doi:10.1006/jmbi.2001.4776 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **310**, 311–325

JMB

Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes

Gordana Apic[1]*, Julian Gough[1] and Sarah A. Teichmann[2]

[1]*MRC, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[2]*Department of Biochemistry & Molecular Biology, University College London, Darwin Bldg., Gower Street London WC1E 6BT, UK*

**Corresponding author*

There is a limited repertoire of domain families that are duplicated and combined in different ways to form the set of proteins in a genome. Proteins are gene products, and at the level of genes, duplication, recombination, fusion and fission are the processes that produce new genes. We attempt to gain an overview of these processes by studying the evolutionary units in proteins, domains, in the protein sequences of 40 genomes. The domain and superfamily definitions in the Structural Classification of Proteins Database are used, so that we can view all pairs of adjacent domains in genome sequences in terms of their superfamily combinations. We find 783 out of the 859 superfamilies in SCOP in these genomes, and the 783 families occur in 1307 pairwise combinations. Most families are observed in combination with one or two other families, while a few families are very versatile in their combinatorial behaviour; 209 families do not make combinations with other families. This type of pattern can be described as a scale-free network. We also study the N to C-terminal orientation of domain pairs and domain repeats. The phylogenetic distribution of domain combinations is surveyed, to establish the extent of common and kingdom-specific combinations. Of the kingdom-specific combinations, significantly more combinations consist of families present in all three kingdoms than of families present in one or two kingdoms. Hence, we are led to conclude that recombination between common families, as compared to the invention of new families and recombination among these, has also been a major contribution to the evolution of kingdom-specific and species-specific functions in organisms in all three kingdoms. Finally, we compare the set of the domain combinations in the genomes to those in the RCSB Protein Data Bank, and discuss the implications for structural genomics.

Keywords: domain combinations; gene duplication; recombination; structural genomics

Introduction

Domains are units of evolution[1–3] and all proteins consist of one or more domains, with the exception of some disordered proteins. There is probably a limited repertoire of domain families,[4,5] so that we can study gene duplication and recombination by investigating the domain combinations in the proteins of completely sequenced genomes.[6,7]

We use structural assignments to genome sequences, because proteins of known three-dimensional structure have clearer domain definitions and evolutionary family relationships than sequences of unknown structure. The domain and superfamily definitions in the Structural Classification of Proteins (SCOP)[1] are used to establish the pairs of superfamilies that are adjacent to each other in the proteins of the 40 genomes. This allows us to survey and compare the set of domain family combinations present in the 40 archaeal, bacterial and eukaryote genomes listed in Table 1. We study the phylogenetic distribution of pairwise domain combinations, the N to C-terminal orientation of domain pairs and domain repeats. We compare the set of the domain combinations in the genomes to those in the RCSB Protein Data Bank

Abbreviations used: POB, Protein Data Bank; SCOP, Structural Classification of Proteins; HMM, Hidden Markov Model.

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

0022-2836/01/020311–15 $35.00/0

Table 1. Forty genomes and their structural assignments

Bacteria	Archaea	Eukarya
A. *Forty genomes in three kingdoms*		
Aquifex aeolicus	*Archaeoglobus fulgidus*	*Arabidopsis thaliana*
Bacillus halodurans	*Aeropyrum pernix*	*Caenorhabditis elegans*
Bacillus subtilis	*Halobacterium*	*Drosophila melanogaster*
Borrelia burgdorferi	*Thermoplasma acidophilum*	*Homo sapiens*
Buchnera	*Methanobacterium thermoautotrophicum*	*Saccharomyces cerevisiae*
Campylobacter jejuni	*Methanococcus jannaschii*	
Chlamydia muridarum	*Pyrococcus horikoshii*	
Chlamydia trachomatis		
Chlamydophila pneumoniae J138		
Deinococcus radiodurans		
Escherichia coli		
Haemophilus influenzae		
Helicobacter pylori		
Lactococcus lactis		
Mycobacterium leprae		
Mycobacterium tuberculosis		
Mycoplasma genitalium		
Mycoplasma pneumoniae		
Neisseria meningitidis		
Pasterella multocida		
Pseudomonas aeruginosa		
Rickettsia prowazekii		
Synechocystis sp.		
Thermotoga maritima		
Treponema pallidum		
Ureaplasma urealyticum		
Vibrio cholerae		
Xylella fastidiosa		

B. *Summarized results of structural assignments to 40 genomes*

Genome group	Total no. of proteins	Matched proteins (%)	Matched amino acids (%)	Complete matches out of matched proteins (%)	No. of families
Archaea	1478-2694	31-54	33-45	46	286-320
Bacteria	480-5570	40-67	33-49	47	170-453
Yeast	6297	44	33	27	461
Metazoa	14,331-29,303	40-52	29-39	28	537-597

(PDB) in order to establish which pairs of adjacent domains have interfaces of unknown three-dimensional structure, and are thus interesting targets for structural genomics projects.

Results and Discussion

Structural assignments and domain composition of protein sequences

As described in Methods, it is easy to obtain an overview of the domain composition of the sequences from the different groups of organisms (archaea, eubacteria and uni- and multicellular eukaryotes) using our domainmap format files (Figure 1) of structural assignments to protein sequences. Previously, Teichmann *et al.*[8] observed that about two-thirds of the sequences of the bacterium *Mycoplasma genitalium* consist of more than one domain. Gerstein[9] calculated sequence length distributions from eight completely sequenced unicellular organisms, including archaeal, eubacterial and eukaryote genomes, and showed that more than two-thirds of the proteins are longer than an average domain length, implying that these are multi-domain proteins.

As shown in Table 2, we find that about 35 % of the matched protein sequences of the archaea and bacteria are completely covered by one domain. In the eukaryotes, the proportion of single-domain proteins is only about 20 %. The fraction of sequences completely matched by two SCOP domains is about 10 % of the sequences in the prokaryotes and 5 % of the matched sequences in the eukaryotic genomes. Sequences completely matched by more than two SCOP domains also form 1-2 % of matched sequences. All the partially matched sequences must have two or more domains, and these make about one-half of the sequences in prokaryotes and about three-quarters in eukaryotes.

Since only one-fifth to one-third of the genomic proteins are single-domain proteins, we conclude that the majority of the genomic proteins are multi-domain proteins, which have evolved through gene duplication, recombination and fusion. Proteins formed by combinations of domains are particularly abundant in eukaryotes, representing more than 80 % of all matched sequences. In pro-

```
>gi|2621052  28  29_2.35.4_94  1  95_2.30.7_164  77
>gi|2621610  1  2_3.54.1_127  18  145_3.29.1_403  14  417_3.39.1_472  40
```

Figure 1. Domainmap format for structural genome assignments. Domainmap format is useful when analysing domain structure of proteins on the large scale. This is an example from our MT structural assignments. The format contains following information for each protein in a given order: > followed by a protein identifier, number of unassigned residues, start residue of structural assignment followed by _ and the SCOP identifier of the assigned domain, followed again by the _ and the end residue of the assignment, ending with the number of unassigned residues at the C terminus.

karyotes they are somewhat less abundant, but still represent the majority of matched sequences (about 65 %).

These results and all our conclusions are derived from the set of proteins with assigned structures, which are about one-third to one-half of the predicted proteins in the genomes. Our structural assignments represent a larger fraction of the globular proteins in the genome, however, as a significant fraction, 20-30 %, of proteins are predicted to be transmembrane domains or disordered regions.[9–12] A comparison of the sequence length distributions of sequences with and without assigned structures suggests that we are dealing with a representative set of sequences, except for the regions between 100 and 150 amino acid residues (Figure 2). There are disproportionately few sequences with assigned structures in this region. This may be because the fraction of single-domain proteins is larger than what we observe using structural assignments, or the short sequences could be artefacts. In yeast, it has been suggested that many short predicted proteins are mispredictions[9] and in the multicellular eukaryotes, there is evidence that there are many fragmentary gene predictions.[13,14]

With the extensive information we have on the domain structures and evolutionary relationships of the proteins in 40 completely sequenced genomes, we want to investigate the patterns of domain combinations and thus reveal the evolutionary mechanisms that created complex proteins. First, we will turn our attention to tandem domains in proteins, and then to combinations of different types of domains. Finally, we ask whether more complex proteins have evolved by the creation of the new protein families, or the recombination of existing families. (Whenever the term family is used here, SCOP superfamily is meant.)

Tandem domains from the same family in polypeptide chains

Tandem domains from the same family within one polypeptide chain, also called domain repeats, may have evolved by recombination or fusion, in the same way as adjacent domains from two different families. However, tandem domains may have evolved by a different mechanism, internal duplication. Therefore, we consider this type of domain combination separately from combinations of different domain types. We define tandem domains as adjacent domains from the same family with less than 30 residues between them. We are interested in how frequently evolutionary mechanisms have resulted in tandem domains in the different groups of genomes, and the particular protein families involved.

Only a small fraction of genomic sequences contains tandem domains, and only a small fraction of genomic families form tandem repeats (Table 3A). In multicellular organisms, the fraction of sequences containing tandem domains is 11 %, on average, whereas in unicellular organisms this fraction is only half as much. Out of all families in multicellular organisms, about one-quarter forms tandems. In the unicellular organisms, again only half as many, about one-tenth of families, occur in tandems.

In the 40 genomes studied here, there are 203 domain families that form repeats. All these 203 families also occur combined with other families.

Domains of the same family can be duplicated internally just once, resulting in a tandem of two domains, or they can be duplicated more times, resulting in many consecutive domains. The number of consecutive domains from the same family is plotted against the number of occurrences in the groups of genomes in Figure 3. From the distri-

Table 2. Single and multi-domain proteins

	Matches of genomic sequences by *n* SCOP domains (%)					
	One domain matches		Two domain matches		Three or more domains matches	
Genome group	Complete	Partial	Complete	Partial	Complete	Partial
Archaea	36	43	9	8	2	2
Bacteria	35	42	10	9	2	2
Yeast	22	57	5	12	1	3
Metazoa	23	52	4	13	1	7

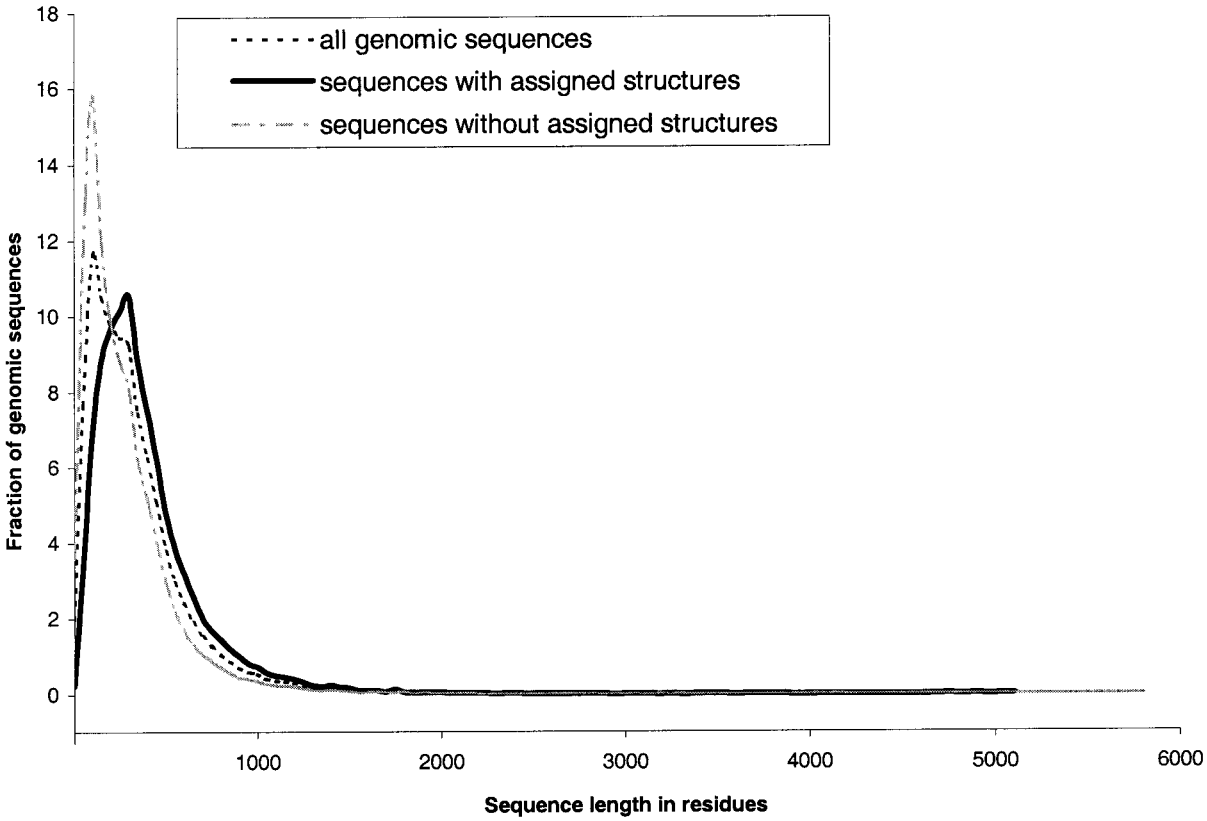

Figure 2. Sequence length distributions of all genomic sequences and sequences with assigned structures. The sequence length distribution of all genomic sequences is given as a black dotted line; that of sequences with assigned structures is a continuous black line; that of sequences without assigned structures is a dotted grey line. The lines are an average of the 40 genomes studied here. Each individual genome has a very similar distribution to this one. The *x*-axis represents the sequence length in residues and the *y*-axis shows the fraction of genomic sequences of a certain sequence length. Sequence lengths range from 100 to 5600 residues. The distributions of sequences with structural assignments and the one of all genomic sequences are very similar. The distribution of all genomic sequence lengths has a peak at 100 to 150 residues, which is probably an artefact of fragmentary gene predictions, particularly in eukaryotic genomes. The longest sequences are in eukaryotes, and only a small fraction of sequences are very long. Thus, sequences with assigned structures cover all sequence lengths of all genomic sequences and they are distributed in a very similar way.

bution, we observe that two adjacent domains occur most frequently in all genomes. The longest repeat length for the unicellular organisms are about five to ten domains long. There is not much difference between the number of domains in a tandem repeat, tandem repeat lengths, in the unicellular Archaea and Eubacteria and unicellular eukaryote (SC). An obvious difference between the unicellular and multicellular organisms is that the latter have much longer repeats. There are a significant number of proteins that have repeats of 30 or even 50 domains in the metazoan genomes.

The question arises as to the nature of these long repeats of domains in the metazoa. Are the repeats of domains from families in the protozoa just longer in metazoa, or are they repeats of families that arose later in evolution, and are thus specific to metazoa? There are 32 families that are seen in repeats in all three phylogenetic groups, and the 14 families that have repeats longer than three domains in at least one organism, or three-domain repeats in at least two organisms, are listed in Table 3B. Seven of the 14 common families listed are enzymatic families. The longest repeat of all the common families is eight domains of the DNA-binding homeodomain-like family.

The families involved in the top ten longest repeats in the metazoan organisms are given in Table 3C. Eight of the ten families are specific to metazoa, and seven of the ten families are mainly extracellular and involved mainly in cell adhesion and signaling. For the families involved in cell adhesion or other functions related to the cellular or physiological structure of an organism, the domain repeats provide a structural role, such as immunoglobulin domains in muscle proteins or laminin domains in proteins in the extracellular matrix. The three intracellular families are spectrin, a cytoskeletal protein, and two nucleic acid-binding families (zinc finger and KH domains). Cell adhesion, complex signaling and regulatory mechanisms, and large-scale structural components

Table 3. Tandem repeated domains from a same family

A. *The frequency of tandem domains in the phylogenetic groups and the fraction of families forming tandems*

Genome group	Genomic sequences containing tandem repeats (%)	Families forming tandem repeats out of all families in genome (%)
Archaea	5	10
Eubacteria	4	14
Yeast	6	11
Metazoa	11	24

B. *Families in repeats common to all three phylogenetic groups*

SCOP Family	No. of consecutive domains in a longest tandem repeat		
	Eubacteria	Archaea	Eukarya
Nucleic acid-binding proteins	6	5	5
PYP-like sensor domain	6	5	2
Thiolase-like	4	2	2
Rhodanese/cell cycle control phosphatase	4	2	2
CBS domain	3	4	4
Phosphoglucomutase	3	3	3
Metal-binding domain	3	2	7
Cupredoxins	3	2	6
PKD domain	2	5	2
P-loop hydrolases	2	3	3
Homeodomain-like	2	2	8
Aspartate/ornithine carbamoyltransferase	2	2	5
Actin-like ATPase domain	2	2	4
Regulatory domain of the amino acid metabolism	2	2	4

C. *The ten families with the longest repeats in the metazoan genomes*

SCOP Family	No. of consecutive domains in the longest repeat of the family		
	Metazoa	Archaea	Bacteria
Immunoglobulins	43	0	2
Spectrin	37	0	0
Cadherin	34	0	0
EGF/laminin	25	0	0
Spermadhesin	25	0	0
Classic zinc finger C2H2	23	0	0
Complement control module/SCR domain	21	0	0
Lipoprotein receptor	17	0	0
Ovomucoid/PC1-like inhibitors	16	0	0
KH-domain	14	2	0

became important as multicellular organisms evolved. We show here that some of the additional demands in these organisms were met by internal duplication.

Many of the families specific to metazoa involved in long repeats are flexible in the number of domains adjacent to each other in the proteins. Here, for instance, the immunoglobulin repeats in metazoa are present in 15 different lengths, ranging from two to 43 domains. It was shown for the cadherin family in *Caenorhabditis elegans* and *Drosophila*[13] and immunoglobulins in *C. elegans*[14] that gene predictions for the long genes involving these families were often incomplete. Therefore, the eukaryotic domain repeats may be even longer than shown here.

Combinations of domain families

Neighbouring domains

As for the tandem domains from the same family, we consider a pair of domains to be neighbours, in general, if they are not more than 30 residues apart. If domains are neighbours in one polypeptide chain, we say that they have combined with each other in the course of evolution, as the definition of a domain in SCOP is an independent evolutionary unit.

In the 40 genomes from the three kingdoms of life, we observe that only a small fraction of families combines with more than one other family. About one-third to one-half of the families does not combine with any other family, one-third of the families are seen as neighbours to only one other family, about one-ninth of families are adjacent to only an uncharacterised region and a variable fraction is never adjacent to any other families or regions. The details are given for each individual kingdom in Table 4A. Thus, for the majority of families that have domain neighbours, the adjacent domains are from one or two types of families.

A few families are very versatile in their combination partners, however, as shown in Table 4B. Most of these families are also the most abundant in genomes.[15] The reason for the abundance and

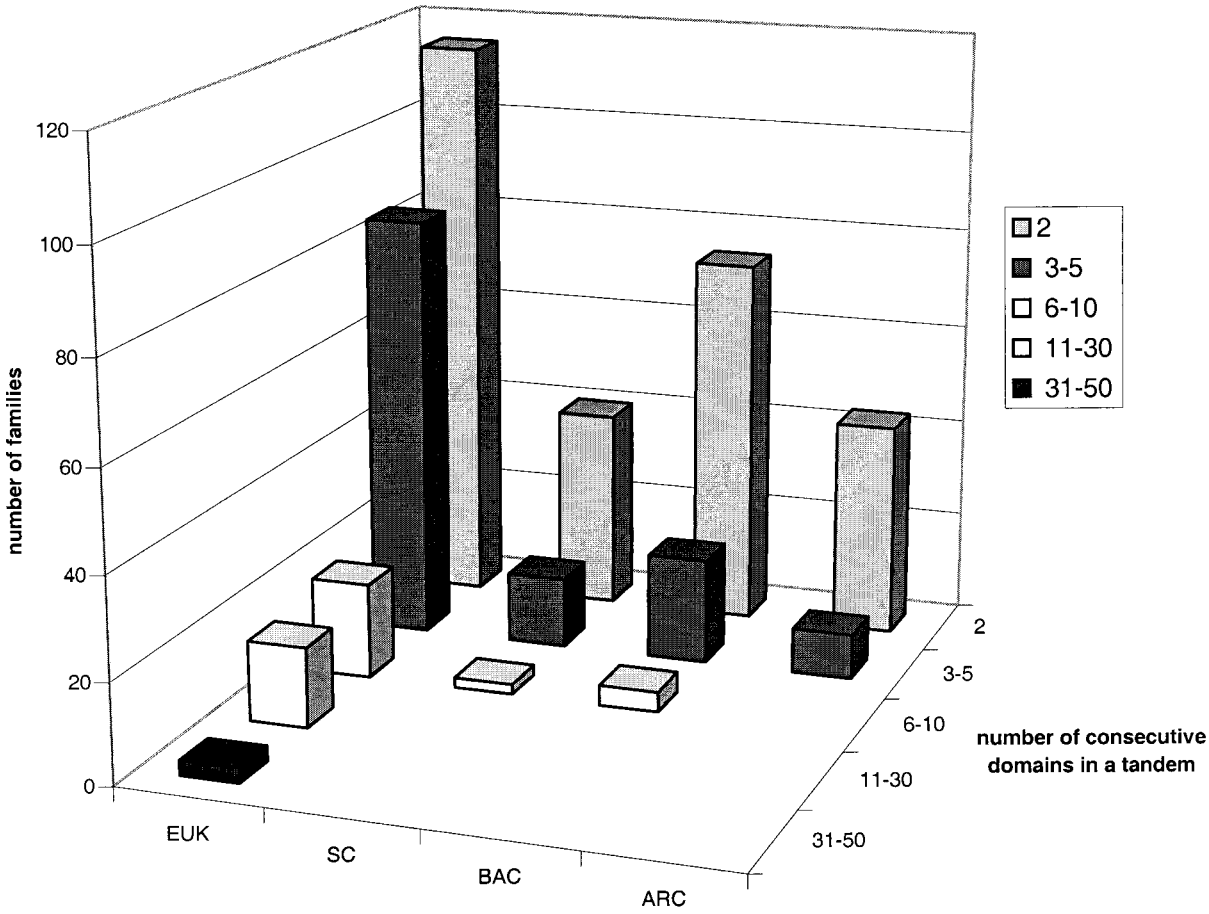

Figure 3. Distribution of the sizes of repeated domains in four groups of genomes. The distributions of tandem domains are plotted here for each genome group. The number of families that form repeats of a certain length determines the height of the bar. We observe that a repeat length of two domains occurs most frequently in all genomes. The distribution is very similar for all unicellular organisms (Archaea, Bacteria and yeast, SC) with very few repeats that are five to ten domains long. Multicellular organisms (EUK) have a significant fraction of much longer repeats, from ten to 50 domains long.

versatility of these families is their function. For instance, the energy for motion and reactions in the cell is often provided by the P-loop nucleotide triphosphate hydrolases. Domains from this family hydrolyse ATP or GTP and can act as kinases and transferases on their own or combined with different families. Rossman domains are similar, in that they provide oxidising or reducing energy through oxidation or reduction of the NAD(P)(H) cofactor. Transcription and translation are tightly regulated by proteins that consist of nucleic acid-binding motifs, such as ''winged helix'' DNA-binding domains or RING finger domains, combined with other domains responsible for the specificity of the regulation. In the metazoa, several families involved in signal transduction are among the most versatile, such as the intracellular domains of the protein kinase, EF-hand and PH domain families and the mainly extracellular EGF/laminin and immunoglobulin domains.

The pattern of few families combining with many other domains and most families having one or few partners is that of a power law, as shown in Figure 4(b). This power law implies that the graph of domain combinations is a scale-free network. The part of the graph surrounding the five most versatile families in *Archaeoglobus fulgidus* is shown in Figure 4(a). A scale-free type of network, recently described for the World Wide Web[16] and metabolic pathways,[17] has a few key nodes that are connected with many other nodes. All the other nodes have only a few connections. By definition the key nodes, or hubs, in this type of network have a fairly unique repertoire of nodes connected to them. This means that the versatile families, such as P-loop nucleotide triphosphate hydrolases and Rossmann domains, have a unique repertoire of combination partners that they do not share with other key families. With the future progress in homology detection and new protein structures, the domain combination network is going to expand, but there is no reason to believe that its form will change significantly.

If a family combines with many other families, then there must be at least as many domains in the family as there are partner families. The relationship between the size of a family and the number of neighbouring families in all seven

Table 4. Combination partners of domain families

A. *Percentage of families that combine with one, two or three or more families*

	Archaea		Bacteria		Yeast		Metazoa	
Number of partner families	n[a]	n + unc	n	n + unc	n	n + unc	n	n + unc
0	37	9	29	10	42	10	24	9
1	22	9	19	11	24	12	16	15
2	5	7	3	11	4	2	3	10
3 or more	2	9	2	15	2	4	0	23

B. *The most versatile families*

	Archaea	**Bacteria**	**Yeast**	**Metazoa**
Rank	Domain family (number of combination partners)[b]			
▲	P-loop hydrolase (21)	P-loop hydrolase (47)	P-loop hydrolase (16)	P-loop hydrolase (44)
	Rossmann domain (18)	Rossmann domain (29)	Rossmann domain (8)	Protein kinase (37)
	„Winged helix" DNA binding domain (14)	„Winged helix" DNA binding domain (23)	Tetratricopeptide repeat (6)	EGF/Laminin (32)
	Nucleic acid binding domain (11)	FAD/NAD(P) binding domain (19)	ARM repeat (6)	Immunoglobulin (28)
	Gluthatione synthetase ATP domain (8)	Homeodomain like (18)	Protein kinase (6)	EF-hand (25)
	4Fe-4S Ferredoxin (8)	Tetratricopeptide repeat (14)	Gluthatione synthetase ATP domain (5)	PH domain (24)
	FAD/NAD(P) binding domain (7)	CheY (14)	Class II amino acyl tRNA synthase (5)	RING Finger (24)
	PKD (6)	PYP-like sensor domain (13)	Ferredoxin reductase (5)	RNI-like domain (20)

[a] n stands for the percentage of families with the number of partner families given in the left column, and unc stands for uncharacterised region. We define an uncharacterised region as a sequence region more than 30 residues long without a structural assignment, which may be one or more domain families without a known representative structure.

[b] The number of combination partners for each family is given in parentheses next to the family name. The most versatile family in all genomes is the P-loop hydrolase family. The Rossmann domain family is also versatile across all genomes, having 11 different neighbouring domain families in Metazoa.

genomes is shown in Figure 4(c). The median of this relationship also follows a power law. One-half of the observed points follow the median for the region up to ten types of domain neighbours. The fluctuations in the region of more than ten combination partners per family are probably due to very small number of data points, since the genomes studied have relatively small number of very versatile families. Thus, the larger a family, the more likely it is to combine with more different types of domains, roughly according to a power law.

In order to check whether versatile superfamilies tend to combine with superfamilies of the same shape, we studied the combination of folds in addition to superfamilies. As described in Methods, the fold level in SCOP groups superfamilies that have the same topology of secondary

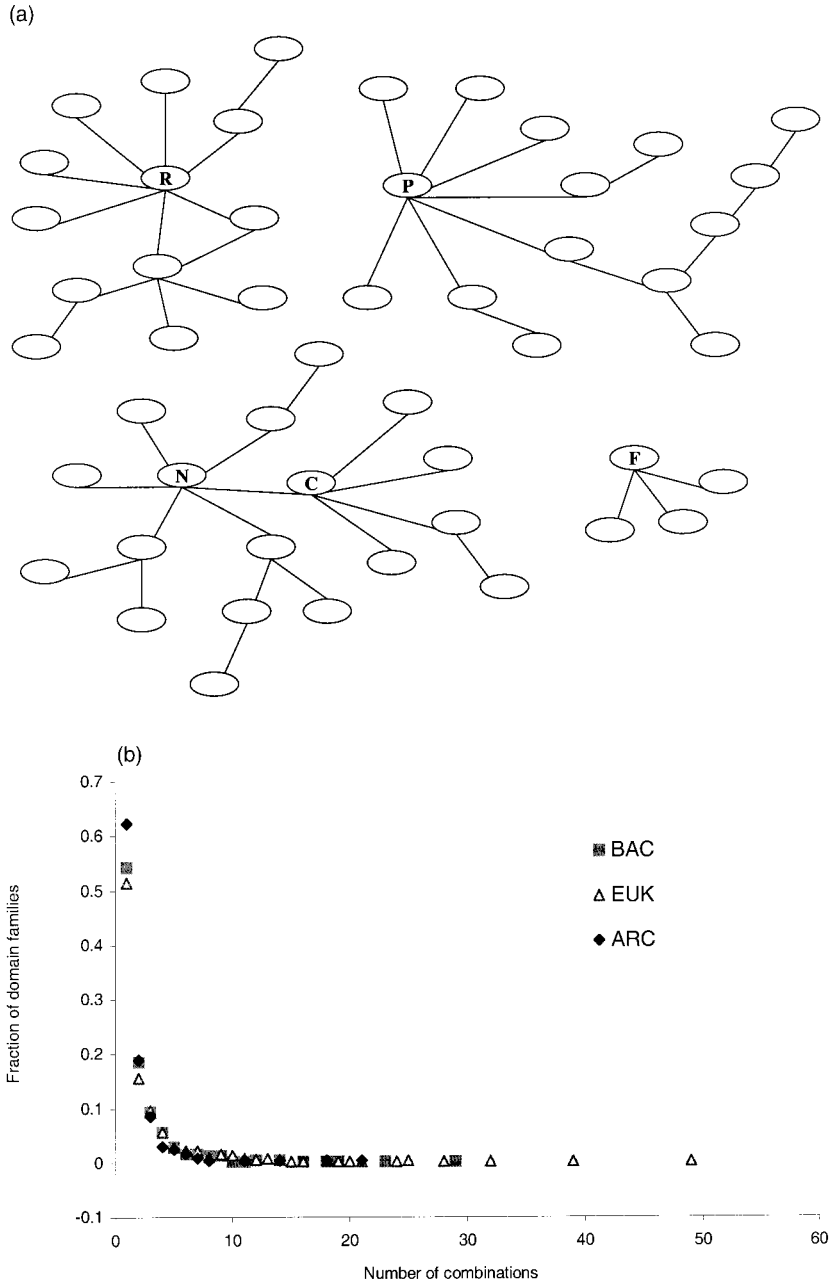

Figure 4 (*legend opposite*)

structure elements, but that do not show evidence of evolutionary relationship. Figure 5 illustrates the fact that folds are combined in a very similar manner to domain families (i.e. superfamilies), also according to a power law. Therefore, domain combinations are not subject to extensive geometric constraints: versatile families are combining with families from a variety of folds.

From the results presented above, we conclude that the number of types of neighbours is small for most families. Only a few protein families are very versatile in their combination partners, and these versatile families are also the largest families in the genomes.

N to C-terminal orientation of domain pairs

In accordance with the scale-free form of the domain combination graph, the number of types of neighbouring domains is limited for most of the families. The N to C-terminal orientation of the domains to each other in the sequence is also very limited, so that more than 90 % of domain pairs are seen in only one orientation to each other. A few families are seen in both orientations and we call

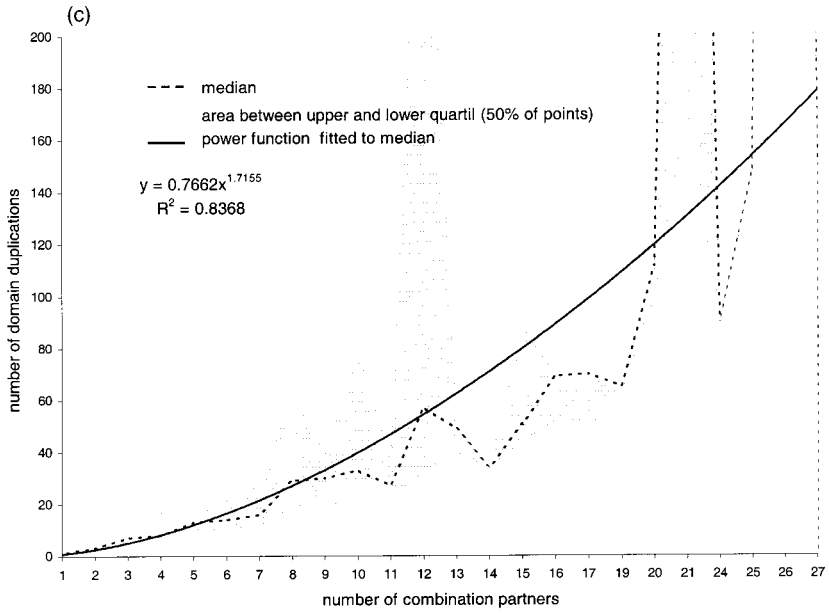

Figure 4. (a) Network of domain combination partners for the AF genome. The graph representing the five most versatile domain families and their combination partners is given here for the AF genome. Each ellipsoid represents a different protein family. Ellipsoids with a letter inside them are the most versatile families (R for Rossmann domain, P for P-loop hydrolases, N for nucleic acid-binding proteins, C for class II synthases, and F for FAD/NAD(P)-binding domain). For example, Rossmann domains combine with seven different types of domains, and one of these combines with four different types of domains. The graph resembles a scale-free network. The hubs are the most versatile families, which have unique repertoires of combination partners connected to them. AF is the simplest genome studied here in terms of domain combinations. The four key groups of connections surrounding the hubs are not interconnected here. In Eubacteria and Eukarya the hubs have more combination partners and the key groups are more interconnected, still preserving the fairly unique repertoire of combination partner for each key family. (b) Graph of fraction of domain families against number of combination partners. The number of partner families is given on the x-axis. The fraction of domain families out of all families that combine with other domains is plotted on the y-axis. We observe that about one-third of the families combine with only one family in all phylogenetic groups. ARC is Archaea, BAC is Eubacteria, and EUK is Eukarya. Only a few families in the Eubacteria and Eukarya combine with more than ten families, representing less than 1 % of families. A power law can be fitted to the data as follows: for ARC $y = 0.47x^{-1.8}$ and $R^2 = 0.92$; for BAC $y = 0.52x^{-1.8}$ and $R^2 = 0.93$; for EUK $y = 0.74x^{-1.8}$ and $R^2 = 0.92$. (c) Graph of number of family combinations against family size. For each family, the number of neighbour families in all genomes is plotted on the x-axis, while the family size in all forty genomes together is given on the y-axis. (Plots for the individual genomes follow the same trend). The grey area is the area between lower and upper quartile, meaning that 50 % of all data points are within it. The median is given as a dotted line; the power function fitted to the median is a continuous black line and its equation is given above. The fluctuations in the region of more than ten combination partners are probably due to the very small number of very versatile families.

such domain pairs inverted domain pairs. We studied the extent of inversion of domain pairs in each individual genome of the 40 genomes and across different genomes.

The fraction of inverted domain pairs and of domain families forming these pairs within each genome is very low, ranging from 1 %, on average, in bacteria to 4 % or 5 % in Eukarya (Table 5). When comparing domain combinations within one phylogenetic group or across all 40 genomes, the fractions increase a little (Table 5), but the conservation of domain pair orientation is still very strong. It is worth noting that about one-third of the inverted domain pairs are part of larger gene structures: domains combined as ABA, ABAB, BAB or BABA. Therefore, the actual frequency of the ''real'' inverted domains, those that are not part of these larger domain structures, is even smaller. The ABAB domain arrangements are probably internal duplications or gene fusions of two similar genes rather than the result of recombination.

To check whether this conservation of orientation is simply the result of a lack of recombination, or whether it is due to selection for one of the two possible orientations, we counted the number of times each domain pair occurs in different combinatorial contexts in an AB or BA orientation. By combinatorial context, we mean whether there are different domains neighbouring the AB or BA domain pair. The underlying assumption is that A and B have the chance to reshuffle each time there is a new combinatorial context. We compared the number of times AB or BA occurs in a combinatorial context to the expected distribution using a χ^2 test. (The expected distribution is that AB and BA

Table 5. Inverted domain pairs and family combinations

Group	Within one phylogenetic group (cross-genomic)		Average of individual genomes (intragenomic)	
	Inverted pairs (%)	Families in inverted pairs (%)	Inverted pairs (%)	Families in inverted pairs (%)
Archea	3	3	2	2
Bacteria	8	5	1	1
Eukarya	10	8	4	5
All 40 genomes	8	15	2	3

are equally frequent with deviations from the expected values being normally distributed.) The observed distribution differs from the expected distribution at the 0.1 % significance level. Assuming that our model using combinatorial context is reasonable, we can conclude that there is strong selection for one orientation of a domain pair in the process of domain recombination. The evolution of the interface and the function of a domain pair are evidently too costly to evolve twice, in two orientations, as compared to simply copying and altering the domain pair in one orientation by gene duplication and mutation.

In order to deduce which types of domain families tend to appear in inverted domain pairs, we classify them into catalytic families (e.g. P-loop hydrolases or DNA/RNA polymerases), regulatory families (e.g. SH2, SH3, zinc finger) and spacer families (e.g. ankyrin repeat, Ig, tetratricopeptide repeat). Since spacer families and regulatory families are not as directly coupled to the function of the pair as catalytic families, it is not surprising that more than 80 % of the inverted pairs involve these families. Only about 20 % of the inverted pairs consist of two catalytic domains. Two-domain proteins consist of one domain pair, and most probably the functions of these two domains are tightly coupled in producing the protein's activity. The composition of these pairs in terms of the domain family type is very similar to that described above for all inverted pairs. Out of 119 inverted domain pairs that we find in cross-genomic comparison of all 40 genomes, only 13 of the pairs are such two-domain proteins. Examples of these pairs are: SH2 and PH-like domains, P-loop hydrolase and EF-hand domains, and Rossmann and GroES-like domains.

In the PDB, there are 11 domain pairs whose structures are known for both orientations, and one of them is a two-domain proteins. From the PDB structures it is evident that some inverted domain pairs may actually have the same three-dimensional orientation to each other, if one domain pair is connected *via* a sufficiently long linker.

Thus, the appearance of inverted domain pairs is only marginal in genomes. These results imply that domain combinations are limited in their orientation as well as type of neighbour.

Conservation and variation of domain combinations among the three phylogenetic groups

In order to assess the extent to which domain combinations are variable in the different phylogenetic groups, we class families into those common to all three phylogenetic groups, and those present in only one or two of the phylogenetic groups. We then analyse how the two different types of families combine in each group of genomes. The three sets of domain combinations from the different kingdoms of life share 383 common families. As shown in Table 6A, most of the families in Archaea (88 %), more than a half in Eubacteria (63 %), and about half of the eukaryotic families are common families.

The domain combinations that involve common families represent most of the domain combinations present in each phylogenetic group, ranging from four-fifths in Eukarya to 98 % in Archaea (Table 6B). The 383 common families occur as combinations among themselves, in other words common family-common family combinations, or in combination with families that are not common to all three phylogenetic groups. Of the common family-common family combinations, 127 are found in Archaea, Eubacteria and Eukarya, as shown in Figure 6. Most of these 127 combinations represent families involved in macromolecule and small molecule metabolism.

Over half of the combinations involving common families in Eubacteria and Eukarya, and one-fifth in Archaea, are common family combinations specific to that kingdom. There are 804 combinations of pairs of common families, and only 127 of these are actually found in all three kingdoms, as shown in Figure 6. For instance, of all the common family-common family combinations in eukaryotes, 80 % are eukaryote-specific. As also shown by the statistics in Table 6B, this means that the common families recombine with each other much more than expected if one assumes that both common and non-common families have a likelihood of combining that is proportional to the fraction of common and non-common families present in the individual genomes in each kingdom. The difference between the expected and observed pattern of combinations between common and non-common families is significant at the 0.5 % level in

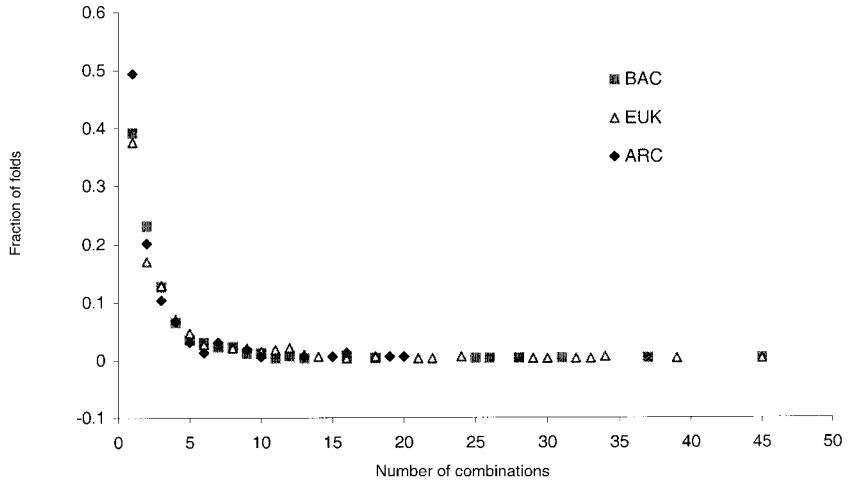

Figure 5. Graph of fraction of domain folds against number of combination partners. The number of folds is plotted on the x-axis, the fraction of folds out of all folds in that phylogenetic group that combine with other folds is given on the y-axis. About half of the folds are combined with only one other fold. ARC is Archaea, BAC is Eubacteria, and EUK is Eukarya. A power law can be fitted to the data as follows: for ARC $y = 0.46x^{-1.6}$ and $R^2 = 0.91$; for BAC $y = 0.4x^{-1.4}$ and $R^2 = 0.9$; for EUK $y = 0.68x^{-0.65}$ and $R^2 = 0.93$.

all three phylogenetic groups as tested with the χ^2 statistic.

Assuming that at least a large fraction of the kingdom-specific common family-common family combinations is genuinely not present in all three kingdoms, we are led to conclude that the common families have undergone more recombination than the kingdom-specific families. This could be because the large and very versatile families, such as P-loop nucleotide triphosphate hydrolases and Rossmann domains, are present across all genomes and are especially useful, and hence involved in many domain combinations, including species-specific ones. It is possible that there has been more evolutionary time for recombination between common families to take place, as they may be more ancient than non-common families. Evidently, novel combinations of domains among common families are an important part of the process of divergence of genomes that includes invention of new families, sequence divergence, and expansion and contraction of domain families.

Domain combinations in genomes only: targets for structural genomics

The whole set of domain combinations that we observe in all 40 genomes comprises 1307 combinations among 783 families. Of these combinations, 298 are observed in the PDB, which contains 356 domain combinations. Thus, out of all combinations in the PDB, we detect over four-fifths of these in the 40 genomes. There are 58 combinations in the PDB that we have not seen in the 40 genomes (Table 7). We consider the species distribution of these 58 pairwise domain combinations, but must remember that PDB entries and parsing of PDB entries can confuse the original organism of the protein and the organism where the protein is expressed for the crystallisation purposes. In any case, some of the 58 domain combinations are in organisms other than those in our set, while others are not present in the protein predictions that we have, such as several human proteins. Over three-fifths of the domain combinations found only in PDB are from different types of bacteria.

This means that we detect 1009 novel combinations of structural superfamilies, not seen in PDB. Structural genomics projects have the aim of extending the fold library and hence solving the structures of sequences without assigned PDB structures,[18,19] but another aim of the structural genomics efforts should be to elucidate the way domains interact with each other. This will provide important insights into the function of individual proteins and protein-protein interactions. The 1009 novel combinations of structural superfamilies, not seen in PDB, are suitable to as targets for structural genomics and details of these and the proteins are available in our electronic appendix†.

Conclusion

Here, we provide a first quantitative insight into the domain combinations in the three kingdoms of life. The majority of genomic proteins, two-thirds in unicellular organisms and more than 80 % in metazoa, are multi-domain proteins created as a result of combinations of domains. We observe a pattern of domain combinations where only few domain families combine with many types of domains and most families have one or few combi-

† http://www.mrc-lmb.cam.ac.uk/genomes/DomCombs/

Table 6. Combinations involving families common to all three kingdoms

A. *Common families in the three phylogenetic groups*

Phylogenetic group	No. of all families	No. of common families	No. families in combinations	No. common families in combinations
Archaea	437	383	234	192
Eubacteria	604		366	
Eukarya	704		477	

B. *Domain combinations involving common families*

Phylogenetic group	Total no. combinations	Combinations of common families with common families (%)	Combinations of common families with non-common families (%)	Combinations of non-common with non-common families (%)
Archaea	265	93/80	5/19	2/1
Eubacteria	546	77/47	17/44	6/9
Eukarya	894	55/31	27/50	18/19

Common families are those present in all three kingdoms, while non-common are those observed in only one or two kingdoms. The percentage of observed combinations for a given type out of all combinations is given, followed by the expected percentage. The expected percentage is calculated from the number of common and non-common families out of all families in the kingdom given in A, taking into account the combinations that are present in the individual genomes. There are many more common family-common family combinations than expected.

nation partners. This pattern can be described as a scale-free network.

The majority of domain combinations in all three kingdoms of life are those involving families that are shared among the three kingdoms, the common families. We show that there are many kingdom-specific combinations of families common to all three kingdoms. This means that recombination of common families among each other contributes more to the process of divergence at the level of domain combinations than common families combining with kingdom-specific families. This implies that in creating new functions, nature has, over time, combined more common building blocks with each other than kingdom-specific ones. Therefore, the common families also tend to be versatile and large, and so of use to organisms in many different contexts. At the same time as observing many protein family combinations, and some versatile families, we see that the preservation of the N to C-terminal orientation of the combined domain pair is almost absolute, with only a few examples of domain pairs appearing in both orientations. The analysis of domain repeats shows that special functional requirements that appeared with the evolution of metazoa, such as cell adhesion and cell-signalling, were fulfilled by internal duplication of domains of metazoa-specific families.

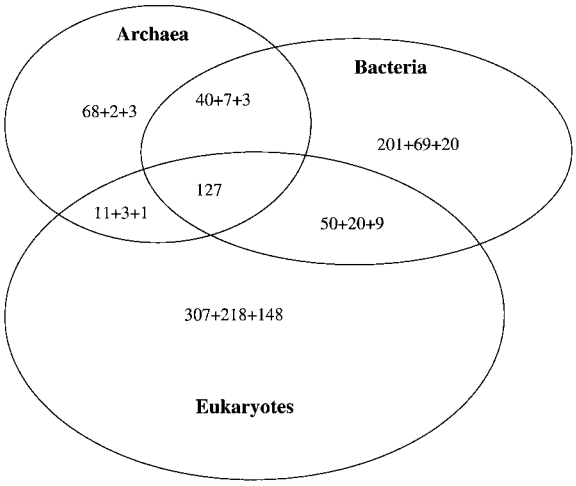

Figure 6. Unique and shared domain combinations. The diagram shows the pattern of domain combinations involving domain families common to the three phylogenetic groups. For each section of the Venn diagram, the number of domain combinations is split into three types: from left to right domain combinations in which both families are common to all three kingdoms, domain combinations in which one family is common to all three kingdoms and the other is not, and domain combinations in which neither family is a common family. It is obvious that the majority of domain combinations involve common families, even though most of these common family-domain combinations are not shared between kingdoms.

Methods

Domains in the structural classification of proteins (SCOP) database

As mentioned above, the domain definition used here is that of the Structural Classification of Proteins (SCOP) database developed by Murzin and co-workers.[1] SCOP

Table 7. The 58 domain combinations from PDB that are not detected in the 40 genomes

Group of organisms	Number of domain combinations from PDB not detected in the seven genomes
Virus	6
Plants	4
Mammals	7
Fungi	1
Protozoa	1
Other eukaryote	2
Bacteria	37

contains domain definitions and describes evolutionary relationships between the proteins whose three-dimensional structures have been determined. The domains in SCOP are evolutionary units: they are classified individually only if there is evidence from known protein structures that the domains can undergo independent duplication and recombination. Small proteins and most medium-sized proteins consist of one domain. If structural domains from a multi-domain protein are, up to now, seen only linked to each other, they are classified as a single unit in SCOP, and these entries form less than one-tenth of the superfamily entries in the current SCOP database. This is because it is possible that these domains are functional only if they are linked to each other. In most cases, however, later protein structure data have provided evidence for the independent evolution of these domains.

The SCOP domains are clustered together into families if they are close evolutionary relatives, usually detectable at the sequence level. Families are brought together into superfamilies of more distant evolutionary relatives. Domains in superfamilies may have low sequence identity but their structural and sometimes functional features strongly suggest a common evolutionary origin. Folds are the next level of the hierarchy, grouping together superfamilies that have the same secondary structures in the same arrangement, but without evidence for evolutionary relationships between the superfamilies.

We used SCOP version 1.53 containing 25,164 structural domains clustered into 859 superfamilies in SCOP classes 1 to 7. (These sequences are available from the supplementary website.) In this work, we studied the 859 superfamilies assigned to genomic sequences in order to elucidate the domain structure and combinations in 40 genomes. We use the term family in this text meaning a SCOP superfamily.

Forty genomes and four phylogenetic groups

SCOP domain assignments to genome sequences are the data set analysed here, and the set of completely sequenced genomes we chose are diverse, so that we hope to cover much of the domain family and domain combinatorial space in nature. As listed in Table 1A, there are seven archaeal genomes, 28 eubacteria, one unicellular eukaryote (*Saccharomyces cerevisiae*, SC) and four multicellular eukaryotes. The four metazoa are one plant (*Arabidopsis*), one vertebrate (*Homo sapiens*) and two invertebrates (*C. elegans*, CE; *D. melanogaster*, DM). It should be noted that the quality of the protein predictions in the multicellular genomes is variable, and the human protein predictions undoubtedly contain many fragments, as can be seen from the short average protein length of the human protein data set. Nevertheless, we have included the Ensembl human protein predictions in our analysis as part of the set of multicellular eukaryotes.

The unicellular organisms come from very different external environments: for instance their optimal temperatures range from room temperature (e.g. *S. cerevisiae*) to 85 °C in deep marine subsurface oil areas (e.g. *Archaeglobus fulgidus*), and they range from from autotrophs (e.g. *Methanobacterium thermoautotrophicum*) to optional parasites (e.g. *Escherichia coli, Bacillus subtilis*) and obligate parasites (e.g. *M. genitalium*).

In some parts of this investigation, it is convenient to compare three or four different phylogenetic groups. To create one phylogenetic group, we take all the genomes from this group and make a non-redundunt union of the features studied in the two genomes, for instance domain combinations or tandem domains. When we refer to the phylogenetic group in the text we mean these non-redundant unions. The ''Archaea'' group the seven archaeal genomes, ''Eubacteria'' groups together the 28 bacterial genomes, the unicellular eukaryote is *S. cerevisiae* and the multicellular eukaryotes are the four metazoan genomes.

Assigning domains to genome sequences

The SCOP domains were assigned to the genome sequences using Hidden Markov models (HMMs).[20,21] HMMs are one of the most sensitive sequence comparison methods currently available.[22] For each non-identical SCOP domain, at the 95 % sequence identity level, an HMM was generated by Gough and co-workers[23] using the iterative SAM-T99 method.[24] The genome sequences were scanned against this Hidden Markov model library, finding single or multiple matches with significant scores.

A match was accepted if the expectation value was at least 0.01, as calibrated to a 1 % error rate in an assessment of the Hidden Markov model library using the known evolutionary relationships of the domains in the SCOP database. Adjacent matches were allowed to overlap up to 20 % of the length of the shorter match. In the case of longer overlaps, the match with the better score was retained and the match with the worse score was rejected.

With this procedure, domains that are inserted into discontinuous domains are detected as well as normal continuous domains. The fraction of proteins that contain discontinuous and inserted domains is small, ranging from 0.5 to 7 % of all proteins with structural assignments in a genome and 4.2 %, on average, in our set of 40 genomes. As the process that produces inserted domains is different from the process that leads to combinations of adjacent domains, we have excluded all inserted domains from our analysis. In proteins with inserted domains, only the discontinuous domain and its adjacent domains are retained. We have completely discarded all proteins with more than one domain inserted into another domain next to each other, which are less than 1 % of all proteins with structural assignments.

As summarized in Table 1B, between one-third and one-half of all residues in the genomes have a structural assignment with this procedure. The detailed results for each genome are given in the electronic appendix. Out of the 859 superfamilies in SCOP version 1.53, 783 occur in at least one of the genomes. Members of about 300 families are assigned to the archaeal genomes, while the number of families increases by about a third in the large bacterial genomes to 400 or 450 families and 461 in the unicellular yeast genome. The multicellular organisms have almost double the number of families as the Archaea, ranging between 537 for *C. elegans* and 594 for the human protein predictions.

Analysing domain combinations

In order to be able to analyse the domain structure of the proteins more easily, we translated the structural assignments into ''domainmap'' format. The domainmap format contains the following information for each pro-

tein: protein identifier, SCOP superfamily identifier of assigned domains from N to C terminus, with flanking residue numbers for the domain region, and the lengths of unassigned regions. An example is given in Figure 1.

Using domainmap format files, the domain properties of the proteomes can be ascertained easily. Some of the matched proteins have long regions that have not been assigned SCOP domains. We consider protein sequences to be matched by SCOP domains completely if the regions flanking the matched domains are less than 30 residues long. We also consider domains to be adjacent to each other only if there are less than 30 residues between the two domain assignments. If there is a longer unmatched region adjacent to an assigned domain, then the neighbouring domain is considered to be of an unknown character. The conservative threshold of 30 residues was chosen, because this makes it unlikely that either a transmembrane region or a SCOP domain is present in the remaining regions, as all but 0.4 % of SCOP domains (at less than 95 % sequence identity) are longer than 30 residues. In terms of domain composition of individual proteins, changing the threshold increases the fraction of completely matched proteins out of all matched proteins by about 10 % for every 30 residues that the threshold is increased from 30 to 90 residues. In terms of domain combinations, increasing the threshold also adds roughly 10 % of new combinations every 30 residues, though some of these could be errors, such as combinations divided by a transmembrane region or a small globular domain. With our 30-residue threshold, we minimize the likelihood of obtaining incorrect combinations.

About one-half of the sequences matched by a SCOP domain in the prokaryote genomes and a little less than a third in the eukaryote genomes are complete matches, as described in Table 1B. Therefore, we have a precise description of the domain composition for 10-30 % of the proteins in these genomes and an insight into about one-half of the proteins in these genomes.

When studying the phylogenetic distribution of domain combinations, we consider pairwise combinations of neighbouring domains according to the 30 residue criterion for neighbourhood. We do not take into account the N to C-terminal orientation of a pair of domains, as this is conserved in most cases. The exceptions are treated in a separate section. The number of different families that are adjacent to the domains of one family can include repetition of the family itself in the case of internal duplications. (These are also treated in a separate section on tandem domains from the same family.) The domains in some families are always observed as completely covering a protein without any domain neighbours, but in some families one or more of the members are observed adjacent to an unassigned region. These two cases are distinguished throughout our analysis.

Acknowledgements

We are grateful to Cyrus Chothia for helpful discussions and encouragement. We thank Graeme Mitchison and Karsten Neuhoff for discussions about the statistics of inverted domains. G.A. has an LMB Cambridge Overseas Fellowship and S.A.T. has a Beit Memorial Fellowship for Medical Research.

References

1. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
2. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH: a hierarchic classification of protein structure. *Structure,* **5**, 1093-1098.
3. Riley, M. & Labedan, B. (1997). Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857-868.
4. Chothia, C. (1992). One thousand families for the molecular biologist. *Nature,* **357**, 543-544.
5. Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897-905.
6. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding proteins. *Nature,* **250**, 194-199.
7. Patthy, L. (1994). Introns and exons. *Curr. Opin. Struct. Biol.* **4**, 383-392.
8. Teichmann, S. A., Park, J. & Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA,* **95**, 14658-14663.
9. Gerstein, M. (1998). How representative are the known structures of the proteins in complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497-512.
10. Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.
11. Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Letters,* **423**, 281-285.
12. Wallin, E. & von Heine, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean and eukaryotic genomes. *Protein Sci.* **7**, 1029-1038.
13. Hill, E., Broadbent, I., Chothia, C. & Pettitt, J. (2001). The cadherin superfamily of proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *J. Mol. Biol.* **305**, 1011-1024.
14. Teichmann, S. A. & Chothia, C. (2000). Immunoglobulin superfamily proteins in *Caenorhabditis elegans*. *J. Mol. Biol.* **296**, 1367-1383.
15. Teichmann, S. A., Chothia, C. & Gerstein, M. (1999). Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390-399.
16. Albert, R., Jeong, H. & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature,* **406**, 378-382.
17. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organisation of metabolic networks. *Nature,* **407**, 651-654.
18. Brenner, S. A. (2000). Target selection for structural genomics. *Nature Struct. Biol.* **7**, **(Suppl.)**, 967-969.
19. Burley, K. S. (2000). An overview of structural genomics. *Nature Struct. Biol.* **7**, **(Suppl.)**, 932-934.
20. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **3**, 361-356.

21. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications in protein modelling. *J. Mol. Biol.* **235**, 1501-1531.
22. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
23. Gough, J., Chothia, C., Karplus, K., Barrett, C. & Hughey, R. (2000). Optimal hidden Markov models for all sequences of known structure. In *Currents in Computational Molecular Biology* (Miyano, S., Shamir, R. & Takagi, T., eds), vol. 124-125, Universal Academic Press, Tokyo, Japan.
24. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846-856.

Edited by G. von Heijne

(Received 14 December 2000; received in revised form 8 May 2001; accepted 9 May 2001)