

Protein fold and family occurrence in genomes: power-law behaviour and evolutionary model

Running title: Power-law behaviour and evolutionary model

Jiang Qian^{*}, Nicholas M Luscombe^{*} & Mark Gerstein[†]

*Department of Molecular Biophysics and Biochemistry,
Yale University, 266 Whitney Avenue, PO Box 208114,
New Haven CT 06520-8114, USA.*

mark.gerstein@yale.edu

^{*}These authors contributed equally to this work

[†]To whom correspondence should be addressed

*Brief communication to J Mol Biol
Revised manuscript 1 September 2001*

Summary

Global surveys of genomes measure the usage of essential molecular parts – defined here as protein families, superfamilies or folds – in different organisms. Based on surveys of the first 20 completely sequenced genomes, we observe that the occurrence of these parts follows a power-law distribution. That is, the number of distinct parts (F) with a given genomic occurrence (V) decays as $F = aV^b$, with a few parts occurring many times and most occurring infrequently. For a given organism, the distributions of families, superfamilies and folds are nearly identical, and this is reflected in the size of the decay exponent b . Moreover, the exponent varies between different organisms, with those of smaller genomes displaying a steeper decay (*i.e.* larger b). Clearly, the power law indicates a preference to duplicate genes that encode for molecular parts which are already common. Here, we present a minimal, but biologically meaningful model that accurately describes the observed power law. Although the model performs equally well for all three protein classes, we focus on the occurrence of folds in preference to families and superfamilies. This is because folds are comparatively insensitive to the effects of point mutations that can cause a member protein to diverge beyond detectable similarity. In the model, genomes evolve through two basic operations: (i) duplication of existing genes and (ii) net flow of new genes. The flow term is closely related to the exponent b and can accommodate considerable gene loss; however, it reproduces the observed data best with a net inflow, *i.e.* with more gene gain than loss. Moreover, we show that prokaryotes have much higher rates of gene acquisition than eukaryotes, probably reflecting lateral transfer. A further natural

outcome from our model is an estimation of the fold composition of the initial genome, which potentially relates to the common ancestor for modern organisms. Supplementary material pertaining to this work is available from www.partslist.org/powerlaw.

Keywords: bioinformatics, genomics, proteomics, evolution, protein folds, protein superfamilies, protein families, power law

Introduction

The power-law behaviour is frequently observed in different population distributions.

Also known as Zipf's law, it was first widely recognised for word usage in text documents¹. By grouping words that occur in similar amounts, Zipf observed that a small number of words such as "the" and "of" are used many times, while most are used infrequently. When the size of each group is plotted against its usage, the distribution follows a power-law function. That is, the number of words (F) with a given occurrence (V) decays according to the equation $F = aV^b$ – a distribution that has a linear appearance when plotted on double-logarithmic axes. Mandelbrot² suggested that the observation is connected with the hierarchical structure of natural languages and further work by Zipf described the behaviour for the relative sizes of cities, income levels, and the number of papers per scientist in a field¹.

Subsequent to this, power laws have also been reported for the construction of large networks. Examples include social interactions³, the World Wide Web^{4,5}, metabolic pathways⁶, and the intermolecular interactions of different proteins⁷. Barabasi and Albert⁸ proposed that this results from self-organising development, in which new vertices preferentially attach to sites that are already well connected.

Interestingly, the power-law behaviour is intrinsic to the usage of short nucleotide sequences⁹ and the populations of gene families in genomes^{10,11}. Here we report this for the occurrence of protein superfamilies and folds also. However, despite these observations in genomic biology, there have been few attempts to explain them. Using analogies to text documents, Mantegna and co-workers⁹ implied that the behaviour

originated from similarities of DNA sequences with natural languages. We argue this to be highly unlikely, and instead propose a simple, biologically reasonable model of evolution. Furthermore, although the sequence of any individual genome provides only a snapshot of evolutionary time, constructing such a model increases our understanding of how it arrived at the current state.

Protein family, superfamily, and fold occurrence in genomes

Most proteins that are encoded in a genome can be grouped according to their similarity in three-dimensional structure or amino acid sequence. Here we consider the fold, superfamily and family taxonomies that provide a hierarchical classification of proteins¹²; each class is a subset of the one before, and proteins are grouped with increasing similarity between them. First, proteins are defined to have a common fold if their secondary structural elements occupy the same spatial arrangement and have the same topological connections. Next, proteins are grouped into the same superfamily if they share the same fold, and are deemed to share a common evolutionary origin, for example owing to a similar protein function. Both the fold and superfamily classes aim to group proteins that are structurally related, but whose similarities cannot necessarily be detected only by their sequences. Finally, proteins are grouped into the same family if their amino acid sequences are considered to be similar, most commonly by their percentage sequence identities or using an *E*-value cut-off. Alternatively they can also be characterised by the presence of a particular sequence “signatures” or “motifs”. In this study we have used the fold and superfamily classifications from SCOP¹² and the family classifications from InterPro^{13,14}; the three will be collectively termed as molecular “parts”.

One way to represent the contents of a genome is to count the number of times that each part occurs and then group together those with similar occurrences (Figure 1a). In ranking parts by their occurrence, it is clear that for all organisms most occur just once, some occur several times and a few are found many times^{10,15-17}. For example, the 229 folds assigned in the *E. coli* genome have an average occurrence of seven: 72 (31.4%) folds are found just once, 45 (19.7%) are found twice, and only 10 (4.4%) occur more than 30 times. The most common, the TIM-barrel fold, occurs 93 times in the genome. Similar observations can be made for families: 303 (23.6%) occur once, 222 (17.3%) occur twice and the most common, the AAA-ATPase proteins, occurs 86 times. Previous studies have shown these distributions to be an actual feature of genomes, rather than a result of bias in the contents of classification databases¹⁸.

Significantly, this relationship is described by a power-law, in which the number of parts (F) with a certain genomic occurrence (V) decays with the equation $F = aV^b$. The distribution has a linear appearance when plot on a log-log graph, where $-b$ defines the slope. The power-law function that produces the smallest residual has the best fit with the genomic data (Figure 1b). Figure 2 shows the distributions of families, superfamilies and folds for *E. coli* and *S. cerevisiae*, and plots for the 18 other organisms are available in the supplementary website. We note that the overall distributions of all three classifications are very similar within each genome, and share near-identical gradients. In general, the smaller the genome, the steeper the gradient: the two eukaryotes have shallow gradients ($b = 0.9-1.2$) while the prokaryotes with smaller genomes have steep gradients ($b = 1.2-1.8$).

An evolutionary model

Given a mathematical description that is common to all organisms, we can simulate the observed distributions. Our model is based on an evolutionary process in which genomes start from a small size with a limited number of genes and grow to their current states by gaining new ones. The main sources for new genes are from duplication of existing ones, and introduction of completely novel genes via lateral transfer from other organisms or *ab initio* creation.

Although our model would apply equally well for any of the three protein classifications (families, superfamilies and folds), we have decided to define the relationships between different genes by their folds. This is mainly because of how proteins are grouped into families or superfamilies. With families, as proteins accumulate mutations over time, they can diverge too much for their common origin to be detected by sequence comparison. Thus individual proteins give the impression of “breaking away” to form a new group during the course of evolution. Superfamily classifications, particularly those in SCOP, try to address this issue by grouping together all proteins that have diverged from a common ancestor, even if they have diverged beyond detectable sequence similarity. These more distant groupings usually depend on proteins sharing distinct structural features, for example, the same active site location in the fold. However, much of the classification is subjective and does not involve the application of uniform thresholds. In contrast, folds do not suffer from these drawbacks. Genes can accumulate significant changes to their sequences without affecting their fold classification¹⁹. Furthermore, although there is the danger that two proteins have adopted similar

structures through convergent evolution, membership to a fold class can be determined objectively by applying clear thresholds to structural comparison algorithms²⁰⁻²². Therefore we focus on folds as a taxonomy that groups the most divergent collection of genes that we consider to have evolved from a common ancestor. In practise though, given the similarity in the genomic distributions of families, superfamilies and folds, the choice of classification will not greatly affect the simulations.

Consider a genome that originally comprises N_0 folds, with only one copy of each, *i.e.* the number of genes equals the number of folds. The genome grows by a step-wise duplication event, randomly selecting a gene. The probability of fold duplication is proportional to its current occurrence in the genome; so all folds initially have equal probabilities for duplication, $1/N_0$. After the first time-step, the duplicated fold occurs twice and has a new probability for duplication $2/(N_0+1)$, while the remaining folds have a decreased probability $1/(N_0+1)$. At each time-step, we also consider the introduction of new folds and the deletion of existing ones. The rate of fold flow, r (= fold acquisition – fold deletion), is measured as the ratio with respect to duplication events. Here we consider the average rate over the entire course of evolution and the parameter remains constant throughout the simulation. We assume that multiple copies of a given fold arise only through duplication. Thus, the behaviour of this simple model is governed by three parameters: the initial size of the genome (N_0), the number of duplication events or generations (t), and the rate (r) of fold flow per generation. A schematic of the model is shown in Figure 1c.

In effect, the model follows the development of a single genome from the root of the phylogenetic tree to the branch that defines the organism (Figure 1d). Therefore, although the simulation may appear to treat each genome independently, we can imagine that the evolution of different organisms relate by their divergence from a common ancestor.

Gene loss is also a major factor during evolution and we could easily include this as an explicit parameter in our model. However we are reluctant to do so, because introducing another free parameter would seriously compromise our comparisons of the model against the observed genomic data. Instead gene loss is incorporated within r and t . If a fold only occurs once, then a deletion would remove it from the genome, thus giving a smaller r . Here, positive values of r signify a net inflow of folds and negative values indicate an overall loss. Alternatively, if a fold has multiple copies, deleting a gene would simply reduce its occurrence, which effectively amounts to stepping back in the simulation or “unduplicating” the gene.

The behaviour of the model

Different parameter values result in distinct distributions. It is important to remember that these values correspond to averages for the whole evolutionary process and are fixed for each simulation. (i) When $r \leq 0$, the distribution never approaches a power law even after a large number of generations (t). This eliminates the possibility of a negative r as the actual genomic data follows the power law. Furthermore, such a process would leave the genome without any folds that have single occurrences. Thus henceforth, we will refer to fold flow as fold acquisition. (ii) For a fixed $r > 0$ and N_0 , the distribution is exponential for small numbers of generations (t), but converts to a power law for large values of t (Figure 3a). (iii) Provided that r and t meet the requirements, a power-law distribution can be obtained for any initial genome size, $N_0 > 0$. (iv) Once the simulation reaches the power-law phase, the gradient (b) of the distribution stays constant with continued increments in t . Therefore the slope is mostly attributed to the size of r .

The main observations are summarised in the phase diagram of Figure 3b, which plots the transition between the two distributions types for different r and t . While the conversion is gradual, distributions are exponential below the transition boundary and power law above. We note that the threshold t -value required for the power-law phase is inversely correlated to r , so smaller numbers of generations are required for higher rates of fold acquisition. We also highlight the rightward shift of the transition line for increased N_0 , so for a given rate r , larger initial genomes require more generations to reach a power law. In summary, we show that the dominant trend during the initial stage of evolution is towards an increase in the number of genes rather than a decrease and

that gene deletion is not a major effect. In addition, we find that exponential distributions result from a “sampling” process in which the existing folds are duplicated, but the power-law distribution derives from a sufficiently long “evolutionary” process in which folds are also periodically acquired.

Estimation of parameters

Having described the general features of the model, we now simulate the actual distributions that are observed in different genomes. The three parameters we just discussed are all related to the evolutionary history of an organism and appear unattainable. However, we can obtain estimates by looking at the current states of each genome, which provide “snapshots” of the evolutionary process.

The total number of folds (N_{folds}) in the genome at the end of a simulation is the sum of the initial number of folds (N_0) and the number of new folds (rt) acquired during the course of evolution, $N_{folds} = N_0 + rt$ (eq. 1). Next the total number of genes in the final genome (N_{genes}) is the sum of the initial number of genes (N_0), the number of new folds acquired (rt) and the number of duplicated genes (t), $N_{genes} = N_0 + rt + t$ (eq. 2). Finally, N_{genes} and N_{folds} can be linked by the average level of fold duplication, *i.e.* the average fold occurrence (C) in the genome,

$$C = \frac{N_{genes}}{N_{folds}} = \frac{N_0 + rt + t}{N_0 + rt} = \frac{N_{folds} + t}{N_{folds}} \quad (\text{eq. 3})$$

By rearrangement, we obtain the number of duplication events $t = (C-1)N_{folds}$ and the relationship between N_0 and r ,

$$r = \frac{N_{folds} - N_0}{(C-1)N_{folds}} \quad (\text{eq. 4})$$

For each genome, we obtain C by dividing the number of genes that have structural assignments with the number of folds that have been assigned ($C = N_{genes}^{\text{assigned}} / N_{folds}^{\text{assigned}}$, *eq. 5*). We then estimate the total number of folds in the genome – including those yet to be assigned – by dividing the total number of genes with C , *i.e.* $N_{folds} = N_{genes} / C$. In calculating this value, we assume that the distribution of folds among the structurally assigned genes is equal for unassigned genes – *i.e.* that they too follow a power law – and we apply the same value of C across the whole genome. By doing this, we suggest that most of the folds encoded by these remaining genes occur very few times, with a small possibility that some may occur many times. This is a realistic assumption, as in fact, homology studies of the unassigned genes indicate that a significant fraction has no, or very few related genes within the genome²³. Wolf *et al.*²⁴ have previously used a more complex method to obtain N_{folds} , in which they counted the number of protein families in a genome, and then extrapolated this number to estimate the population of folds. Despite the differences in procedures, the estimates for N_{folds} are similar to ours. Finally, having obtained values for C and N_{folds} , only N_0 in *eq. 4* needs to be adjusted for an optimal fit between the model and actual data.

The model versus the genomes

In Figure 4, we present the results of our simulations compared to the distributions of four representative genomes, including the smallest *M. genitalium* and largest *C. elegans* in the dataset. (Plots for the remaining 16 organisms are available from the supplementary website.) Here the genomic data do not give smooth lines because they comprise single population samples rather than averages. Nevertheless, we highlight the remarkable resemblance between the modelled and observed distributions, including the varying gradients for different organisms. The fits are particularly good for low fold occurrences; our predictions for the numbers of folds with fewer than five copies fall within 5% of the actual data. The plots for *A. fulgidus*, *M. tuberculosis* and *M. genitalium* (Figures 4b-d) closely follow the power law throughout. However the distribution for *C. elegans* diverges from a straight line for fold occurrences above 10 (Figure 4a), an observation we also make for *S. cerevisiae*.

Discussion

The parameter values we used for the 20 organisms are summarised in Table 1. The average occurrence of folds (C) ranges 2.4-31.6 and indicates that there is a higher level of duplication in larger genomes. This a simple reflection of the observation that higher organisms have more paralogous genes that provide a fuller spectrum of related, but modified protein functions.

The number of folds (N_{folds}) spans 200-613 and though the trend is less marked, also increases with genome size. Our figures are in broad agreement with the predictions provided of Wolf *et al.*²⁴ and we note that none of the genomes exceed the estimates for the universal population of folds^{25,26}. Previous work have extensively analysed the

degree to which the most common folds are conserved across different genomes^{10,11,15,27}. These studies found that many of the largest folds and families are conserved across closely related genomes, although their precise occurrences differ. Clearly, this stems from both the evolutionary relationship, and functional requirements that such organisms share. For example in *M. genitalium* and *M. pneumoniae*, the ferrodoxin-like fold dominates. In the eukaryotic *S. cerevisiae* and *C. elegans*, the protein-kinase and α - α superhelix folds are common. Although our current model only simulates the development of individual genomes, each simulation follows the evolution of a single genome from the root to of the phylogenetic tree to the branch (Figure 1d). If we believe that organisms originate from a common ancestor, then separate genomes would have undergone the same evolutionary process before diverging. Therefore, the protein families that grew during this stage of evolution would be large for all organisms concerned, whereas families that grew after divergence would only be prominent in a few genomes. Without this common ancestry we would not observe the same level of conservation of the largest fold families in different genomes.

Of greatest interest are t , r , and the free parameter N_0 . First considering t , ranging 280–13727, we find that larger genomes have experienced more duplication events than smaller ones and reflect the variations in fold occurrences. Whereas the real time between successive generations may differ between organisms, it leads to the logical conclusion that complex genomes such as *C. elegans* experienced much longer development times before becoming full organisms, than simpler genomes such as *E. coli*.

Next we turn to the rate of fold acquisition; here we consider the reciprocal of r , which gives the number of duplication events that must occur for each acquisition. Here we emphasise that this parameter gives the average rate of fold acquisition for the entire course of the simulation, and rates at particular time-points most likely vary during evolution. Like t , there is a wide spread of values, from 1.6 to 83.3 and fold acquisitions are more frequent in smaller genomes. New folds are mainly introduced from two sources: the intra-genomic creation of a new fold and lateral transfer of existing folds from other organisms. Although the model does not distinguish between the two, given the high rates, we expect the latter to account for most acquisition events. This is further supported by the view that lateral transfer is more prevalent in prokaryotes than eukaryotes²⁸⁻³³, and we observe that the rate of fold acquisition is lowest in *S. cerevisiae* and *C. elegans*. Using the model, we calculate that about 30% of the *E. coli* genes are descended from acquisitions; this provides a crude estimate for the maximum number of genes introduced through lateral transfer, and compares to the 18% of genes that have been introduced since the organism's divergence from the *Salmonella* lineage²⁸.

Finally we look at N_0 , which ranges 20-280, and like the other parameters has a tendency to increase with genome size. In this study we have assumed that there is only one of each type of fold in the starting genome. Although we might expect some folds to be present in several copies, further extensions to the model indicate that the initial distribution does not greatly affect the final power-law distribution³⁴. In addition, even if there were multiple copies of some folds, given their small starting sizes of the genomes, the difference between the most and least common folds would be very small

– *i.e.* some folds occurring three or four times at most. Therefore, the starting state we use is a good approximation.

Although N_0 is a free parameter, its value is effectively fixed by the current appearance of each genome. So as a natural outcome of our model, it is worth discussing the evolutionary implications. However, as the origin of modern organisms remains a fiercely debated subject, we must treat this matter with care. For the sake of argument, suppose that N_0 corresponds to the size of the genome that emerged out of the progenetic entities³⁵. Then, if all organisms originated from a common ancestor, the variation in N_0 requires consideration. So far, we have included gene loss as a factor during the expansive stage of evolution. However, it is also a major effect after the genome has reached its maximum size. Commonly termed reductive evolution, this is almost certainly responsible for the subsequent contraction of many bacterial genomes. A detailed investigation into *H. influenzae* and *E. coli* showed that their last common ancestor is likely to have been at least as large as the latter³⁶. Similar studies of other obligate parasites *Mycoplasma*, *Mycobacterium*, *Rickettsia*, *Chlamydia* and *Buchnera* indicate that the effect has been equally acute in these genomes³⁷⁻³⁹. So given the reduced genomes, we inevitably underestimate their initial sizes. On the other hand, there is so far little evidence that *E. coli*, *S. cerevisiae* and *C. elegans* have undergone such radical changes and their current genomes could be near their maximum sizes. Therefore as our simulation more accurately models the evolution of these organisms, their N_0 values suggest that the universal ancestral genome contained 170-280 folds and by default, genes. In a plot against N_{genes} , N_0 appears to converge to about 300 (supplementary material), thus setting a ceiling for the initial genome sizes of organisms

that are larger than *C. elegans*. We also note that these numbers are similar to the size of the minimal gene set predicted by Mushegian and Koonin ⁴⁰. Simulations for the most recently sequenced *D. melanogaster*, *A. thaliana*, and *H. sapiens* are work in progress, and their N_0 values will be of great interest.

We examined two further models for evolution after the organisms reached their maximum sizes. Both of these involve a larger amount of gene loss than the expansive model we focussed on above. For reductive evolution, we started with a simulated genome resembling *E. coli*, and randomly removed genes until it was the size of the smaller bacteria. For steady-state evolution where genomes maintain the maximum sizes, random gene duplications and losses were made at the same rate. In both cases, the power-law distributions remain intact.

The last main point for discussion is whether the probability of fold duplication should depend only on its current occurrence. Clearly, there are further factors that could contribute. For instance, some genes are located where genomes recombine more rapidly ⁴¹, and folds with high occurrence tend to have more symmetric structures, as seen in lattice models ⁴². Above all, selective pressure in the duplication and deletion of particular genes is perhaps the most dominant factor. However, tests of different “usefulness” functions that modify the probabilities of fold duplication had little effect on the outcome of the simulation. This is reinforced by the observation that folds with high occurrences in yeast are not necessarily associated with many functions ^{43,44}. For example, the highest-ranking fold, the 7-bladed β -propeller occurs 140 times, but is so far only linked with two distinct functions. Therefore in conclusion, selective pressure is

probably the single most important factor in determining the fate of individual genes, and our results clearly do not apply to the exact occurrence of an individual fold. Nevertheless, we demonstrate that the power-law distributions seen at a genomic scale – across many different organisms – result from an underlying stochastic process of evolution involving random duplications and an overall acquisition of folds.

Conclusion

In conclusion, this paper has demonstrated that the occurrence of protein folds, superfamilies, and families in genomes follow a power-law distribution. At first glance this may appear to be nothing more than a mathematical curiosity. However, the behaviour summarises an important feature in biology that a few members often dictate the overall appearance of a population – in this case, that most genes in a genome encode one or a few fold types. By designing a minimal, but biologically realistic model, we showed that such a distribution stems from how these organisms evolved. Furthermore, it allows us to estimate the rate at which different genomes have acquired new folds, and speculate on the size of the universal common ancestor. As the current state of the genomes provides only a snapshot in evolutionary time, these values are hard to obtain by other methods. In limiting the complexity of the model, we have identified the essential components of evolution that produce the power-law behaviour: gene duplication and flow. We anticipate future studies to incorporate further parameters for factors such as gene deletion and selective pressure to build a more biologically exact model. However, initial analyses suggest that such factors are most important at the level of individual proteins, families or folds, rather than at a genomic level.

Supplementary material

Supplementary material is available from <http://www.partslist.org/powerlaw>.

Acknowledgements

We thank Sarah Teichmann, Paul Harrison, Roman Laskowski, Annabel Todd, Ronald Jansen and Jimmy Lin for helpful discussions during preparation of the manuscript.

References

- 1 Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA.
- 2 Mandelbrot, B. B. (1953). *Symposium on Applications of Communications Theory, London*.
- 3 Wasserman, S. & Faust, K. (1994). *Social Network Analysis*, Cambridge University Press, Cambridge.
- 4 Editorial. (June 1999). Members of the Clever project. In *Sci. Am.*, pp. 54.
- 5 Albert, R., Jeong, H. & Barabasi, A. L. (1999). Internet: Diameter of the World-Wide Web. *Nature* 401, 130-131.
- 6 Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* 407(6804), 651-4.
- 7 Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol* 307(3), 929-38.
- 8 Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509-512.
- 9 Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C., Simons, M. & Stanley, H. E. (1994). Linguistic features of noncoding DNA sequences. *Phys Rev Lett* 73(23), 3169-3172.
- 10 Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* 274(4), 562-76.

- 11 Huynen, M. A. & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* 15(5), 583-589.
- 12 Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res* 28(1), 257-9.
- 13 Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M. & Servant, F. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 29(1), 37-40.
- 14 Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E. V., Mittard, V., Mulder, N., Phan, I. & Zdobnov, E. (2001). Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res* 29(1), 44-8.
- 15 Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* 33(4), 518-34.
- 16 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 U S A* 95(25), 14658-63.
- 17 Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome Res* 9(1), 17-26.

- 18 Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des* 3(6), 497-512.
- 19 Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 136(3), 225-70.
- 20 Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng* 6(5), 485-500.
- 21 Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 25(1), 231-4.
- 22 Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A* 95(11), 5913-20.
- 23 Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nat Struct Biol* 8(6), 559-66.
- 24 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(4), 897-905.
- 25 Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357(6379), 543-4.
- 26 Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature* 372(6507), 631-4.

- 27 Lin, J. & Gerstein, M. (2000). Whole-genome trees based on the occurrence of
folds and orthologs: implications for comparing genomes on different levels.
Genome Res 10(6), 808-18.
- 28 Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the
Escherichia coli genome. *Proc Natl Acad Sci U S A* 95(16), 9413-7.
- 29 Doolittle, W. F. (1999). Phylogenetic classification and the universal tree.
Science 284(5423), 2124-9.
- 30 Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer
and the nature of bacterial innovation. *Nature* 405(6784), 299-304.
- 31 Kidwell, M. G. (1993). Lateral transfer in natural populations of eukaryotes.
Annu Rev Genet 27, 235-256.
- 32 Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could
account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14(8),
307-11.
- 33 de la Cruz, I. & Davies, I. (2000). Horizontal gene transfer and the origin of
species: lessons from bacteria. *Trends Microbiol* 8(3), 128-133.
- 34 Kamal, M., Luscombe, N. M., Qian, J. & Gerstein, M. (Manuscript in
preparation). Power law and genomes: numerical analysis of evolutionary
models.
- 35 Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc Natl
Acad Sci U S A* 97(15), 8392-6.
- 36 de Rosa, R. & Labedan, B. (1998). The evolutionary relationships between the
two bacteria Escherichia coli and Haemophilus influenzae and their putative last
common ancestor. *Mol Biol Evol* 15(1), 17-27.

- 37 Andersson, S. G. & Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends Microbiol* 6(7), 263-8.
- 38 Andersson, J. O. & Andersson, S. G. (1999). Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* 9(6), 664-71.
- 39 Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & Kurland, C. G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396(6707), 133-40.
- 40 Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 93(19), 10268-73.
- 41 Barnes, T. M., Kohara, Y., Coulson, A. & Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141(1), 159-79.
- 42 Li, H., Helling, R., Tang, C. & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* 273(5275), 666-9.
- 43 Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 288(1), 147-64.
- 44 Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307(4), 1113-43.

- 45 Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W.
& Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs. *Nucleic Acids Res* 25(17), 3389-402.

Table Legend

Table 1. Parameter values that are used for the 20 organisms, listed from the smallest genome to the largest.

Figure Legend

Figure 1. Representations of the occurrence of protein folds in complete genomes and a schematic diagram of the evolutionary model. The main computational method for making family, superfamily or fold assignments to the protein products of genomes (proteomes) is to detect sequence homologies between the genes and proteins whose sequence or structures have been classified. Superfamily and fold matches were made using PSI-BLAST⁴⁵ against the SCOP database⁴⁵ and assignments are available for 17.6-34.6% of the gene sequences in the 20 genomes under consideration. These are available from <http://www.partslist.org>. Family assignments were obtained from the InterPro database from <http://www.ebi.ac.uk/interpro>^{13,14}. **(a)** The structural contents of a genome can be represented by counting the number of times different protein folds occur and then grouping together those with similar occurrences. **(b)** This relationship is described by a power-law, in which the number of folds (F) with a certain genomic occurrence (V) decays with equation $F = aV^b$. The vertical axis gives the number of folds, normalised by the total number of fold types in the genome, against the occurrence of each fold. We show the data for *E. coli* (■), and the fitted power-law function (—). **(c)** The observed distribution can be simulated by a model in which genomes grow from an initial small sizes to their current states by duplicating existing genes and acquiring new ones by lateral transfer or *ab initio* design. **(d)** The model effectively simulates the evolution of individual genomes from the root of the

phylogenetic tree to the branch representing the organism (—). (1) The model starts with an initial genome. (2) During evolution genomes diverge into different organisms by starting their own course of evolution. (3) Simulations end when the current size of the genome is achieved. (4) Genomes retain the same set of genes that were present at the point of divergence. Therefore, genomes that diverged more recently retain a more similar set of genes.

Figure 2. The occurrence of InterPro families (\diamond), SCOP superfamilies (\times) and folds (\blacksquare) in (a) *E. coli* and (b) *S. cerevisiae*. The distributions for the three gene classifications are similar and follow power-law behaviour (—).

Figure 3. (a) Average distributions obtained after 500 simulations. For small t , the distribution is exponential: $N_0 = 100$, $r = 0.6$, $t = 100$ (\diamond), for large t , the distribution is power-law: $N_0 = 100$, $r = 0.6$, $t = 30,000$ (\blacksquare). The solid lines are the fitted exponential ($y = ae^{-x}$) and power-law ($y = ax^b$) functions. **(b)** Phase diagram depicting the transition between exponential and power-law distributions for different r , t and N_0 . We do not show distributions for negative values of r and t because they are biologically implausible. The nature of the simulated distributions is determined from the residuals of optimally fitted functions, *i.e.* if the residual for the exponential function is smaller than the residual for the power-law function, the distribution is considered exponential and vice versa. The transition boundaries follow the parameter values for which the residuals are equal for both functions. While the conversion between the two phases is gradual, simulated distributions are exponential below the transition line and power-law

above. The threshold t required for a power-law is inversely correlated with r , and there is a rightward shift in the transition boundary for larger initial genome sizes (N_0).

Figure 4. The actual and simulated distributions for **(a)** *C. elegans*, **(b)** *A. fulgidus*, **(c)** *M. tuberculosis* and **(d)** *M. genitalium*. In each, the actual fold occurrences are represented by points (■) and the average distributions for 500 simulations are depicted by solid lines (—).