Fold usage on genomes and protein fold evolution

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · SEPTEMBER 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20506 · Source: PubMed

2 AUTHORS:

Sanne Abeln
VU University Amsterdam
24 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Charlotte M Deane
University of Oxford
110 PUBLICATIONS **3,016** CITATIONS
SEE PROFILE

Fold Usage on Genomes and Protein Fold Evolution

Sanne Abeln and Charlotte M. Deane*
Department of Statistics, University of Oxford, United Kingdom

ABSTRACT **We review fold usage on completed genomes to explore protein structure evolution. The patterns of presence or absence of folds on genomes gives us insights into the relationships between folds, the age of different folds and how we have arrived at the set of folds we see today. We examine the relationships between different measures which describe protein fold usage, such as the number of copies of a fold per genome, the number of families per fold, and the number of genomes a fold occurs on. We obtained these measures of fold usage by searching for the structural domains on 157 completed genome sequences from all three kingdoms of life. In our comparisons of these measures we found that bacteria have relatively more distinct folds on their genomes than archaea. Eukaryotes were found to have many more copies of a fold on their genomes. If we separate out the different fold classes, the alpha/beta class has relatively fewer distinct folds on large genomes, more copies of a fold on bacteria and more folds occurring in all three kingdoms simultaneously. These results possibly indicate that most alpha/beta folds originated earlier than other folds. The expected power law distribution is observed for copies of a fold per genome and we found a similar distribution for the number of families per fold. However, a more complicated distribution appears for fold occurrence across genomes, which strongly depends on fold class and kingdom. We also show that there is not a clear relationship between the three measures of fold usage. A fold which occurs on many genomes does not necessarily have many copies on each genome. Similarly, folds with many copies do not necessarily have many families or vice versa. Proteins 2005;60:690–700.**

Key words: structural genomics; evolution; protein folds; alpha/beta folds; power law

INTRODUCTION

Over 150 complete genomes have now been sequenced including examples from all three kingdoms of life, allowing us to begin the genomic study of fold evolution.[1] Several studies have now been made which assign structural domains to genes on completed genomes.[2–7] Profile based sequence alignment methods provide a fast tool for finding these relationships.[8] In particular they can show whether a fold occurs on a genome[2] and can give us the number of copies of a structural domain per genome.[9] Other studies have concentrated on assigning functional relationships between proteins using occurrence patterns,[5,10] classifying folds by cellular function,[4] analyzing domain combinations,[7] and developing evolutionary fold trees using occurrence and copy patterns.[6] We will investigate different measures of fold usage on the completed genomes to gain insights into evolutionary relationships between folds.

The set of currently known structures in the Protein Data Bank (PDB)[11] limits the set of genes for which we can assign a structure. Furthermore, it is probable that the PDB is biased toward certain structures, for example those which are easily crystallizable or are of significant biological interest. This could mean that the set of known structural domains is not an evenly distributed sample of all structures in nature. Even so, profile-based methods such as PSI-BLAST[8] have been demonstrated to assign structures to about 40% of the proteins in a bacterial genome.[12] Algorithms based on hidden Markov models have been shown to assign structural domains to around 50–60% of the genes on bacterial genomes.[1,13] This shows that a significant proportion of genes on genomes can be annotated with a currently known structure. In this study we assigned 37% of all genes from archaea, bacteria, and eukaryotes to a structural domain using a fast searching method, PSI-BLAST.

In order to obtain single structural domains for proteins from the PDB and group these together into evolutionary related groups or folds, we need a classification system for protein structures such as SCOP,[14] CATH,[15] or the Dali database.[16] We use SCOP, which is based on a four-level hierarchy. The top level in the hierarchy groups structure on the basis of their secondary structure elements, such as all alpha helical structure, all beta strand, alternating alpha and beta elements, or a mixture of alpha and beta. The second level is the fold level which describes the topology of a protein structure; at this level we cannot presume the protein structures are evolutionarily related but they do share common secondary structure elements in the same order. The next level divides folds into superfamilies; these are thought to be evolutionarily related by evidence of similar active sites and/or function. The final level is the sequence family, in which the proteins have a high level of sequence identity, indicating

Grant sponsor: EPSRC; Grant sponor: the Wellcome Trust

*Correspondence to: Charlotte M. Deane, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK; E-mail: deane@stats.ox.ac.uk

Received 7 July 2004; Revised 22 October 2004; Accepted 20 January 2005

Published online 6 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20506

clear evolutionary relations. Divergent and convergent evolution of superfamilies under a fold have been discussed in many analyses of protein structure evolution. Although some particular examples of convergent evolution are possible and some examples for convergence of local structure are known, divergent evolution is suggested for most superfamilies under a fold.[17] In particular a model has recently been developed which explains the power law distribution of edges per node in a network of structural similarity between domains. The model shows that the desired distribution can be created assuming only divergent evolution.[18] A related discussion is the total number of folds in nature. It is now generally accepted that there is a limited set of folds in nature.[17,19–23] However, it is not clear whether there are more fold topologies possible than those which currently exist in nature. Nature might have reached a limit on the total number of possible fold topologies[24] or there might be more topologies to be explored, e.g., for larger alpha/beta proteins not all possible secondary structure connections have yet been found.[25]

In order to investigate the evolutionary relationships between protein structures, we will consider the usage of folds. It has been noticed that some folds are found far more commonly than others.[2,9,19] Fold usage can be explored with several different measures such as occurrence across the genomes, copies per genome and the number of sequence families using the same fold. Comparing these measures for different folds can give us an indication of the age of each fold and can give us insights into the mechanisms behind fold evolution. First we will describe common measures of fold usage and how these have been used previously:

(1) *The number of sequence families per fold.* The distribution for the number of families per fold has been described to follow many different functions including an exponential,[22] a highly stretched exponential,[23] a logarithmic function,[21] and a power law.[17] In order to create a good estimate for this distribution, folds have also been separated in three different groups: unifolds with only one sequence family per fold, super folds with very many sequence families per fold, and an intermediate group, mesofolds.[19] This measure has also been used to estimate the total number of folds and families we can expect in nature.[19,21,22]

(2) *The number of copies of a fold per genome*. In other words this is the number of genes that occur on a genome classified under the same fold. The distribution for the number of copies of a fold on a single genome is thought to follow a power law.[9] This would mean that most folds have only one copy on each genome, and a very small number of folds have a very large number of copies, the distribution tails off asymptotically to zero. This distribution can be created by self dependency of a variable, e.g., a fold with many copies is more likely to duplicate again than a fold with only a single appearance on a genome. Qian et al. showed that a power law distribution can arise when the number of genes on a genome can increase without restriction.[9] However Karev et al.[26] provide a birth, death, and innovation model (BDIM) for genomes in equilibrium, such that the number of genes on the genome is stable, which can also generate a power law distribution for fold copies on a genome. For some superfamilies on bacterial genomes the number of copies is correlated with genome size;[27] these superfamilies seem to perform specific cellular functions.[27,28]

(3) *The number of genomes on which a fold occurs.* Only a few studies and on a limited number of genomes have been carried out on this measure[2,3], making it hard to draw further conclusions. Nevertheless, this might in fact be a measure for the relative evolutionary age of a fold: if a fold occurs on all genomes in a set, it is likely that it already existed on the last common ancestor of this set (not taking into account gene loss and lateral gene transfer). This can give us an indication how far back in evolution this fold first occurred.

We investigate how these measures behave on different sets of folds and genomes. We find power law behavior in most kingdoms and fold classes for both the distribution of fold copies on a genome and the number of superfamilies per fold. Surprisingly the distribution of fold copies on eukaryotes seems to have a much worse fit to a power law. The distribution of occurrence across genomes is more complicated and depends heavily on kingdom and fold class. We also investigate whether ancient folds have many copies and have more sequence families and superfamilies per fold. We find that many folds which occur on a large set of genomes only have a low number of copies and/or superfamilies, whereas folds with many copies usually occur on many genomes and are likely to be old. We explore whether the chance to create a new superfamily becomes higher when there are many copies of a fold, but find that this is not supported by the data. We also show that the distribution of these fold measures is different for the alpha/beta class as compared to all other fold classes. In order to verify our results we performed a similar analysis with SUPERFAMILY[13] assignments. In general the results from the PSI-BLAST and SUPERFAMILY assignments agree.

METHODS

Search for Structural Domains

We looked for structural domains in the protein sequences of 157 completely sequenced genomes, containing 10 eukaryotes, 130 bacteria, and 17 archaea. The protein sequences were obtained from ftp://ftp.ncbi.nih.gov/genomes/. In order to identify the fold(s) of each gene, we used structural domains from the SCOP classification[14] and PSI-BLAST.[8] The amino acid sequences for the SCOP domains were obtained from a filtered database[29] containing sequences with 95% sequence identity or less from SCOP release 1.63. To check for related sequences we ran the SCOP domains with PSI-BLAST against a merged nonredundant database of the 157 genomes, using a maximum of five PSI-BLAST iterations per domain and an

e-value threshold of 10^{-5} for the inclusion of protein sequences.

Power Laws

The distribution of fold copies and families per fold both appear to follow a power law. However, demonstrating a power law is difficult with sparse data. Using the "descending rank" we can use all observations in our set of data.

A problem for using frequency bins or plotting all frequencies to show a power law distribution occurs in regions of very low frequency. In these regions there will be many bins not containing any data (folds). However, this is usually not shown in the diagram, hiding the fact that observations with low frequencies occur less and less. This results in horizontal lines of observations at the lowest frequencies. To overcome this problem we could use variable bin sizes[20] or the descending rank of the observations rather than the frequency.

It is easy to show that the descending rank of a power law distribution also follows a power law, with power increased by one. A probability distribution function following a power law has the form: $f(y) = cy^{-\tau}$ where c and τ are constants. We can get the descending rank for element x, by a summation over all the elements $> x$. This summation can be approximated (for $\tau \neq 1$) by an integral, i.e., the number of observations (N) times the surface under the probability distribution curve form x to ∞ represents the number of observations greater than x. We get:

$$r(x) = N\sum_{x}^{\infty} cy^{-\tau} \approx N\int_{x}^{\infty} cy^{-\tau}\,dy = Nax^{-\tau+1}$$

where $a = \frac{c}{1-\tau}$ is another constant. Hence the descending rank of a power law distribution should follow a straight line on a log–log plot with a negative slope. Furthermore we can obtain the power of the distribution function by subtracting one from the power of the descending rank function. More detailed analysis for the use of ranking methods to show a power law distributions can be found in Adamic and Huberman[30] and Troll and Graben.[31]

Measures of Fold Usage

In order to analyze the fold content for each genome we created a table with the number of copies for each fold per genome. We created similar tables for copies per superfamily and per sequence family as classified by SCOP. We can calculate the number of distinct folds per genome and the number of genomes on which a fold occurs from these tables. To obtain the average number of copies of a fold per genome, we divide the total number of copies for a fixed set of genomes by the occurrence for the fold on the same set of genomes, so that the average number of copies is calculated for only those genomes on which the fold occurs.

To investigate the effect of grouping superfamilies into folds, we created data sets where superfamilies were assigned to random folds, with the same distribution as in the real data.

SUPERFAMILY Assignments

The SUPERFAMILY database[13] uses Hidden Markov Models to assign superfamilies to genes on completed genomes. These assignments in general cover almost 60% of the genes with a false positive rate $< 1\%$. The data we used from SUPERFAMILY contains 19 archaea, 120 archaea, and 37 eukaryotes and does not contain different strains of a genome (in contrast with the data we obtained for PSI-BLAST).

RESULTS AND DISCUSSION

Genomes

Using PSI-BLAST, we found one or more SCOP domains for 37.4% of the genes in our set of genomes. In total over 200,000 hits were found. Some high ratios of up to 66% were found for Bacterial symbionts, such as *Buchnera sp.*, *Blochmannia floridanus*, and *Wigglesworthia*. On the other hand, *Leptospira interrogans* (22.8%), *Pirellula sp.* (23.2%), *Borellia burgdoferi* (23.2%), and *Plasmodium falciparum* (22.7 %) have a relatively low percentage of matched genes. Approximately 10% of all families, superfamilies, and folds from the SCOP database were not found on any of the 157 genomes. Many of the PDB entries for these families are viral proteins. In the all alpha and small protein class some eukaryotic proteins are not found on any of the genomes in our set (e.g., toxins and pollen allergens); it is likely that these are lineage specific folds. All folds of the alpha/beta class are found on at least one of the genomes in our set. There are a few (< 10) undetected protein domains derived from one of the genomes in our set. These false negatives might be caused by annotation errors on the genomes or not detected since PSI-BLAST was used with a low complexity filter.

First we examine how the number of distinct folds found on a genome changes with genome size. Figure 1(a) shows that it increases in a slower than linear fashion. The differences in fold coverage of structural domains on the genomes create noise in the data. This was filtered by using the number of genes on the genome, that had a structural hit, rather than just using the total number of genes on the genome. This creates a much better fit with a lower residual and a good approximation of a straight line when putting the number of genes on a logarithmic scale [Fig. 1(b)]. Similar results were found by Wolf et al. using a much smaller set of genomes.[3] The trend does not change when instead of folds, superfamilies are plotted against genome size, but this similarity may be solely the effects of clustering superfamilies into folds (see discussion in section: Distribution of (Super) Families per Fold, below).

Figure 1(b) shows that archaea and eukaryotes seem to have relatively few distinct folds compared to similarly sized bacteria or have larger genomes with the same number of distinct folds. (The points representing archaeal and eukaryotic genomes lie below the linear regression line.) In archaea this may be due to their extreme living environments, which will only allow a subset of physiochemically very stable fold structures.[32] However, this effect is not seen for bacterial extremophiles, therefore not all species in extreme environments necessarily have

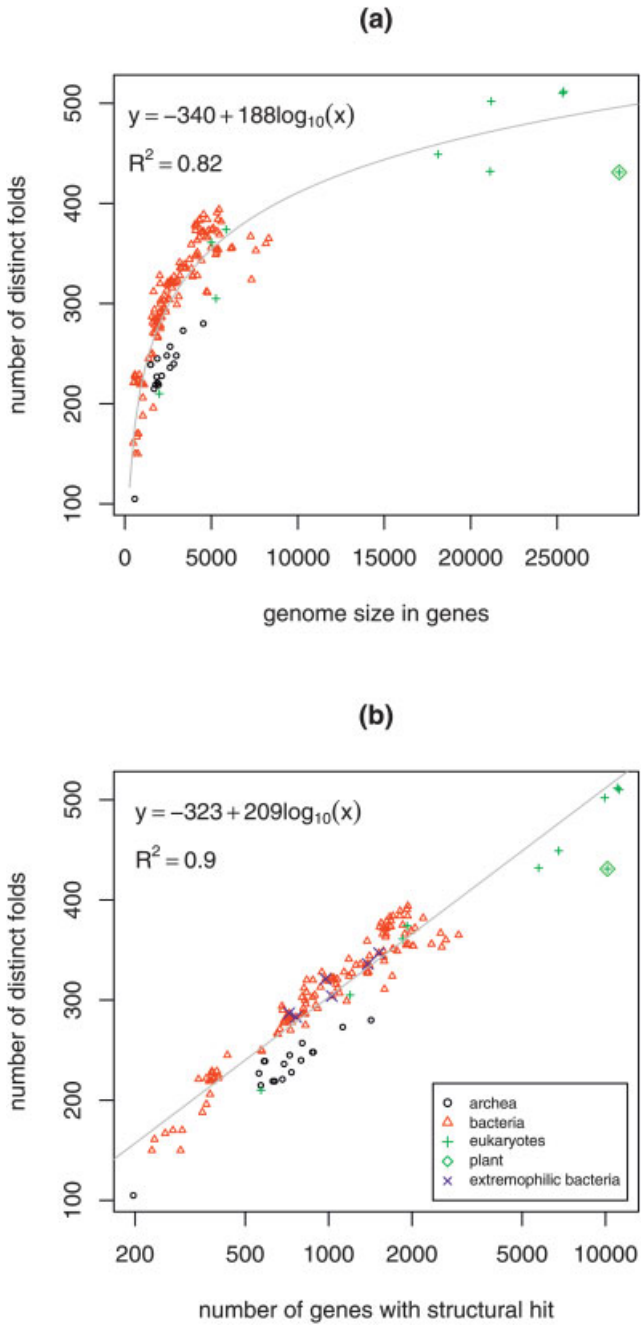

Fig. 1. Genome size in genes against number of distinct folds (**a**) linear-linear plot (**b**) log-linear plot with genome size replaced by the number of genes with a structural hit. Archaeal genomes are represented by circles, bacterial by triangles, and eukaryotes by pluses. The grey line indicates the best fitted logarithmic approximation. Note that archaeal and eukaryotic genomes lie below the bacterial genomes, indicating that archaea and eukaryotes have relatively fewer distinct folds per genome than bacteria. The diamond represents the only plant, *Arabidopsis thaliana*, and the crosses represent the bacterial extremophiles: *Bacillus halodurans*, *Oceanobacillus iheyensis*, *Thermoanaerobacter tengcongensis*, *Thermosynechococcus elongatus*, and *Thermotoga maritima*.

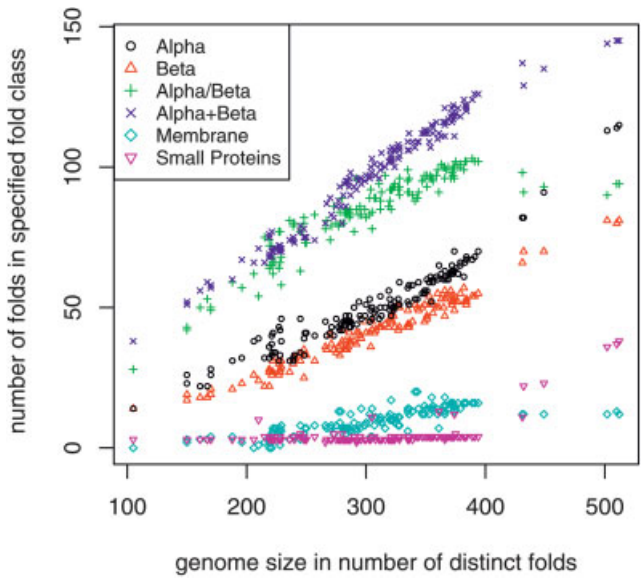

Fig. 2. The number of distinct folds on a genome versus the number of distinct folds on a genome per fold class. All alpha folds are represented by circles, all beta folds by triangles, alpha/beta folds by pluses, alpha+beta folds by crosses, membrane protein folds by diamonds, and small protein folds by downward pointing triangles. Note that alpha/beta folds have relatively fewer folds on large genomes, and that all alpha and small protein folds have relatively more folds on larger genomes.

fewer distinct folds. Archaea are known to have relatively few unique folds, i.e., folds only occurring in the archaeal kingdom, compared to bacteria and eukaryotes;[3,6] our data showed similar results. This might suggest that it is hard for archaeal proteins to diverge into other folds due to extreme living environments. Note that the PDB contains fewer structures for archaeal proteins, than for eukaryotes and bacteria, which might result in finding fewer folds unique to archaea and possibly fewer distinct folds on archaea.

Most proteins consist of at least two domains[33] and eukaryotes appear to have more multi-domain proteins than archaea or bacteria.[7] Since we compare the number of genes on a genome with the number of distinct folds, rather than the total number of domains on a genome, the proportion of multi-domain proteins might influence our results. For eukaryotes this could mean that fewer distinct folds per gene are observed because we found structural hits for a smaller proportion of all domains on eukaryotes.

Fold Classes

The fold classes defined by SCOP each relate differently to genome size (Fig. 2). The number of distinct all-alpha and small protein folds grows relatively faster on large genomes, i.e., there seems to be more room for developing new folds in these classes on the larger genomes. Small proteins usually do not have a hydrophobic core, which means they need either metal ions or disulphide bridges to create a stable fold. On the other hand the number of distinct alpha/beta folds seems to reach a maximum after which despite increasing genome size, no further increase in distinct folds is seen (Fig. 2). This might indicate that there are fewer opportunities to create novel folds or superfamilies within this class. Changes in behavior on larger genomes seem specific to eukaryotes. Note that no conclusions should be drawn for membrane proteins since

they are underrepresented in the PDB. This should not affect the general results significantly, since only 30 out of the 695 folds found were membrane proteins. Very similar patterns are found for the alpha/beta class, the alpha class and the small proteins class, when we compare superfamilies instead of folds with genome size.

Distribution of Fold Copies

The distribution of copies of a fold per single genome has been described to follow a power law.[9] We looked if this distribution depends on the structural class or kingdom. Simply binning the data for a power law will give bad estimates at low frequencies and does not make use of all available data. Therefore we plotted the number of copies of a fold against the folds descending in rank according to their number of copies (see methods); this method will also show as a straight line on a log-log plot, if the underlying distribution is a power law. In order to use the data from all genomes the average was taken only over those genomes a fold occurs on.

The average number of copies of a fold per genome does appear to follow a power law distribution for archaea and bacteria (Fig. 3). Comparing the different fold classes, we can see that the slope of the alpha/beta class in bacteria is more gradual than that of the other fold classes, implying bacteria have relatively more copies of alpha/beta folds than of other classes [Fig. 3 (a)]. This effect is weaker in archaea and not seen in eukaryotes.

We also found that the distribution of copies of a fold in eukaryotes does not appear to follow a clear power law: the line seems bent [Fig. 3(b)]. This might be caused by the ability of eukaryotes to have longer genomes, allowing genes in eukaryotes to duplicate with less restriction than in either bacteria or archaea. If this is the case, than the limiting genome size possibly creates the power law distribution. This would argue in favor of the model created by Karev et al., where a power law is obtained by restricting the genome size,[26] as opposed to the model by Qian et al. where genes are allowed to duplicate freely.[9] The results look very similar (not shown) when calculated for superfamilies rather than folds.

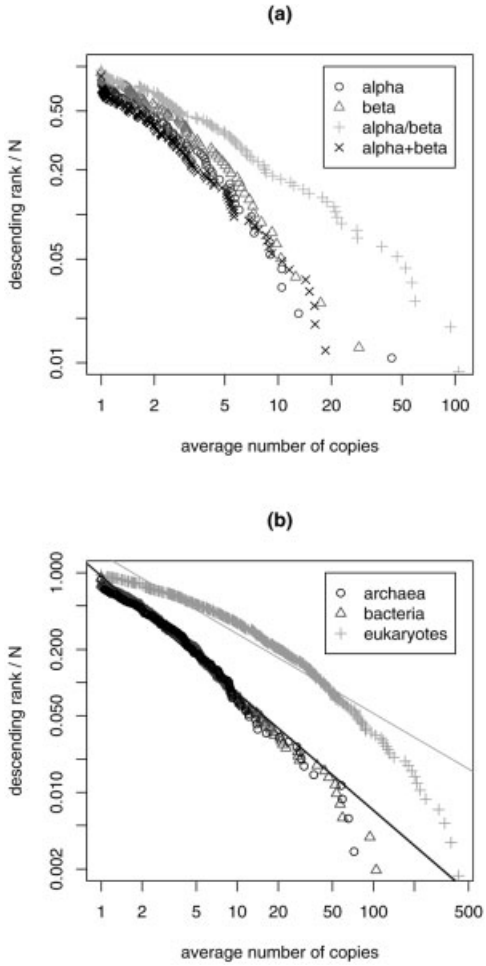

Fig. 3. The descending rank of the average number of copies per genome (**a**) in bacteria for different fold classes (**b**) in different kingdoms. In (**a**) all alpha folds are represented by circles, all beta folds by triangles, alpha/beta folds by pluses and alpha+beta folds by crosses. In (**b**) copies for folds on archaea are represented by circles, bacteria by triangles and eukaryotes by pluses. The lines indicate the best fitted power law approximations to the data from all genomes. The descending rank seems considerably lower than the linear regression line for folds with a high number of copies. This might be caused by a lack of data in this region. Alternatively the power law distribution might only hold for the lower range of copies. Note that alpha/beta folds have many more copies of a fold per bacterial genome than the other fold classes. Eukaryotes have more copies of folds and the fit is worse than for archaea and bacteria, indicating perhaps a different distribution trend.

Distribution of (Super) Families per Fold

In order to compare the number of families under a fold with the other variables, only families occurring on a fixed set of genomes are used, e.g., on archaea, on bacteria, on eukaryotes or on all 157 genomes. The number of superfamilies or families under a fold as well as the number of families under a superfamily seem to follow power law behavior within a fixed set of genomes. It has previously been suggested that a power law is an overly simplistic approximation of the distribution of families under a fold.[19–21] However looking at our data using descending rank there is a straight line on a log–log plot (Fig. 4), similar to the descending rank for fold copies. It is important to note that the power law-like distribution might be caused by the classification system rather than a biological process. It is probable that in developing a classification system like SCOP, it is easier to classify a superfamily under a fold with more superfamilies, since there are more superfamilies with which similarity can be observed. This process of the rich get richer can create a power law distribution.[34] The power law does not appear to hold for folds with only one superfamily [Fig. 4(a)], which are overrepresented. This group may correspond to the unifolds as described by Coulson and Moult.[19] Unifolds contain only one sequence family per fold. In our data these folds contain a single superfamily an analogous outcrop is not seen for family per fold or family per superfamily [Fig. 4(b,c)]. Alternatively, the overrepresentation of folds with only one superfamily may be caused by a bias in SCOP, or a biological process (see discussion in section: Folds with one Superfamily and Many Copies, below).

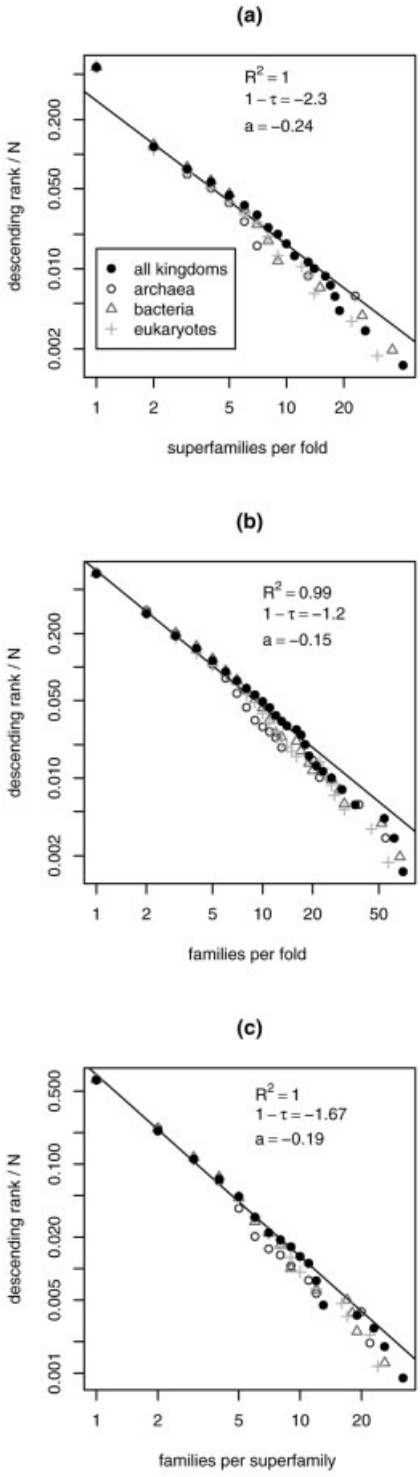

Fig. 4. Descending rank plots with the lines indicating the best fitted power law approximations to the data from all genomes: (**a**) number of superfamilies per fold; (**b**) number of families per fold; (**c**) number of families per superfamily. Folds on all genomes are represented by closed circles, archaea by open circles, bacteria by triangles and eukaryotes by pluses. There are more folds only containing one superfamily than expected by a power law distribution. In this region we would expect the data to be most accurate, since there are many observations. We disregarded these folds for the approximation of a power law by linear regression.

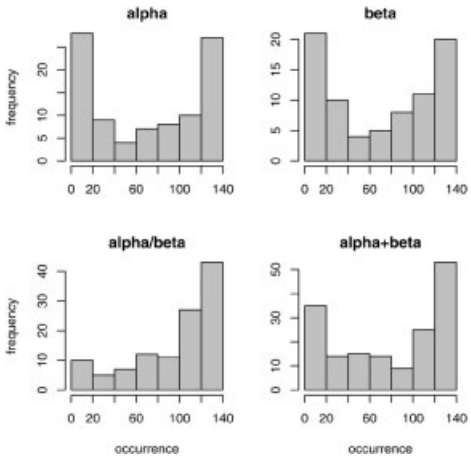

Fig. 5. Histograms for genomic occurrence on bacteria for the different fold classes. In general we can see that there is a peak for folds occurring on all genomes and a peak for folds occurring on only a few genomes. For the alpha/beta class the latter peak is not observed.

The power laws on different levels, i.e., family, superfamily, and fold [e.g., Fig. 4(a–c)], may be caused by the property of a power law distribution to be scale free. This is a fractal property implying that the level of detail in a distribution does not change the overall distribution. This might also explain why power laws are seen both for copies of a fold, as well as for a superfamily and also for sequence families.[35] It could also account for the very similar results for distinct folds and superfamilies versus genome size. However, truly scale-free behavior should also result in identical powers for the power law distribution, which is not observed in our results.

Distribution of Fold Occurrence

If we now consider the distribution of occurrence of distinct folds across the genomes we do not observe a power law, as we did for the distribution of number of copies on a genome and the number of families per fold (Fig. 5). For most fold classes and kingdoms, distribution is characterized by a peak for folds occurring at very few genomes, and another peak for folds occurring on many genomes. This distribution pattern can be explained by combining two sets of folds (1) relatively old folds occurring on most genomes except for some deletions and (2) relatively new folds occurring on only a few genomes, with an occurrence depending on how recently it has evolved. When we compared the distribution of different fold classes and kingdoms, we found that the height of the two peaks in the distribution changes strongly with kingdom, fold class. Note that the relation between the age of a fold and the number of genomes it occurs on is not strict (see discussion in section: Folds Occurring in All Kingdoms, below). Comparing the distributions for the three kingdoms of life, for eukaryotes the peak for folds occurring at very few genomes appears to be much lower (except for all alpha folds). Although this might be due to the small number of eukaryotic genomes (10) in our study, more recent evolution of eukaryotes is also a likely explanation.

Comparing the distributions of different fold classes separately, we see that the alpha/beta class has relatively more folds occurring on all genomes in each kingdom (Fig. 5); the distribution only has a peak at folds occurring on many genomes and no peak for folds occurring on only a few genomes. This effect is seen for all kingdoms. This might indicate an older age or fewer opportunities to create novel folds or superfamilies within this class. Either of these explanations could be supported by the results for the number of distinct folds versus genome size (Fig. 1), where it appears that the number of distinct alpha/beta folds is not able to keep increasing after a certain genome size.

Relations between Fold Occurrence, Superfamilies per Fold, and Fold Copies

So far we have discussed the distribution for our measures of fold usage, now we will investigate if these measures are related. We would probably expect that the number of copies of a fold per genome increases if it occurs on more genomes. Similarly we would expect the number of superfamilies per fold to increase with occurrence and copies on assumption of divergent evolution: a superfamily has more chance to diverge into a new superfamily, if there exist more copies of a fold. However the results below show that there are no clear relations between these measures.

Figure 6(a) shows the relation between fold occurrence and fold copies. The relation appears to be restrictive, such that a fold with a low number of occurrences can not have many copies on a genome. This could be explained if we consider occurrence across the genomes as an indication for the age of a fold: only older folds, i.e., folds occurring on many genomes, will have had the time to duplicate significantly. However, Figure 6(a) also shows that folds occurring in all genomes do not necessarily have many copies. The age of a fold seems to create an upper limit for the number of copies of a fold per genome but not a lower limit. This indicates that the number of copies on a genome alone might not be a good estimator for the age of a fold. Within the restriction of time the distribution of copies again seem to follow a power law. Overall the number of copies on a genome appears to be more self dependent than related to the fold occurrence. From the power law distribution it follows that folds with many copies are more likely to duplicate again. In addition, if we assume that occurrence can give us an estimate for the age of a fold, the maximum number of copies for a fold is restricted by the time the fold has had to duplicate.

A similar restrictive relation seems to occur if we compare the number of superfamilies per fold with fold occurrence across genomes [Fig. 6(b)]. We could again argue that time limits the number of new superfamilies a fold can create. However, the observed relation might also be caused by the grouping of superfamilies into folds. The larger the group size the higher the chance that there exists a superfamily in the group which has a high number of occurrences. Indeed, if we check for this by randomly grouping superfamilies into folds we get similar results. This would probably still indicate that the number of superfamilies per fold in general is self dependent according to the power law distribution, so that a fold with many superfamilies is more likely to gain, either in SCOP or evolutionarily (see discussion in section: Distribution of (Super) Families per Fold, above), a new superfamily than a fold with only one superfamily. The number of superfamilies per fold compared to the number of copies gives the

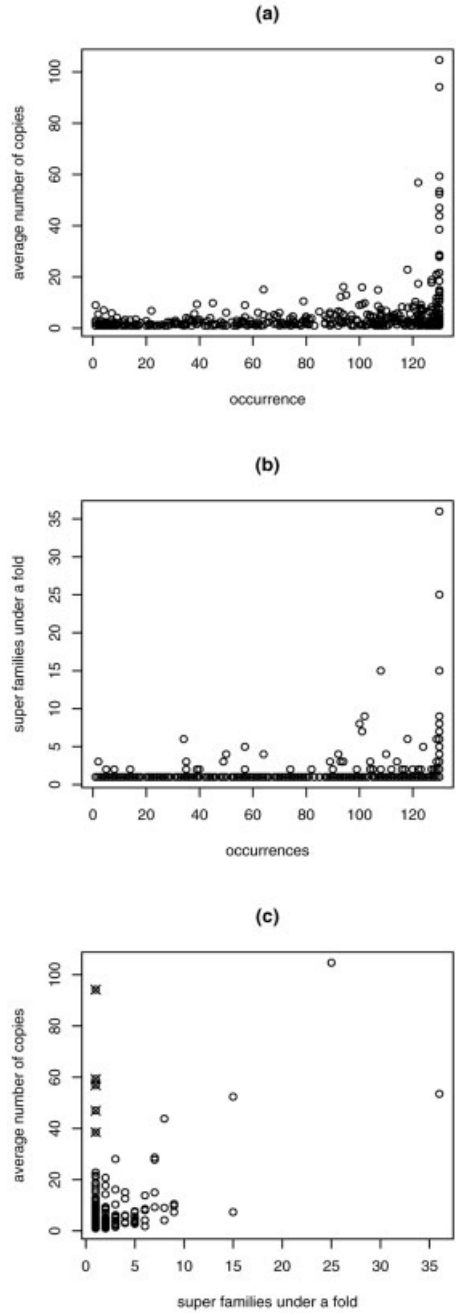

Fig. 6. Relations between measures of fold usage: (**a**) fold occurrence versus the total number of copies divided by the number of occurrences; (**b**) fold occurrence versus the number of superfamilies per fold; (**c**) number of superfamilies per fold versus the total number of copies divided by the number of occurrences. All three figures show data from bacterial genomes only, to provide a more uniform set of genomes. The crosses indicate folds with many duplications and only one superfamily.

TABLE I. Folds with Many Copies, but Only one Superfamily†

SCOP	Fold	CATH	Number of families	Average copies			Occurrence		
				arc	bac	euk	arc	bac	euk
c.2	NAD(P) binding Rossmann fold	3.40.50.720	10	36	59	128	17	130	10
c.37	Ploop containing nucleoside triphosphate hydrolases	3.40.50.300	19	59	94	371	17	130	10
c.66	Sadenosy lLmethionine dependent methyltransferases	3.40.50.150	23	58	47	64	17	130	10
c.67	PLP dependent transferases	3.40.640.10	6	27	39	53	16	130	10
c.69	Alpha/beta Hydrolases	3.40.50.1820	26	30	57	201	13	122	10

†The columns show the SCOP identifier, the name of the fold as defined in SCOP, the CATH identifier, the number of families under the fold, the number of copies for each kingdom, the number of genomes the fold occurs on for each kingdom.

result we would expect when two power law distributions without a relation are plotted against each other [Fig. 6(c)], as before very similar results are shown if we group superfamilies randomly. We would probably expect a higher chance of forming a new superfamily if more copies of a fold are present. It seems, once again, that the number of superfamilies on a fold is more self dependent than related to the number of copies. However, there might be small unseen effects due to the grouping of superfamilies.

All results are very similar if we look for relations between copies, occurrence, and families in a superfamily. At this level evolutionary relations between genes clustered together in a superfamily are more reliable, arguing that most of the effects described above are real.

Folds with One Superfamily and Many Copies

We noticed that there are more folds with only one superfamily, but many copies, than randomly expected [see crosses in Fig. 6(c)]. This is not seen when we compare the number of copies of a superfamily with the number of families per superfamily. The existence of these folds might correlate with the over-representation of folds with only one superfamily compared to the general power law distribution for superfamilies per fold [Fig. 4(a)]. Table I shows that these folds with only one superfamily and many copies occur in all three kingdoms, and are therefore likely to be old. All belong to the alpha/beta class, and have many sequence families under a single superfamily. If we compare the classification for these folds between SCOP and CATH, we can see that CATH classifies all these domains under the same architecture [3Layer (aba) Sandwich]. All but one domain are also classified under the same topology (Rossmann fold, Table I). However the Rossmann fold as defined in CATH has 105 homologous superfamilies under it. Within SCOP there are no folds with such a high number of superfamilies (Fig. 4).

In an analysis of SCOP domains and EC numbers,[36] George et al. found four superfamilies (SCOP codes: c.2.1, c.37.1, c.67.1 and c.69.1) that have more different enzymatic functions assigned to them than other superfamilies.[37] We see that all these four superfamilies are also found in our list of folds with only one superfamily and many copies.

This effect could be caused by a bias in SCOP. If relatively more evolutionary relations between sequence families were spotted within these folds than in other folds, it would create a single superfamily with a high number of sequence families; this could also explain why so many different EC numbers are assigned to these superfamilies. Here we assume that many other folds have evolutionarily related sequence families, which are not (yet) classified in the same superfamily, i.e., divergent evolution. We could also reason that these folds are "trapped" within a certain superfamily, i.e., they are unable to diverge away from this superfamily into another superfamily under the same fold; the similar topology of these folds might lead us to believe that these fold types are evolutionarily more conserved. This reasoning would also assume that most superfamilies under a fold are divergently related. Note that the two explanations presented above do not exclude each other: if the structure of these folds is particularly conserved, it might be easier to spot evolutionary relations.

Folds Occurring in All Kingdoms

Fold occurrence alone is not an accurate measure for the age of a fold. However, folds that are found on genomes in each of the three kingdoms are likely to be older than the split of the kingdoms. Table II shows there are in total 303 folds occurring in all three kingdoms, which is 44% of the total number of folds described in SCOP. Some of these folds might have been introduced into a kingdom by lateral gene transfer, so are likely to occur only on very few genomes in a kingdom. To reduce the chance of lateral gene transfer affecting our results, we also looked for folds occurring on at least half of the genomes in each kingdom. This still estimates that around 27% of all folds existed on the last common ancestor between the three kingdoms. These results could be caused by convergent evolution of two superfamilies within the same fold. However, most of the folds which occur on many genomes are due to one of the superfamilies in this fold occurring on multiple genomes. There are only a few cases where several superfamilies each occur on a few genomes leading to an overall high occurrence for that fold.

Only 6% of folds occur on all genomes. Hence there are a large number of folds that occur in all kingdoms but are not detected on all genomes. This could be caused by (1) high divergence in sequence similarity, or unknown superfamilies in a fold, which makes the fold on certain genomes

TABLE II. Numbers of Old Folds and Old Superfamilies in the Different Fold Classes[†]

	Occurrence on genomes in all kingdoms					
	All		> 50%		> 1	
Folds	Number	%	Number	%	Number	%
Class						
All alpha	5	3	26	17	44	28
All beta	6	6	27	27	48	48
Alpha/beta	13	11	67	57	91	78
Alpha + beta	18	9	55	26	88	42
Multi domain	2	6	9	26	18	53
Membrane	0	0	3	10	10	33
Small	1	2	3	7	4	9
Total	**45**	**6**	**190**	**27**	**303**	**44**
Superfamilies						
All alpha	4	2	28	11	56	22
All beta	6	3	29	14	76	38
Alpha/beta	15	8	91	47	136	71
Alpha + beta	15	5	73	24	129	42
Multi domain	2	6	9	26	18	53
Membrane	0	0	2	4	11	20
Small	1	2	3	5	7	11
Total	**43**	**4**	**235**	**21**	**433**	**39**

[†]Occurrence on all genomes, on more than half of the genomes in each kingdom and on more than one genome in each kingdom is shown.

undetectable by sequence alignment; (2) lateral gene transfer; (3) convergent evolution, although similar results are shown for superfamilies; or (4) deletion of a fold on a genome. Since we do not detect folds, which are likely to be old, on all genomes, the age of a fold might not directly be predicted by the number of genomes on which a fold occurs. It is possible that better searching methods for detection make this a more reliable measure. On the other hand, deletions have been observed at gene level, particularly in bacteria.[38] It is therefore likely that deletions also occur at fold level, so that folds are not detected on some genomes because they have been deleted during evolution.

The alpha/beta class has relatively more folds which occur in all kingdoms than folds from other fold classes. It is possible that 78% of the folds in the alpha/beta class might originate from before the split of the kingdoms, and a more conservative estimate still shows that half of the alpha/beta folds are probably older than the split between kingdoms (Table II). All alpha folds and small proteins show the opposite result occurring relatively infrequently in all three kingdoms. This might be caused by deletion of the folds early on in evolution or by not being able to detect these folds on some genomes. It is more likely however that many small protein and all alpha folds evolved later.

There are a number of folds and superfamilies that occur on all genomes in our set and have only very few copies (Table III). These might be of particular interest as they may have important biological functions and therefore evolutionary mechanisms may prevent these folds from being deleted on a genome. Furthermore, they may have stayed particularly close in sequence, either by structural or functional constraints, allowing them to be detected more easily than folds which have diverged more. Many of these superfamilies are ribosomal proteins or are involved in the translation process. This agrees with results from Ranea et al., showing that most of the superfamilies which do not increase their number of copies with genome size are in the functional category of translation, ribosomal structure and biogenesis.[27]

SUPERFAMILY Data

In all of the analysis above we used PSI-BLAST for all assignments; we carried out a similar analysis using assignments from the SUPERFAMILY database.[13] These assignments have higher coverage on a more diverse set of genomes than the PSI-BLAST assignments (see section: SUPERFAMILY Assignments, in Methods). On this larger data set, we found that the general trends for the distributions and relations are unaltered, except for the following cases: It becomes more evident that the number of distinct alpha/beta folds does not increase on larger/eukaryotic genomes. Comparing the number of distinct folds with genomes size we still observe that archaea have relatively fewer distinct folds than bacteria. However eukaryotes appear more similar to bacteria, using the SUPERFAMILY assignments. This might have been caused by a better coverage of the genomes by the SUPERFAMILY assignments, since Figure 1(b) was not normalized for sequence coverage, and eukaryotes and bacteria are likely to have a different proportion of multi-domain proteins (see section: Genomes, above).

Conclusions

We have shown that the fold usage on genomes can give us an idea of the mechanisms behind protein fold evolution. In particular it can give us an indication of the evolution of folds in different kingdoms or in different fold classes.

Comparing different kingdoms we can conclude that eukaryotes and archaea have relatively fewer distinct folds on their genomes than bacteria. In eukaryotes this might be explained by a higher content of multi-domain proteins; for archaea this might be caused by their extreme living environments making it harder for folds to evolve into other stable folds. Furthermore eukaryotes appear to have a different distribution for the number of copies of a fold per genome than archaea and bacteria, which both seem to follow a power law.

The structural class alpha/beta behaves differently from the other fold classes in several ways. The number of distinct folds of the alpha/beta class does not increase with genome size on eukaryotes. The number of copies on bacteria is much higher than those for other fold classes. Many alpha/beta folds occur on nearly all genomes. These results indicate that perhaps many alpha/beta folds are very old. Furthermore as many as 78% of the alpha/beta folds might have originated from before the split of kingdoms. The fact that the number of distinct alpha/beta folds does not increase in eukaryotes in line with other fold classes might also indicate that this fold class has explored

TABLE III. Superfamilies Which Occur on All Genomes in Our Set, with on Average Less Than Two Copies per Genome in at Least one Kingdom[†]

		Number of families			Average copies		
SCOP id	Superfamily	arc	bac	euk	arc	bac	euk
a.60.7	5′ to 3′ exonuclease, Cterminal subdomain	1	1	1	1	1.3	2.8
a.75.1	Ribosomal protein S7	1	1	1	1	1	2.3
a.156.1	S13like H2TH domain	2	2	2	1.8	2	5.1
b.39.1	Ribosomal protein L14	1	1	1	1	1	3.2
b.44.1	elF2gamma Cterminal domain	1	1	1	2.4	1.5	11.4
c.12.1	Ribosomal proteins L15p and L18e	1	1	1	1.9	1	6.1
c.20.1	Initiation factor IF2/elF5b, domain 3	1	1	1	1	1	2.5
c.23.15	Ribosomal protein S2	1	1	1	1	1	6.6
c.55.4	Translational machinery components	2	1	2	3.1	2	7.6
d.41.4	Ribosomal protein L10e	1	1	1	1	1	5.3
d.50.1	dsRNAbinding domainlike (ribosomal S5)	2	2	3	1.1	2	17.5
d.53.1	Ribosomal protein S3 Cterminal domain	1	1	1	1	1	1.2
d.55.1	Ribosomal protein L22	1	1	1	1	1	7.5
d.58.11	EFG/eEF2 domains III and V	1	1	1	1	2.7	6.7
d.66.1	AlphaL RNAbinding motif (ribosomal S4)	3	4	3	1.5	5.2	4.2
d.74.3	RBP11 like subunits of RNA polymerase	1	1	2	1	1	3.5
d.77.1	Ribosomal protein L5	1	1	1	1	1	3
d.131.1	DNA clamp	1	1	1	1.5	1.1	2.1
d.140.1	Ribosomal protein S8	1	1	1	1	1	3.6
d.141.1	Ribosomal protein L6	1	1	1	1	1	5.3
e.24.1	Ribosomal protein L1	1	1	1	1	1	5.7
g.41.3	Zinc betaribbon	1	2	1	2.8	1.4	6.6

[†]The columns show the SCOP identifier for the superfamily, the name of the superfamily, the number of families under the superfamily for each kingdom, the average number of copies on a genome for each kingdom.

most of its fold space. This could be caused by a more restrictive definition of the topology for this class.

On the other hand the number of distinct folds for small proteins and all alpha folds is shown to increase relatively fast on the larger genomes and a relatively smaller number of these folds occurs in all three kingdoms. This could indicate that most of these folds appeared more recently in evolution.

It will remain difficult to prove that the distributions for copies and families per fold are true power laws. However, power law behaviour where "the rich get richer" seems likely for copies of a fold on genomes as well as the number of families per fold, although in the latter case it is possible that this effect is caused by the creation of a classification system, rather than by a biological mechanism. All results in our analyses look very similar at fold and superfamily level (except for occurrence distributions). The grouping of (super) families into a fold with a power law distribution might cause this and could be an important factor for creating the final results at fold level.

Assuming a power law like distribution for superfamilies per fold, we noticed that there are more folds with only one superfamily than expected. This is also visible when we compare the number of superfamilies for a fold with the number of copies. This could be caused either by trapped folds, i.e., folds which are unable to diverge from a single superfamily or by a bias in the available databases.

We observe no clear relationship between fold occurrence, copies, and the number of sequence families under a fold. There exist folds which occur on all genomes, with only one copy per genome on average. Similarly there exist folds with only one superfamily, but many copies of these superfamilies. Although more subtle relations might be hidden due to a grouping effect, it is clear that a high occurrence across genomes does not necessarily imply a high number of copies or a high number of superfamilies.

When estimating the age of a fold we have to be careful with deriving it solely from number of copies of a fold. In general a high number of copies does indicate an old age, apart from eukaryotes where it seems that some folds have started duplicating more recently. However there are several folds that have on average only one copy per genome, which do occur on almost all genomes, and are likely therefore to be old. The relation for occurrence versus copies and the distribution for occurrence suggest that occurrence might give a rough indication for the relative age of a fold. Thus, a combination of measures for fold usage (occurrence, copies, and families per fold) can possibly be used to obtain an indication for the relative age of a fold when we would consider the taxonomy of the genomes on which the folds occur.

Direct evolutionary relations between two folds or superfamilies are even more difficult to identify from these measures. We must remember when one fold or superfamily diverges into a new fold or superfamily, the number of copies are set back to one. The diverging event only happens on one genome, setting the occurrence for the newly created fold or superfamily also back to one. Hence

direct evolutionary information from copies or occurrence is erased by such an event. On the other hand it is shown to be possible to reconstruct phylogenetic relations between the species by patterns of fold occurrence,[39] and patterns of deletion on a genome can be important to correlate function or similar functional pathways of two proteins, as they will be likely to be deleted together on a genome.[5] However, similar patterns of deletion/occurrence will not in general indicate an evolutionary relation between the structures of two proteins.

ACKNOWLEDGMENTS

This work was supported by funding from the EPSRC and the Wellcome Trust. We thank Frank von Delft and Lynette Cole for critical comments and discussion.

REFERENCES

1. Lee D, Grant A, Buchan D, Orengo C. A structural perspective on genome evolution. Curr Opin Struct Biol 2003;13:359–369. doi: 10.1016/S0959–440X(03)00079-4
2. Hegyi H, Lin J, Greenbaum D, Gerstein M. Structural genomics analysis: Characteristics of atypical, common, and horizontally transferred folds. Proteins 2002;47:126–141. doi:10.1002/prot.10078
3. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. Genome Res 1999;9:17–26.
4. Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci 2001;10:1970–1979. doi:10.1101/ps.10101
5. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.
6. Caetano-Anolles G, Caetano-Anolles D. An evolutionarily structured universe of protein architecture. Genome Res 2003;13:1563–1571.
7. Apic G, Gough J, Teichmann S. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310:311–325. doi:10.1006/jmbi.2001.4776
8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman J. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
9. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 2001;313:673–681. doi:10.1006/jmbi.2001.5079
10. Wu J, Kasif S, DeLisi C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 2003;19:1524–1530. doi: 10.1093/bioinformatics/btg187
11. Berman H, Westbrook Z, Feng J, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
12. Muller A, MacCallum RM, Sternberg M. Benchmarking psi-blast in genome annotation. J Mol Biol 1999;293:1257–1271. doi:10.1006/jmbi.1999.3233
13. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 2002;30:268–272.
14. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
15. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
16. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
17. Koonin E, Wolf Y, Karev G. The structure of the protein universe and genome evolution. Nature 2002;420:218–223. doi:10.1038/nature01256
18. Deeds E, Shakhnovich B, Shakhnovich E. Proteomic traces of speciation. J Mol Biol 2004;336:695–706. doi:10.1016/j.jmb.2003.12.066
19. Coulson AF, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71. doi:10.1002/prot.10011
20. Liu X, Fan K, Wang W. The number of protein folds and their distribution over families in nature. Proteins 2004;54:491–499. doi:10.1002/prot.10514
21. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol 2000;299:897–905. doi:10.1006/jmbi.2000.3786
22. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305. doi:10.1006/jmbi.1998.2282
23. Govindarajan S, Recabarren R, Goldstein R. Estimating the total number of protein folds. Proteins 1999;35:408–414.
24. Crippen G, Maiorov V. How many protein folding motifs are there? J Mol Biol 1995;252:144–151. doi:10.1006/jmbi.1995.0481
25. Taylor W. A "periodic table" for protein structures. Nature 2002;416:657–660. doi:10.1038/416657a
26. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evol Biol 2002;2:18. doi: 10.1186/1471.2148.2.18
27. Ranea J, Buchan D, Thornton J, Orengo C. Evolution of protein superfamilies and bacterial genome size. J Mol Biol 2004;336:871–887. doi:10.1016/j.jmb.2003.12.044
28. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci USA 2004;101:3160–3165.
29. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Res 2000;28:254–256.
30. Adamic LA, Huberman BA. Zipf's law and the internet. Glottometrics 2002;3:143–150.
31. Troll G, beim Graben P. Zipf's law is not a consequence of the central limit theorem. Physical Review E 1998;57:1347–1355.
32. Danson M, Hough D. Structure, function and stability of enzymes from the archaea. Trends Microbiol 1998;6:307–314. doi:10.1016/S0966–842X(98)01316-X
33. Vogel C, Bashton M, Kerrison N, Chothia C, Teichmann S. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 2004;14:208–216. doi:10.1016/j.sbi.2004.03.011
34. Barabasi AL. Linked: the new science of Networks. New York: Perseus. 2002.
35. Huynen M, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 1998;15:583–589.
36. IUBMB. Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. 1992.
37. George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB. SCOPEC: a database of protein catalytic domains. Bioinformatics 2004;20 Suppl 1:I130–I136. doi: 10.1093/bioinformatics/bth948
38. Mira A, Ochman H, Moran N. Deletional bias and the evolution of bacterial genomes. Trends Genet 2001;17:589–596. doi:10.1016/S0168–9525(01)02447-7
39. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res 2000;10:808–818.