinations that match two to three hundred different proteins, and the majority is present in one or two copies. Out of 2063 novel combinations, 1047 different combinations are present in only one copy in the genomes, 245 in two copies, and the ten most frequent combinations occur roughly 200 to 900 times, as shown in Table 3b. For the PDB combinations, only about 23 different combinations have a single homologue in genomes, and the ten most frequent PDB combinations are present in roughly 800 to 14,000 genomic sequences.

Table 3b. The most frequent domain combinations.

| PDB ("known")                                |   | Novel   |  |
|--|---|---|--|
| Internal duplications                        | Combinations  | Internal duplications                         | Combinations   |
| C2C2 and C2HC fingers (14250)                | Homodimeric domain, signal transducing histidine kinase & ATPase domain of HSP90 chaperone (923)        | CBS domain (614)                              | "Winged helix" DNA-binding domain & Periplasmic binding protein-like II (915)          |
| EGF/Laminin (3530)                           | Thioredoxin-like & Glutathione S-transferases, C-term (616)   | PYP-like sensor domain (409)                  | Actin-like ATPase domain & Heat shock protein (HSP70), substrate binding domain (237)  |
| Immunoglobulins (3262)                       | Biotin carboxylase N-terminal domain & Glutathione synthetase ATP binding domain (565)                  | Spermadhesin, CUB domain (318)                | PYP-like sensor domain & Homodimeric domain, signal transducing histidine kinase (234) |
| Cadherins (2714)                             | NAD(P)-binding Rossmann fold domain & Glyceraldehyde-3-P dehydrogenase (516)                            | Tachycitin (241)                              | PLP-binding barrel & Alanine racemase like, C-term dom. (209)                          |
| Fibronectin Type III (1907)                  | NAD(P)-binding Rossmann fold domain & 6-phosphogluconate dehydrogenase C-terminal domain (510)          | Scavenger receptor cysteine rich domain (218) | FYV E/PHD zinc finger & C2C2 and C2HC fingers (179)                                    |
| Complement control module/ SCR domain (1049) | P-loop hydrolase & Translation proteins (463)   | TPR-like (204)                                | ATPase domain of HSP90 chaperone & CheY like (176)                                     |
| Actin-like ATPase (952)                      | Immunoglobulin & Fibronectin type III (449)   | SH3 domain (197)                              | Fibronectin Type III & Immunoglobulin (174)  |
| Spectrin repeat (910)                        | Acyl-CoA dehydrogenase, N-terminal and middle domains & Acyl-CoA dehydrogenase, C-terminal domain (438) | Membrane all alpha (164)                      | C2C2 and C2HC fingers & FYV E/PHD zinc finger (153)                                    |
| RNA-binding domain (787)                     | Thiamin diphosphate binding fold & Transketolase C-terminal domain (339)                                | WW domain (135)                               | TRAF-domain & POZ domain (148)   |
| Homeodomain like (772)                       | Riboflavin synthase domain -like & Ferredoxin reductase-like, C-terminal NADP-linked domain (381)       | PKD domain (123)                              | "Winged helix" DNA-binding domain & Phosphosugar isomerase (130)                       |

The most frequent pairs of domain families are listed in terms of those also in PDB chains and those only in genome assignments. We distinguish two types of domain pairs: i) internal duplications, where the two domains forming a pair are from the same family and ii) combinations of different domain families. The occurrence of domain combinations in the eighty-five genomes is given in brackets. Combinations are given in N-to-C terminal orientation from top to bottom.

### The most frequent novel combinations

The most frequent novel combinations are of particular interest, as elucidating their structure and function provides information about large numbers of multi-domain protein families. The most frequent novel and known combinations are given in Table 3b as internal duplications of one domain family and combinations of two domains from different families. Several of the most abundant known internal duplications are extracellular and involved in signaling, such as Immuno-

globulin, Fibronectin type III or EGF/laminin domains. Others are intracellular catalytic domains such as the Actin-like ATPase family. The most frequent novel internal duplications from the same family are mainly regulatory signaling domains or ligand binding domains.

The most frequent PDB combinations from two different families are mostly two domains forming a small molecule enzymatic function. The bias towards the small molecule metabolism combinations may be due to the biochemical interest in this system: around

one half of all *E.coli* protein sequences have homologous structures, but there is a structural assignment for more than 80% of proteins involved in *E. coli* small molecule metabolism (Teichmann et al. 2001, Jardine et al. 2002).

Amongst the novel combinations, four pairs of domains are involved in transcriptional regulation, and three of these comprise a DNA binding domain and a signaling domain. The most frequent novel combination, with more than nine hundred copies in the eighty-five genomes is present in bacteria and archaea is the "Winged helix" DNA-binding domain and periplasmic binding protein-like II domain. These frequent novel combinations can be viewed as targets for structural genomics projects, as elucidation of the structure of the combined domain pair would lead to a better understanding of many multi-domain proteins in important processes such as transcription regulation.

## Conclusions

The majority of proteins in prokaryote and eukaryote proteomes consist of two or more domains, so understanding the structure and function of domain combinations is vital to our knowledge of the protein repertoire. Using structural assignments to genomes sequences, we identified 13,606 different domain architectures amongst about half the proteins in eighty-five genomes. Since there is evidence that proteins with the same domain architecture are frequently related (Bashton and Chothia, 2002, Hegyi and Gerstein, 2001), we determined the family size distributions of multi-domain proteins based on the domain architectures. The family sizes are distributed approximately according to a power law, like individual domain families. Considering the domain pairs that the multi-domain proteins are composed of, domain pairs occur in four to six domain architectures on average.

A limited repertoire of domain pairs occurs in multi-domain proteins. Previous work showed that larger domain families can be more versatile in the types of domain neighbours they combine with, which suggests a stochastic pattern of domain shuffling. We show here that the domain combinations can be distinguished from a random model of domain recombination, because domain pairs are duplicated more frequently than expected by chance. This is true of all families with more than three domains, and

means that domains are duplicated and selected as pairs or series of domains. Furthermore, there is a wide variation in the number of combination partners with family size, as implied by Figure 2. Thus there is no simple relationship between family size and number of neighbouring families.

The domain combinations that are not present in the PDB, novel combinations, are interesting targets for structure elucidation, because although the structure of the individual domains is known, the geometry and interaction of the domains is unclear. This is important in order to understand both the function and evolution of domains in the context of their neighbours in multi-domain proteins. There are 2063 novel pairwise domain combinations in our set of structural assignments to eighty-five completely sequenced genomes. For 1949 novel combinations it is clear that they are reliable predictions. The most frequent novel combinations represent a large number of multi-domain proteins, and several of these are DNA-binding domains in combination with regulatory domains.

## Acknowledgements

We are grateful to Cyrus Chothia, Ewan Birney and Andre Sali for helpful questions and discussions, and Julian Gough and Martin Madera for the SUPER-FAMILY database structural assignments.

## References

1. Aloy, P. and Russell, R. B. (2002) *Proc. Natl. Acad. Sci. USA.*, **99**, 5896-5901.
2. Aloy P., Ciccarelli F. D., Leutwein C., Gavin A. C., Superti-Furga, G., Bork, P., Bottcher B. and Russell, R.B. (2002) *EMBO Rep.*, **7**, 628-635.
3. Apic, G., Gough, J. and Teichmann, S.A. (2001) *J. Mol. Biol.*, **310**, 311-325.
4. Bashton, M. and Chothia, C. (2002) *J. Mol. Biol.*, **315**, 927-939.
5. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C. (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899-907.
6. Blevins, R.A. and Tulinsky, A. (1985) *J. Biol. Chem.*, **260**, 4264-4268.
7. Blundell, T.L. and Mizuguchi, K. (2000) *Prog. Biophys. Mol. Biol.*, **73**, 289-295.
8. Brenner, S.E. (2001) *Nat. Rev. Genet.*, **2**, 801-809.

9. Chothia, C. (1992) *Nature*, **357**, 543-544.
10. Erdős, P. and Rényi, A. (1960) *Magyar Tud. Akad. Mat. Kutató Int. Kozl.* **5**, 17-61.
11. Geer, L.Y., Domrachev, M., Lipman D. J., Bryant, S. H. (2002) *Genome Res.*, **12**, 1619-1623
12. Gerstein, M. (1998a). *Folding & Design*, **3**, 497-512.
13. Gerstein, M. (1998b) *Proteins*, **33**, 518-534.
14. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) *J. Mol. Biol.*, **313**, 903-919.
15. Gough, J. and Chothia, C. (2002) *Nucleic Acids Res.*, **30**, 268-272.
16. Hegyi, H. and Gerstein, M. (2001) *Genome Res.*, **11**, 1632-40.
17. Jardine, O., Gough, J., Chothia, C. and Teichmann, S.A. (2002) *Genome Res.*, **12**, 916-929.
18. Karplus, K., Barrett, C. and Hughey, R. (1998) *Bioinformatics*, **14**, 846-56.
19. Koonin, E. V., Wolf, Y. I., and Karev, P. (2002) *Nature*, **420**, 218-223.
20. Liu, J. and Rost, B. (2001) *Protein Sci.*, **10**, 1970-1979.
21. LoConte, L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2002) *Nucleic Acids Res.*, **30**, 264-7.
22. Kuznetsov, V.A., Pickalov, V.V., Senko, O.V. and Knott, G.D. (2002) *J. Biol. Systems* **10**, 381-407.
23. Murzin, A., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536-540.
24. Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994) *Nature*, **372**, 631-634.
25. Ponting, C. P. and Russell, R. R. (2002) *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 45-71.
26. Qian, J., Luscombe, N.M. and Gerstein, M. (2001) *J. Mol. Biol.*, **313**, 673-681.
27. Sigler, P. B., Jeffery, B.A., Matthews, B.W. and Blow, D. M. (1966) *J. Mol. Biol.*, **15**, 175-192.
28. Spahn, C. M., Beckmann, R., Eswar, N., Penczek, P. A., Sali, A., Blobel, G. and Frank, J. (2002) *Cell*, **107**, 373-386.
29. Teichmann, S. A., Park, J. and Chothia, C. (1998) *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14658-14663.
30. Teichmann, S. A., Chothia, C. and Gerstein, M. (1999) *Curr. Op. Struc. Biol.*, **9**, 390-399.
31. Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. and Chothia, C. (2001) *Trends Biotechnol.*, **19**, 482-486.
32. Teichmann, S. A., Rison, S. C., Thornton, J.M., Riley, M., Gough, J. and Chothia, C. (2001) *J. Mol. Biol.*, **311**, 693-708.
33. Wolf, Y. I., Grishin, N. V. and Koonin, E. V. (2000) *J. Mol. Biol.* **299**, 897-905.
34. Wuchty, S. (2001) *Mol. Biol. Evol.* **18**, 1715-1723.