Available online at www.sciencedirect.com
ScienceDirect
Gene 402 (2007) 81 – 93
GENE
www.elsevier.com/locate/gene

Evolution of protein families: Is it possible to distinguish between domains of life?

Marta Sales-Pardo *, Albert O.B. Chan, Luís A.N. Amaral, Roger Guimerà

Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA

Received 3 April 2007; received in revised form 18 July 2007; accepted 23 July 2007
Available online 14 August 2007
Received by M. Di Giulio

Abstract

Understanding evolutionary relationships between species can shed new light into the rooting of the tree of life and the origin of eukaryotes, thus, resulting in a long standing interest in accurately assessing evolutionary parameters at time scales on the order of a billion of years. Prior work suggests large variability in molecular substitution rates, however, we still do not know whether such variability is due to species-specific trends at a genomic scale, or whether it can be attributed to the fluctuations inherent in any stochastic process. Here, we study the statistical properties of gene and protein-family sizes in order to quantify the long time scale evolutionary differences and similarities across species. We first determine the protein families of 209 species of bacteria and 20 species of archaea. We find that we are *unable* to reject the null hypothesis that the protein-family sizes of these species are drawn from the same distribution. In addition, we find that for species classified in the same phylogenetic branch or in the same lifestyle group, family size distributions are not significantly more similar than for species in different branches. These two findings can be accounted for in terms of a dynamical birth, death, and innovation model that assumes identical protein-family evolutionary rates for all species. Our theoretical and empirical results thus strongly suggest that the variability empirically observed in protein-family size distributions is compatible with the expected stochastic fluctuations for an evolutionary process with identical genomic evolutionary rates. Our findings hold special importance for the plausibility of some theories of the origin of eukaryotes which require drastic changes in evolutionary rates for some period during the last 2 billion years.

Keywords: Genetics; Evolutionary dynamics; Birth–death–innovation models; Theories of the tree of life; Neomuran

1. Introduction

Species, and their corresponding genomes, evolve and adapt over time. Until recently, it was widely accepted that the lineage of any species could be represented by a branch in a phylogenetic tree. Recent discoveries have pointed out that processes such as lateral gene transfer also play an important role in evolution, making the redefinition of the “tree of life” one of the “biggest challenges in evolutionary biology” (Doolittle, 1999).

The current consensus is that all living species belong to one of three different domains: bacteria, archaea, and eukaryotes (Woese et al., 1990). However, the evolutionary relationships between these three domains of life are still hotly debated (Dagan and Martin, 2006; Kurland et al., 2006, 2007; Martin et al., 2007). Specifically, there is still no agreement neither on where the “root” of the tree of life lies nor on the origin of eukaryotes. Current theories trying to explain the origin of eukaryotes can be roughly divided between those, such as the “Neomuran” theory (Cavalier-Smith, 2002a,b), that surmise that eukaryotes and archaea are “sister” branches originating from a

Abbreviations: BDI, birth–death–innovation; *E*-value, expectation value (returned from BLAST); HAMAP, High-quality automated and manual annotation of microbial proteomes; KS test, Kolmogorov–Smirnov test; MCL, TribesMCL (Markov chain clustering applied to protein-family identification); Myrs, millions of years; NCBI, National Center for Biotechnology Information; **R**, matrix of KS test results; RH, Reed–Hughes; TCL, Transitive clustering algorithm.

* Corresponding author. Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA. Tel.: +1 847 491 7238; fax: +1 847 491 7070.
E-mail address: msales@northwestern.edu (M. Sales-Pardo).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.07.029

bacterial common ancestor, those that surmise that eukaryotes arose from the symbiosis of a bacterium inside an archaean host (Doolittle, 1998; Martin and Müller, 1998), and those that hypothesize that eukaryotes are an ancestral lineage completely independent from prokaryotes (Kurland et al., 2006). At the root of the controversy lies the fact that tracing the evolution of genomes on the time scale of billions of years is extremely difficult.

A large body of literature is devoted to the construction of phylogenetic trees given a pool of taxa (Kuhner and Felsenstein, 1994; Huelsenbeck, 1997; Sanderson, 2002). In particular, many authors developed statistical tools suited to hypothesis testing concerning the point of divergence of a pair of species and the evolutionary rates along the different branches (Huelsenbeck, 1997; Liò and Goldman, 1998; Fares et al., 2006). In the simplest scenario, that of the so-called "molecular clock," genes accumulate mutations over time at a specific and fixed substitution rate. There have been claims that substitution rates are variable and change from branch to branch (Huelsenbeck, 1997; Welch and Bromham, 2005).

Inferring evolutionary parameters at the genomic level from molecular substitution rates is not a trivial task. A common approach is to look at silent site substitution rates between paralogous genes to, for instance, infer gene duplication rates (Lynch and Conery, 2000). A caveat is that silent substitution rates are estimated to take values on the order of 0.005/Myrs (Ochman and Wilson, 1987), so that, over a billion years one would expect a silent site to have changed about five times. Therefore, it would not be possible to infer evolutionary rates on the billion year scale by comparing silent sites, since two sites may be the same, but that would not warrant that they had not changed in a billion years. In addition, due to lateral gene transfer, we know that genes of a single species may have different evolutionary histories (Doolittle, 1999; Zhaxybayeva et al., 2005). Therefore, the comparison between pairs of paralogous or orthologous genes may not provide information about within species evolutionary rates.

A second caveat is that, although it seems reasonable to assume that, for instance, the need for a species to adapt to abrupt environmental changes may trigger a higher rate of change in the species' genome, there is no obvious argument why a high mutation-rate period must extend beyond the short time for adaptation. This second caveat suggests two questions worth pursuing.

First, whether on the time scale of billions of years genomic evolutionary rates are species-dependent, with some species having mutated more frequently over time, or whether all species have evolved in a similar fashion, so that frequent mutation periods resulting in, for instance, more gene duplication events, can be seen as stochastic fluctuations of a single evolutionary process.

Second, if one can determine genomic evolutionary rates for species in the three domains of life on the time scale of billions of years, what can one say about the origin of eukaryotes? Obviously, the analysis of the dynamical evolution of species, cannot by itself solve the debate on the rooting of the tree. Nevertheless, one can investigate differences in rate patterns implied by the main theories of the origin of eukaryotes. For instance, in the Neomuran scenario, archaea and eukaryotes arose from a specific class of bacteria that underwent a dramatic transformation during the period from 850 to 580 Myrs ago. A plausible mechanism for such transformation is that the bacterial ancestor of eukaryotes and archaea evolved with dramatically larger evolutionary rates for a period of 270 Myrs. If this were the case, one would expect to find clear differences between the genomic evolutionary rates of bacteria, and those of archaea and eukaryotes.

Similarly, the theory postulating the existence of a unicellular eukaryote predator as the ancestor of eukaryotes (Kurland et al., 2006) entails major episodes of genome reduction of non-predator unicellular organisms that gave rise to fast-growing prokaryotic species. This hypothesis, despite not having been fully elaborated, could in principle involve sudden changes in the evolutionary rates of some species. In contrast, in the scenario postulating the merging of a bacterial symbiont into an archaean host, there is *a priori* no reason why one should expect bacteria and archaea to have evolved according to different genomic evolution rates.

In order to address these questions, one needs to compare the long time scale evolution of different genomes. A simple, yet powerful, way to quantify the long time scale evolutionary differences and similarities across species is to study the statistical properties of gene and protein-family sizes, that is, groups of genes and proteins with significant similarity at either structural or sequence levels, and which are presumably descendants of a common ancestor (Brenner et al., 1995; Koonin et al., 1995; Huynen and van Nimwegen, 1998; Yanai et al., 2000; Karev et al., 2002, 2003, 2004; Harrison and Gerstein, 2002; Unger et al., 2003; Reed and Hughes, 2004). The advent of high-throughput techniques has enabled researchers to tackle this matter in an unprecedented way. A rapidly increasing number of bacterial genomes have been fully sequenced (van Nimwegen, 2003), making it possible to undertake a large scale comparative study of the genomes of many different species.

1.1. Statistical properties of protein families

In a single species, gene and protein families span a broad range of sizes n (Brenner et al., 1995; Koonin et al., 1995), with cumulative distributions $P(n)$ that decay as a power law (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Unger et al., 2003)

$$P(n) \sim n^{-\alpha}. \tag{1}$$

This power-law behavior has been explained in terms of gene birth (duplication), death (loss), and innovation (*de novo* acquisition) (BDI) models (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Karev et al., 2002, 2003, 2004; Reed and Hughes, 2004). In these models, the exponent α characterizing the protein-family size distribution directly reflects the rates at which genes are duplicated, lost, and acquired *de novo*. The exponent thus provides a proxy for the evolutionary processes that shape the genome of a given species on time scales on the order of a billion years.

Note that, in such an analysis, the specific function or sequence of the members of a given family is not taken into account. A BDI process merely describes the growth of family sizes with some "effective" rates (that is, rates at a much larger scale than molecular rates) that are the same for each element of a family. For the specific case of protein families that are organism/superkingdom specific, we expect these families to have appeared more recently than non-specific protein families. In virtue of the BDI model, younger families will be on average smaller in size than older families. Thus, considering these families or not will only affect the head of the distribution and not the tails of the power law, which carry the information on the evolutionary rates on the billion year time scale. Therefore, in our analysis, we do not need to give any special treatment to families that are organism or superkingdom specific (archaeal or bacterial), since eliminating these families from the analysis does not change the results we report.

Cross-species comparison of protein-family size distributions enables us to test whether different species have evolved with different "overall" evolutionary rates. For example, BDI models have been used in conjunction with empirical distributions of gene and protein families to estimate evolutionary rates for different species and to assess the relevance of each mechanism (birth, death, and innovation) in the evolution of a specific genome (Yanai et al., 2000; Karev et al., 2002, 2004).

Currently, the consensus on this issue is that protein-family size distributions for different species are characterized by different values of α. However, the significance of such difference in the estimated values of α has not been rigorously quantified. Each species typically has only tens to a few hundred families with more than three proteins, so that the tails of the family size distribution and, therefore, the evolutionary parameters inferred from them have, unavoidably, a large uncertainty. To assess whether differences in the empirical distributions are only due to stochastic fluctuations or represent a significant difference in the evolutionary rates of species at a genomic level, one needs to make a systematic comparison between a large pool of organisms. If some groups of species have evolved with different evolutionary rates, one expects to find clear differences in the comparison of protein-family size distributions. Instead, if all species have evolved with the same overall evolutionary rates, one will find no significant differences among the family size distributions for the whole pool of organisms.

Here, we analyze the protein-family size distributions of 229 species—20 archaea and 209 bacteria (see Supplementary Material for a list of species). We find that all species studied have protein families whose sizes are consistent with a common universal distribution, and that this finding is mostly unaffected by the choices of the method being used to identify protein families. Our findings imply that the observed differences in the empirical distribution for different species can be fully accounted for by stochastic fluctuations in the evolutionary process and do not require species-specific trends. Furthermore, we do not observe any significant additional similarities between species in the same phylogenetic branch in comparison to species in different branches. We do not observe any differences between species with the same lifestyle (parasitic and non-parasitic), either. Our empirical findings are supported by numerical simulations of a dynamic BDI model with fixed evolutionary parameters. Therefore, we conclude that the statistical properties of protein-family sizes are consistent with the hypothesis that, for time scales on the order of billions of years, bacteria and archaea evolved with identical genomic evolutionary rates.

Our results have significant implications of our findings on the plausibility of current theories for the tree of life and the origin of eukaryotes. Although not conclusive, the comparison of prokaryotic protein-family size distributions with those of four eukaryotes (*S. cerevisiae*, *P. falciparum*, *C. elegans*, and *D. melanogaster*) is consistent with the hypothesis that all genomes evolved with the same genomic evolutionary rates. Thus, our findings are not consistent with *any* theory that surmises drastic changes in evolutionary rates on the billion year time scale. In order to assess whether from the comparison of protein-family size distributions one could detect differences in species having evolved with different genomic rates for a certain period of time, we perform numerical simulations of a dynamic BDI model for the specific scenario of the Neomuran hypothesis. We find that we should be able to detect large differences in evolutionary rates, even if those differences are restricted to a period of only about 270 Myrs—the time during which the Neomuran revolution is hypothesized to have happened (Cavalier-Smith, 2002a).

2. Materials and methods

2.1. Protein sequence alignment

We study the genomes of 209 bacteria and 20 archaea stored in the GenBank database (Benson et al., 2006). We use BLAST (Altschul et al., 1990, 1997) to compare all pairs of protein sequences for each species, obtaining the expectation values (*E*-values) and bit scores for each comparison. To avoid hits for very common amino acid sequences, we use the low-entropy filter in our comparisons. To estimate *E*, we use the model of random sequences proposed in (Karlin and Altschul, 1990; Dembo et al., 1994).

2.2. Protein families

From either the *E*-values or bit scores *b* obtained for each pairwise comparison, we determine protein families by using two different algorithms: the TribesMCL algorithm (MCL) (Enright et al., 2002, 2003), and a transitive clustering algorithm (TCL) (Brenner et al., 1995). Ideally, for a pair of proteins the *E*-value and *b* should not depend on the order in which the comparison is done. However, it is well known that BLAST comparisons can be slightly asymmetric, thus we symmetrize them by using the most restrictive comparison value, that is, the smallest *E*-value or largest *b*.

The MCL algorithm obtains the protein families from the bit scores *b* for the pairwise comparison of the whole pool of proteins for a single species see (Enright et al., 2002). The TCL algorithm considers that a pair of proteins is related if the E-value is smaller than a certain threshold E_{t}. Since organisms typically have on the order of a thousand proteins, we set $E_{\mathrm{t}}=10^{-6}$ to minimize the rate of false positives. We define a protein family as a set of proteins that satisfy two conditions: (i) each protein is related to, at least, one other protein in the family, and (ii) none of the proteins from one family is related to proteins belonging to another family. This is equivalent to the assumption that evolutionary relationships are transitive—that is, if proteins A and B are evolutionarily related, and A and C are evolutionarily related, then B and C are also evolutionarily related (Brenner et al., 1995; Koonin et al., 1995).

We define the size n of a protein family as the number of proteins it contains. As done in other contexts (Tatusov et al., 2001), we put a threshold n_{t} on the minimum size of the families considered, which for the results we show is $n_{\mathrm{t}}=4$. Obviously such choice limits our analysis to a "small" fraction of the total genome of a species (on average, 29% for TCL and 10% for MCL). There are, however, two main reasons for not including data for small family sizes in our analysis. First, despite many genomes being completely sequenced, the expression *in vivo* of all the sequences that are identified as "proteins" has not been proved. Therefore, it is sensible to assume that proteins belonging to families of a certain size are more likely to be "real" proteins. Thus, by not including data for small families in our analysis, we eliminate the bias due to the way proteins are identified.

Second, we compare empirical results with the outcome of a BDI model (Section 2.5), which is predictive for large values of family sizes, that is, for the tails of the distributions. Therefore, the proper comparison of the distributions has to be made for large family sizes only. Nevertheless, we have analyzed the distributions for $n_{\mathrm{t}}=2$ (see Figs. S-2, S-3, and S-4 in Supplementary Material), which corresponds to analyzing on average 47% of the genome for TCL and 20% for MCL, and the results are very similar to those obtained for $n_{\mathrm{t}}=4$, that is we do not observe any phylogenetic pattern in the protein-family size distributions (see Section 3).

The reason why we use two different algorithms is that they give complementary answers in some specific cases, thus we can evaluate the robustness of our results. Compared to the TCL algorithm, the MCL algorithm generates families that are fairly small, since even some pairs of proteins with significantly large bit scores/small E-values can be classified into separate families. Therefore, if there is a false positive, the TCL algorithm will probably group that pair of proteins within the same family, whereas MCL would not. On the other hand, if there is a weak but real match signal between a pair of protein sequences, that is a high E-value, the TCL algorithm would classify the two proteins within the same family, whereas MCL would not.

We have checked that for the matches between protein sequences we identify, the number of amino acids per match is large enough to truly represent an evolutionary relationship. For a typical species (*B. subtilis*), 95% of the matches we find are 97 amino acids or longer, the shortest match being of 20 amino acids. The median and mean of the distribution of amino acids per match are of 230 amino acids and 286 amino acids, respectively.

2.3. Comparison of protein-family size distributions

For each species A with protein-family sizes $\{n^A\}=\{n_1^A, n_2^A, \ldots\}$, we compute the cumulative distribution $P_A(n)$ of family sizes, which gives the probability of finding a protein family of size equal to or greater than n. The cumulative distribution is defined as $P_A(n)=\sum_{n'=n}^{\infty} p_A(n')$, where $p_A(n)$ is the fraction of families of size n in $\{n^A\}$.

Then, we compare the distributions of all possible pairs (A, B) of species in the set $\{S\}=\{A,B,\ldots,L,\ldots\}$ using the Kolmogorov–Smirnov (KS) test (Press et al., 2002). The significance value r returned by the KS test quantifies the likelihood that the two sets of protein-family sizes $\{n^A\}$ and $\{n^L\}$ have been drawn from the same distribution. An r value greater than 5% is typically taken to indicate that one cannot reject the hypothesis that the two sets of protein-family sizes were drawn from the same distribution.

In our analysis, however, we want to test the null hypothesis of whether the whole pool of family size sets has been drawn from the same distribution. Thus, because we are making a large number of pairwise comparisons—(208 × 19)/2 in total—we expect some r values to be smaller than 5% *even* if the sets of proteins family sizes are in fact drawn from the same distribution.

To determine if the empirical significance values we obtain from the KS test deviate from those expected from the null hypothesis, we generate artificial sets of protein-family sizes by randomly distributing the actual protein-family size values among different species. With this procedure, we generate sets that, by construction, are drawn from the same distribution. The specific way in which we generate such sets is the following. First, we pool the sets $\{n^A\}$ of protein-family sizes of all the N_S species in $\{S\}$ together, thus obtaining a superset $\mathcal{N}=\cup\{n^A\}$. Each random sample comprises N_S random sets of protein-family sizes, such that each set has the same number of protein families as its corresponding empirical set. That is, suppose that we have two empirical sets $\{n^A\}$ and $\{n^B\}$ with p_A and p_B protein families, respectively. Then the superset $\mathcal{N}=\{n^N\}$ has p_A+p_B protein-family sizes. Each random sample consists of two sets $\{n^R{}_A\}$ and $\{n^R{}_B\}$ of p_A and p_B elements, respectively, that are drawn at random from $\mathcal{N}$.

For each random sample, we perform the pairwise comparisons between family size sets and obtain the distribution of r values for that sample. To obtain the expected distribution of r values under the null hypothesis that all protein-family sizes are drawn from a common underlying distribution as well as the 95% confidence intervals, we repeat the same procedure 1000 times. If the empirical distribution of r values falls within the confidence intervals of the distribution expected under the null hypothesis, then we cannot reject the null hypothesis. The comparison between both empirical and randomly generated distribution is thus a stronger test than the comparison with a significance level alone, since it enables us to test a null hypothesis that considers the whole pool of species.

Note that unlike in hypothesis testing for questions involving phylogeny (Huelsenbeck, 1997), we perform a non-parametric statistical analysis. This is because in our analysis *we do not make any assumptions* on the specific shape of the distribution, thus the proper test to use is a non-parametric one. Additionally, it is a well-documented fact that null hypothesis are often rejected when the pool of data is as large as the one we consider (Savage, 1954; Ijiri and Simon, 1977). As we show later, from our results we cannot reject the null hypothesis, that is, we cannot reject that empirical distributions have been drawn from a common underlying distribution, thus enabling us to draw strong conclusions from our analysis.

2.4. Ordering of the matrix of r values

Our goal is to identify whether we can group species in $\{S\}$ in subsets of species with more similar distributions among themselves. A common approach (Tsafrir et al., 2005) is to build a similarity matrix, that is, a square matrix $\mathbf{R}$ of order N (the number of species in $\{S\}$) such that each element $\mathbf{R}_{ij}=r(i, j)$ is the result from the comparison of the distributions of species at positions (rows) i and j.

Species can be ordered in $N!/2$ different ways $\{\mathcal{O}\}$ (considering that a given ordering of the nodes is equivalent to its reverse ordering). For each ordering of the species $\mathcal{O}_k(\{S\})=\{\mathcal{O}_k(A)=i,\ \mathcal{O}_k(B)=j,\ldots\}$, we have a different matrix $\mathbf{R}(\mathcal{O}_k)$. Block diagonal structures in matrices are indicative of clustering of rows. Therefore, to identify different potential clusters, we need to find the ordering $\mathcal{O}$ for which the structure of the ordered matrix $\mathbf{R}(\mathcal{O})$ resembles as much as possible a block diagonal matrix.

To quantify the goodness of each ordering $\mathcal{O}_k$, we define a cost function $C(\mathcal{O}_k)$ that weighs each element in the matrix with its distance to the diagonal (Sales-Pardo et al., in press),

$$C(\mathcal{O}_k)=\frac{1}{N}\sum_{ij}\mathbf{R}(\mathcal{O}_k)_{ij}|i-j|. \tag{2}$$

The best ordering $\hat{\mathcal{O}}$ is the one that minimizes C. Obviously, making an extensive search among the possible $N!/2$ possible orderings is not viable. Therefore, to find an ordering close to the best possible one, we perform simulated annealing—a standard technique to solve optimization problems (Kirkpatrick et al., 1983)—starting from a random ordering of the elements.

Once we find the best possible ordering of the species $\hat{\mathcal{O}}$, we can easily compute the "distance" $d(A, L)$ between any pair of species A and L as the absolute value of the difference between the position each species occupies in the best ordering $d(A, L)=|\hat{\mathcal{O}}(\mathrm{A})-\hat{\mathcal{O}}(L)|$.

To assess whether a group of s species are close in $\hat{\mathcal{O}}$, we compute the average distance between them and we compare it to the 95% confidence interval of the random expectation; this is, we generate 1000 random orderings of N species, we compute the average distance of the subgroup of s species for every sample, and we compute the 95% confidence interval for such distances. If the empirical average distance within a group of s species falls below this interval, then we conclude that at the 5% significance level, these s species are significantly closer in the ordering, and, therefore, their family size distributions are significantly more similar between themselves than to distributions of species in other groups.

2.5. Analysis of group patterns in protein-family size distributions

To assess whether there are any "group" patterns in protein-family size distributions, we group species following two criteria: phylogeny and lifestyle. We classify species into groups defined by each criterion and compute for each group the average distance (see Section 4) and average similarity between distributions. Then, we compare these empirical measurements to the distance and similarity that one would expect to find by chance.

By grouping species using phylogeny, our aim is to assess whether traditional taxonomic classifications capture similarities in evolutionary rates that are not detectable via global analysis. By grouping species by lifestyle, we investigate whether our analysis is affected by the size of the genomes, which, in principle, limits the family sizes and could therefore introduce some bias in the comparisons between distributions.

For the phylogenetic analysis, we group species according to the NCBI taxonomy database (Wheeler et al., 2000). We consider taxonomic groupings at two levels: domain and phylum. The reason for such choice is that in order to make a sensible assessment of whether there are any phylogenetic patterns in family size distributions at a given taxonomic level, one needs to have empirical data for at least two species classified in the same branch and for several branches at that level. For this reason, the phylum level is the deepest taxonomic level that we can analyze.

At the domain level, species are classified into two groups archaea and bacteria. At the phylum level, we only analyze phyla for which we have enough species: Actinobacteria, Bacteroidetes, Deinoccoccus-Thermus, Chlamydiae, Cyanobacteria, Firmicutes, Spirochaetes, and Proteobacteria. Note that these are only phyla of bacteria. All phyla for archaea as well as some phyla of bacteria (Aquificae, Chlorobi, Chloroflexi, and Thermotogae) have a single representative in the pool of species we consider. Therefore, we do not consider species in these phyla at this level of analysis, but we include these species in the results at the domain level (see Table in Supplementary Material for a complete list of all the species considered for each taxonomic group).

For the lifestyle analysis, we have manually grouped species into two groups: *parasitic* and *non-parasitic* (following the HAMAP database (Gattiker et al., 2003) and (Fitz-Gibbon and House, 1999; House and Fitz-Gibbon, 2002)). We classify as *parasitic* all organisms that are listed as symbionts, endosymbionts, or parasites, because organisms living in any of these ecological relationships have been found to have suffered a genome reduction over time (Andersson and Kurland, 1998). We classify as *non-parasitic* those organisms that are either free-living or are animal or plant commensals (see Supplementary Material Table 1 for a complete list of all the species considered for each lifestyle group).

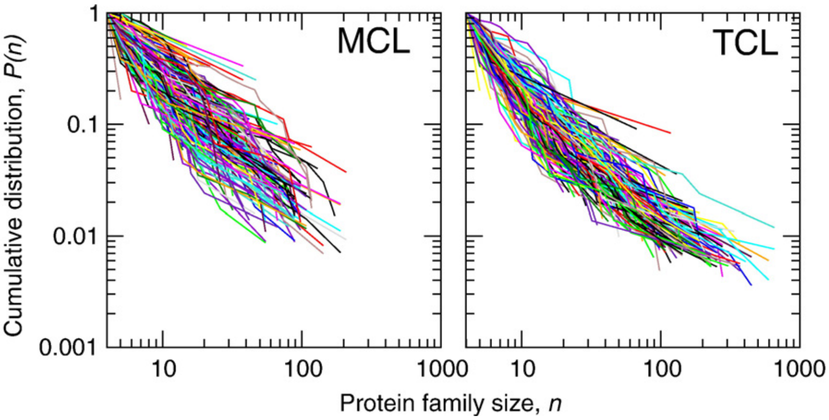

Fig. 1. Distribution of protein-family sizes for archaea and bacteria. Each individual curve shows the cumulative distribution of protein-family sizes for an individual species, that is, the probability of finding a protein family of size equal to or greater than n, with $n \geq 4$. We show the cumulative distributions obtained with two different methods (see Materials and methods): MCL (left) and TCL (right). Due to the small number of protein families per organism, typically 100, the distributions display large variability.

2.6. Dynamical BDI model

In the BDI model for protein-family growth proposed by Reed and Hughes (RH) (Reed and Hughes, 2004), each protein can duplicate with rate λ and die with rate μ. Assuming that at time zero the species has one family with a single protein, and that the ages $\{\tau\}$ of protein families within a single species are exponentially distributed $f(\tau) \sim e^{-\rho\tau}$, one can show that the cumulative $P(n)$ distribution of protein-family sizes decays as a power law $P(n) \sim n^{-\alpha}$ in the stationary state, with $\alpha = \rho/(\lambda - \mu)$ (Reed and Hughes, 2004).

In our dynamical BDI model, each protein can duplicate or die with fixed rates λ and μ, respectively. In addition, we consider that with fixed rate ρ_{sim} a mutation in some family of proteins will be large enough that it will give rise to a new family. Such multiplicative processes result in an exponential growth of the number of families and an exponential distribution of ages, as assumed in the RH model.

Note, however, that the observed rate of growth of the number of families ρ_{obs} is not, in general, equal to ρ_{sim}. Because there is a finite probability for a family to die, the observed growth rate for the number of families will be smaller than the dynamical rate $\rho_{\mathrm{obs}} \leq \rho_{\mathrm{sim}}$ (see Supplementary Material in Appendix).

Selection of parameters—We have selected the values of the rates λ, μ, and ρ as well as the total time for the numerical simulations t_{evol} in order to enable the most accurate comparison with the empirical data.

First, we fix $\lambda = 1$, which sets the unit for time. Then, we fix the exponent of the stationary cumulative distribution of family sizes $P(n)$ to the exponent empirically observed, $\frac{\rho_{\mathrm{obs}}}{(\lambda-\mu)} = \alpha_{\mathrm{emp}}$.

We are thus free to select either ρ or μ with the constraints $\mu < \lambda$ or $\rho_{\mathrm{obs}} \leq \alpha$. Once we select a value of ρ_{obs}—and therefore ρ_{sim}—we can determine the total time of the simulation t_{evol}. We fix t_{evol} so that a species will have on average a number of families equal to the average number of families N_f empirically observed—$N_f \approx 1900$ for the TCL method and $N_f \approx 2500$ for the MCL method. Hence,

$$t_{\mathrm{evol}} = \frac{\ln\left[N_f\right]}{\rho_{\mathrm{obs}}}. \tag{3}$$

With these constraints, we can still have a free parameter, but we have checked that a consistent selection of ρ_{sim}, μ and t_{evol} yields exactly the same results for the family size distributions.

Fig. 7 shows results for $\lambda = 1$, $\mu = 0.1$, and $\rho_{\mathrm{sim}} = 1.405$, which yield a ρ_{obs} such that the exponent of the cumulative distribution is $\alpha = \alpha_{\mathrm{emp}} = 1.4$; the total time for the simulations is, from Eq. (3), $t_{\mathrm{evol}} \approx 5$ time units, where the term “time units” indicates that the time is in units of λ. Simulations for $\mu = 0.5$ yield identical results (data not shown).

Note that for some choices of parameters we recover the correct time scale for the evolutionary process considered. Effective protein duplication rates λ_{eff} are on the order of one protein per billion years (Lynch and Conery, 2000). Therefore, for the total time T of the evolution of life (roughly 4 billion years), $T \times \lambda_{\mathrm{eff}} \simeq 4$. In our model, $\lambda_{\mathrm{eff}} = \lambda - \mu$ and $T = t_{\mathrm{evol}} = \ln N_f / \rho_{\mathrm{obs}}$, therefore $t_{\mathrm{evol}} \times (\lambda - \mu) = \frac{\ln N_f}{\alpha_{\mathrm{emp}}} \approx 5$, which is of the same order of magnitude as estimates in the literature.

3. Results

3.1. Comparison of protein-family sizes

First, we identify the protein families for each of the 229 species of archaea and bacteria using two different methods (see Materials and methods): the TribesMCL (MCL) algorithm (Enright et al., 2002, 2003) and a transitive clustering algorithm (TCL) (Brenner et al., 1995). In Fig. 1, we show the distribution of sizes of families with more than three proteins obtained for each species using both methods MCL and TCL. In both cases, we observe that protein-family size distributions display a large variability from species to species, and yet all the curves display heavy tails; most of the protein families have 10 or less proteins, but there are some protein families with hundreds of proteins (see Supplementary Material). Typically, each species has about 100 different families with more than three proteins; this small number of families accounts for the large fluctuations in the distributions.

To determine whether the differences between the individual family size sets are significant or just an artifact of the small number of families per species, we perform the Kolmogorov–

Smirnov (KS) test for the two sets of protein-family sizes of species A and B—$\{n^A\}$ and $\{n^B\}$—as described in the Materials and methods section. If the significance r value returned by the KS test is high, it means that there is a large probability that the pair of protein-family size sets being tested was drawn from the same distribution.

With the r values obtained from all possible pairwise comparisons between the 229 species of bacteria and archaea, we construct a matrix **R**. Each element in the matrix $R_{ij}=r(i, j)$ is equal to the r value for the comparison between the sets of family sizes of species i and j. To find whether there are any distinct clusters of species with similar distributions, we order the **R** matrix so that the highest r values are closest to the diagonal (see Materials and methods).

From the ordered analysis of the ordered **R** matrices (Fig. 2 A, B), we conclude that there are no distinct clusters of species with higher r values among themselves than with the rest of the species. Rather, we observe that the distributions are all similar to each other for each of the two methods used to identify protein families. Indeed, most significance values are larger than 50%, implying that we can accept the null hypothesis of there being a common underlying distribution with confidence. However, because of the large number of pairs of species being compared, one expects to obtain some low r values despite the family size sets being drawn from the same distribution.

To test whether the low r values obtained from the KS test agree with the expectation for a common underlying distribution in Fig. 3, we compare the empirical r values with those we obtain from constructed sets drawn from a single distribution (see Materials and methods). If we consider the whole pool of species, we find that the observed r value distribution for both MCL and TCL methods is very close to the expectation for a common underlying distribution (Fig. 3 A, B). Nevertheless, it has a slightly lower fraction of values $r \simeq 1$ and a slightly larger fraction of values $r<0.01$ than the expectation for a common underlying distribution. Low significance r values can arise in two possible cases: (i) species comprise a small number of families, and therefore their family size distributions are subject to even larger fluctuations than other species; (ii) species comprise a very large number of families, thus, because one would expect their family size distributions to be smoother, even small fluctuations will result in a small significance value. If we remove the ten species which have the lowest average r values, which constitute only 4% of the species, we find that the empirical distribution agrees remarkably well with the expectation for a common underlying distribution, implying that we cannot reject the hypothesis that all protein-family size distributions are drawn from the same underlying distribution.

3.2. Combined protein-family size distribution

Our results show that there is a high likelihood that protein-family sizes for all bacteria and archaea are drawn from the same distribution. Thus, one can obtain a better estimate of the exponent of the real underlying universal distribution by pooling all sets of family sizes together. Fig. 4 shows the probability density functions obtained for the whole pool of bacteria and for the whole pool of archaea. Note that because the global distributions are significantly smoother than individual ones, here we can directly compare the probability density functions we obtain from the empirical data and the numerical simulations, as opposed to using the cumulative. It is visually apparent that, for both methods, family size distributions for archaea and bacteria are very similar. Indeed, at a 5% significance level, we cannot reject the hypothesis that protein-family sizes for archaea and bacteria are drawn from the same distribution.

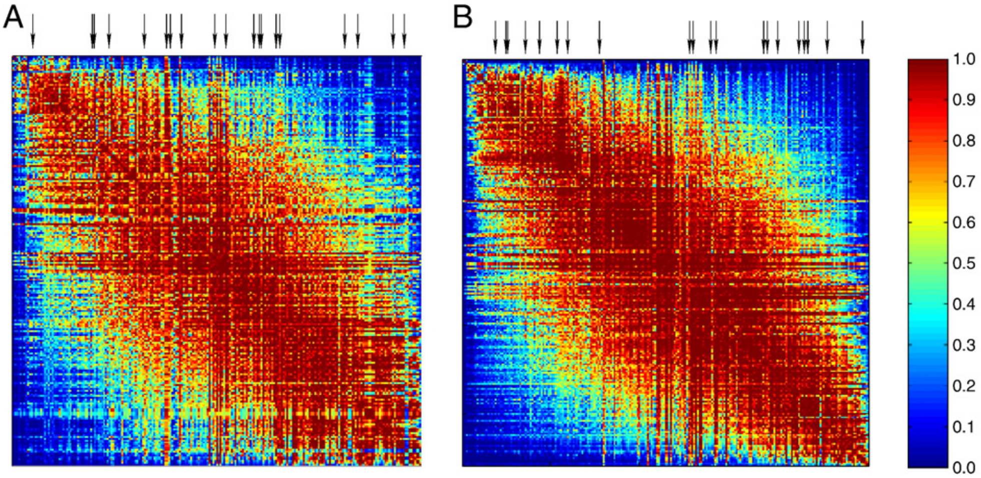

Fig. 2. Ordered **R** matrices. We show the ordered matrices of KS test r values for the pairwise comparisons of family size distributions for the whole pool of species (archaea and bacteria) obtained using: (A) MCL and (B) TCL methods. We color each matrix element according to its r value, following the color code shown on the right hand side of the diagram. $r \rightarrow 1$ means that we cannot rule out the hypothesis that the two samples were drawn from the same distribution, $r \rightarrow 0$ means that the samples could not have been drawn from the same distribution. We have ordered the matrix (see Materials and methods) to cluster organisms with a similar set of r values with any *a priori* assumption on the taxonomic relationships among them. The arrows show the columns corresponding to archaea species. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

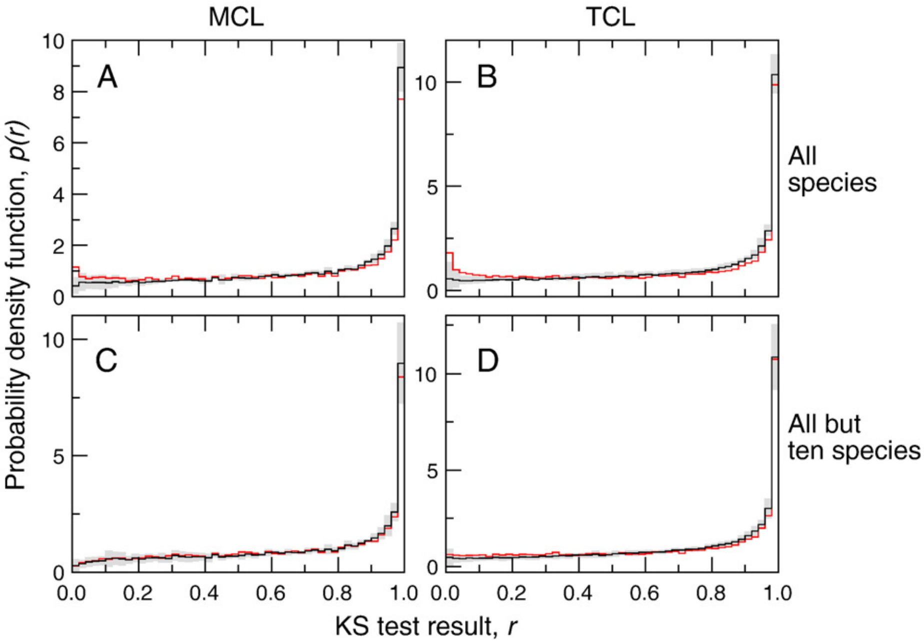

Fig. 3. Probability density function of the KS test result r. In each plot, we show the KS test result distribution (black line) and we compare it with the "null" distribution obtained from 1000 random samples with the same number of species and the same underlying family size distribution (red line). The regions shaded in gray show the 95% confidence interval of the null distribution. If both empirical and null distributions are equal, then one can conclude that family sizes for the different species are drawn from the same underlying distribution. Plots (A) and (B) show the results for the MCL and TCL methods, respectively. Note that in both cases the fraction of values $r \simeq 1$ is slightly lower than the expectation for a common underlying distribution and the fraction of values $r<0.01$ is slightly larger than the expectation for a common underlying distribution. Plots (C) and (D) show the results for the MCL and TCL methods for the case in which we remove the ten organisms which have the lowest average r value. Note that in both cases the agreement of the empirical distribution of r values with the null distribution of r values is excellent.

Additionally, the results we report are robust with respect to the method used, since we cannot reject the hypothesis that protein-family sizes obtained with different methods are drawn from the same distribution. Power law fits to the probability density functions of bacteria (Fig. 4) yield exponents $\alpha_{\mathrm{MCL}}=1.3 \pm 0.1$ and $\alpha_{\mathrm{TCL}}=1.4\pm0.1$ (Fig. 4). Note that if we plot a power law that decays with an exponent 2.4, the line overlaps with the two curves, implying that the exponent for the global cumulative distribution obtained with both methods is compatible with $\alpha=1.4$, even though because of the small region available for the fit to a power law, one can hardly distinguish between curves with exponents in the range [2.3,2.5]. Thus, by combining the protein families of 229 species and based on the statistical analysis of the distributions of protein-families sizes, we are able to obtain the best available estimate of the value of the exponent for the distribution of family sizes.

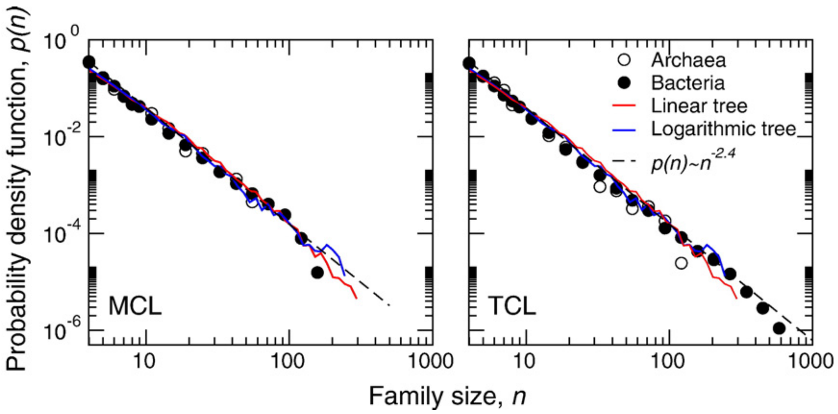

Fig. 4. Combined protein-family size distribution. We show the probability density function of protein-family sizes for the whole pool of archaea and the whole pool of bacteria, with both methods MCL and TCL. Note that in both cases distributions for archaea and bacteria decay with the same slope. The dashed line corresponds to a power-law decay with exponent $1+\alpha=2.4$. The solid lines correspond to the probability density function obtained from the numerical simulations of a BDI model (Materials and methods and Discussion). Note the agreement between simulations and the empirical data.

3.3. Average distance and similarity

The fact that we cannot reject the hypothesis that protein-family size distributions for all archaea and bacteria are drawn from the same distribution, however, does not exclude two possibilities: (i) that species that are closer from a taxonomic point of view have significantly more similar protein-family size distributions; (ii) that species living in a parasitic relationship that have suffered a reduction of their genomes have significantly more similar protein-family size distributions than those of non-parasitic organisms.

To investigate this question, we analyze what is the average *similarity* and *distance* between species classified into the same taxonomic groups at two levels: the domain level (archaea or bacteria), and the phylum level (see Materials and methods). We also analyze the same quantities for species classified as either parasitic or non-parasitic (see Materials and methods) for two cases: considering the whole pool of species, and restricting the analysis to bacterial species.

We consider that the similarity between a pair of species is directly provided by the KS test r value and that the distance between a pair of species is the difference between the position of the species in the best ordering obtained for the **R** matrix (Fig. 2 and Materials and methods). For a set of s species grouped into the same domain or phylum, the average distance/similarity is the average of the $s(s-1)/2$ pair distances/similarities. To assess whether any group of species is significantly close or similar, we compare these average quantities with the 95% confidence interval of the random expectation (see Materials and methods).

Fig. 5 shows that, for both MCL and TCL methods, species belonging to the same group either at a taxonomic (domain or phylum) or lifestyle level are neither more similar nor closer in the ordering than what is expected by chance, except for the species in the phylum Chlamydiae, which are consistently and significantly close in the ordering and similar according to the KS test r values. A closer look at Chlamydiae species reveals that they have a small number of families. Actually, Chlamydiae are known to be obligate intracellular parasites with very short genomes. Statistically, one expects sets with a small number of families to have larger fluctuations. Therefore, if the distributions for these sets are located close to the center of the range of variation of the sample, then any statistical test will return higher significance values from the comparison of these sets than if we considered a set with a large number of families. This is precisely the case for many of the family size distributions of Chlamydiae species. Indeed, for any of the pairwise comparisons with the remaining species, the r values are very high, implying that family size distributions for species in this phylum are not only significantly more similar between themselves but to *any* other species belonging to a different phylogenetic branch. For this reason, many Chlamydiae species are located in the center of the ordered **R** matrix, and therefore the average distance between them is small.

Note that, in spite of such finding, we do not observe that, in general, parasitic organisms with shorter genomes have a larger similarity between their distributions than what one would expect by chance, implying that genome size does not introduce any bias in our analysis. In fact, in order to completely rule out such possibility we have checked that restricting our analysis to those genomes of non-parasitic species yields the same lack of taxonomic pattern found for the whole pool of species: If we divide non-parasitic species into two taxonomic groups, archaea and bacteria, the average similarity for each group is compatible with the random expectation, meaning that family size distributions for non-parasitic archaea are not more similar between themselves than they are to those for non-parasitic bacteria (Fig. S-1 in Supplementary Material).

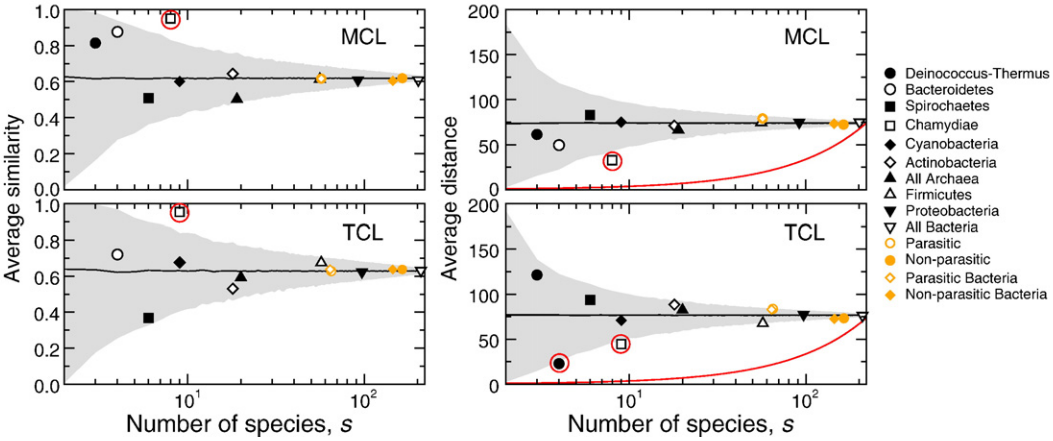

Fig. 5. Average similarity and distance between groups of species classified in the same group at a taxonomic level (phylum or domain), and at a lifestyle level (parasitic and non-parasitic) (Materials and methods). For each set of s species classified in the same group, we plot the average similarity (left column) and the average distance according to the best ordering found for the **R** matrix (right column) obtained for both methods, MCL and TCL. Each symbol corresponds to a different phylogenetic of lifestyle group. The black line shows the average distance/similarity expected for a group of s species picked at random. The gray areas show the 95% confidence intervals for the average distance/similarity of groups of s species picked at random. Red lines correspond to the expected average distance if species in the same group were ordered consecutively. Symbols circled in red indicate groups of species that are either significantly similar or significantly closer in the ordering. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Dynamical BDI Model

The overall form of protein-family size distributions can be understood in terms of birth (duplication with or without mutations), death (loss) and innovation (*de novo* acquisition) (BDI) of genes (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Karev et al., 2002, 2003, 2004; Koonin et al., 2002; Reed and Hughes, 2004).

For the BDI model proposed by Reed and Hughes (2004), one obtains a stationary cumulative distribution of family sizes whose tail tends asymptotically to a power law (Reed and Hughes, 2004) $P(n) \sim n^{-\alpha}$, with $\alpha = \frac{\rho}{\lambda-\mu}$, where λ is the mutation rate, μ is the extinction rate, and ρ is related to the creation rate of new families. Thus, the exponent of the distribution carries the information about the ratio between family creation ρ and protein creation $\lambda-\mu$, but not about the values of these rates themselves. In general, other properties of the distribution depend also on these rates, so that, in principle, one could use the predictions of the model and the empirical distributions to estimate evolutionary rates (Yanai et al., 2000; Karev et al., 2002).

However, our analysis of the empirical distributions reveals that such estimations are in fact misleading. Because distributions of protein-family sizes are statistically indistinguishable from one another—at least for families of size larger than three—it is not possible to infer *any* evolutionary differences between species from the distributions alone. In fact, our results are compatible with all species having evolved with the same evolutionary parameters. Thus the variability empirically observed in the family size distributions is fully accounted by stochastic fluctuations of the same evolutionary process.

It remains to be understood, however, whether simple BDI dynamics can reproduce the absence of phylogenetic similarities in sets of family sizes we observe in the empirical sets of family sizes (Fig. S-6 in the Supplementary Material). To investigate this matter, we perform a numerical analysis of a dynamical BDI model (see Materials and methods). We select values for the evolutionary rates λ, μ, and ρ, such that the expected exponent for the stationary family size cumulative distribution is 1.4 (or 2.4 for the probability density function as shown in Fig. 4).

To test for evolutionary patterns in family size distributions, we generate a pool of 128 species following a hierarchical phylogenetic tree in which at each divergence we duplicate the number of species (Fig. S-6 in Supplementary Material). We let each species evolve for a time equivalent to 3.9 billion years. Then, as we do for the empirical sets of family sizes, we compare the family size distributions—for family sizes larger than three—of the whole pool of species using the KS test and build the corresponding **R** matrix. Then, we compute the average similarity and distance for groups of between species in the same phylogenetic branch and compare them with the 95% confidence interval of the random expectation.

We perform two types of experiment: (i) we space divergences uniformly in time and (ii) we space divergences logarithmically, so that all divergences occur at an early stage of the total evolutionary time. Remarkably, as observed for the empirical data, we find no evidence for stronger correlations between family size distributions of species closer in the phylogenetic tree (see Figs. S-4 and S-5 in Supplementary Material). Indeed, the ordered **R** matrices are very similar to the empirical data, showing no distinct clusters of species with significantly similar distributions.

Furthermore, to assess whether there are significant additional similarities between species in the same phylogenetic branch, we compute the average similarity and distance for groups of s species, with s=2, 4, 8, 16, 32, and 64, and we compare them with the 95% confidence interval of the random expectation (Fig. S-7 in Supplementary Material and Materials and methods). We find that the fraction of points that fall out of the 95% confidence interval is on average 6.1%. Thus, we conclude that, as observed for the empirical data, the family size distributions are neither significantly similar nor significantly closer in the ordering.

Note that, in our numerical simulations, we know *a priori* that all sets of family sizes we obtain are in fact drawn from the same distribution, and yet we can reproduce the same variability in protein-family size distributions that we observe empirically.

4. Discussion

Our findings strongly support the hypothesis that, when viewed in time scales of the order of billions of years, prokaryotic genomes have evolved with the same average genomic evolutionary rates. Indeed, we find no taxonomic-specific trend in the protein-family size distributions, or at least, none that we could measure using the current taxonomic classification of species. We also find that our analysis is not affected by the fact that we consider parasitic species in our analysis. Therefore, rate variations measured for events over scales of millions of years appear to be temporal variations of the same stochastic process.

How does this finding affect current hypothesis on the shape of the tree of life and the origin of eukaryotes? Our findings contradict any theory which puts forth a dramatic change in evolutionary rates of bacteria and archaea at a genomic level. Currently, there are two main classes of theories for the origin of eukaryotes in which drastic changes in evolutionary rates may be implicit: those surmising the bacterial origin of eukaryotes, and those surmising the existence of an independent eukaryotic ancestor. In what follows, we focus on the first class of theories, because they have been extensively elaborated and offer a clear framework to discuss our findings.

For the class of theories that surmise that eukaryotes and archaea emerged from a common bacterial ancestor (Embley and Martin, 2006), such as the Neomuran theory, the surmised changes in the bacterial ancestor happen at a morphological level. Even though, our analysis is restricted to the evolution of protein families in sequence space, a plausible assumption is that major morphological changes are the result of drastic changes in the genome. In that case, one would expect archaea and bacteria to have different evolutionary rates for protein families. The Neomuran theory hypothesizes that around 850 Myrs ago a free-living bacterium suffered drastic changes,

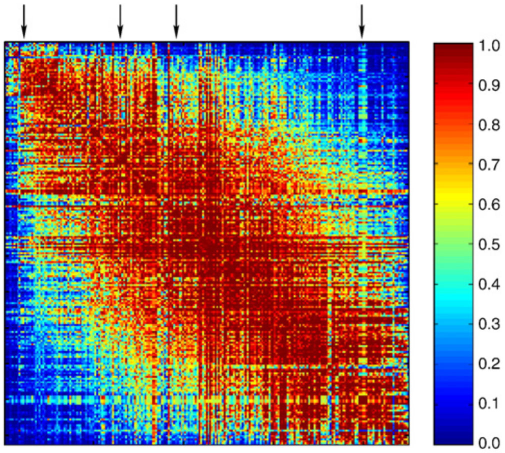

Fig. 6. Ordered **R** matrix including eukaryotes. We show the ordered matrix of KS test r values for the pairwise comparisons of family size distributions for the whole pool of species (archaea and bacteria) plus four eukaryotic species (*P. falciparum*, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*) obtained using the MCL method. We color each matrix element according to its r value, following the same color as in Fig. 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the most important affecting the proteins constituting the cell wall. This event is hypothesized to have given rise to a new clade, the Neomura, which are the common ancestors of archaea and eukaryotes. Such drastic changes in the genome should *in principle* provide a signal detectable by our comparative analysis of archaea and bacteria. We can detect no such signal. Additionally, the comparison of the 229 prokaryotic species with four species of eukaryotes (*P. falciparum*, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*) reveals no differences either (Fig. 6), suggesting that all domains have evolved with the same average genomic rates.

4.1. Limits on the detectability of changes in evolutionary rates

How drastic do changes in gene evolutionary rates need to be in order to provide a detectable signature? To answer this question we have performed numerical experiments using the BDI model described in Section 3.4. In our numerical experiments (Fig. 7 and Fig. S-8 in Supplementary Material), we allow some species to evolve with a larger innovation rate $\rho' > \rho$ during the period of time during which archaea and eukaryotes are hypothesized to have experienced the most drastic changes, approximately between 850 and 580 Myrs ago (Cavalier-Smith, 2002a). We assume that "drastic changes" in the genome happened by creating new protein families needed for performing new functions, thus, we select to keep the same gene duplication and mortality rates while increasing the rate of creation of new families. We have performed numerical experiments for rates $\rho'/\rho = 2$, 4, 8. Our results demonstrate that by comparing protein-family size distributions one would

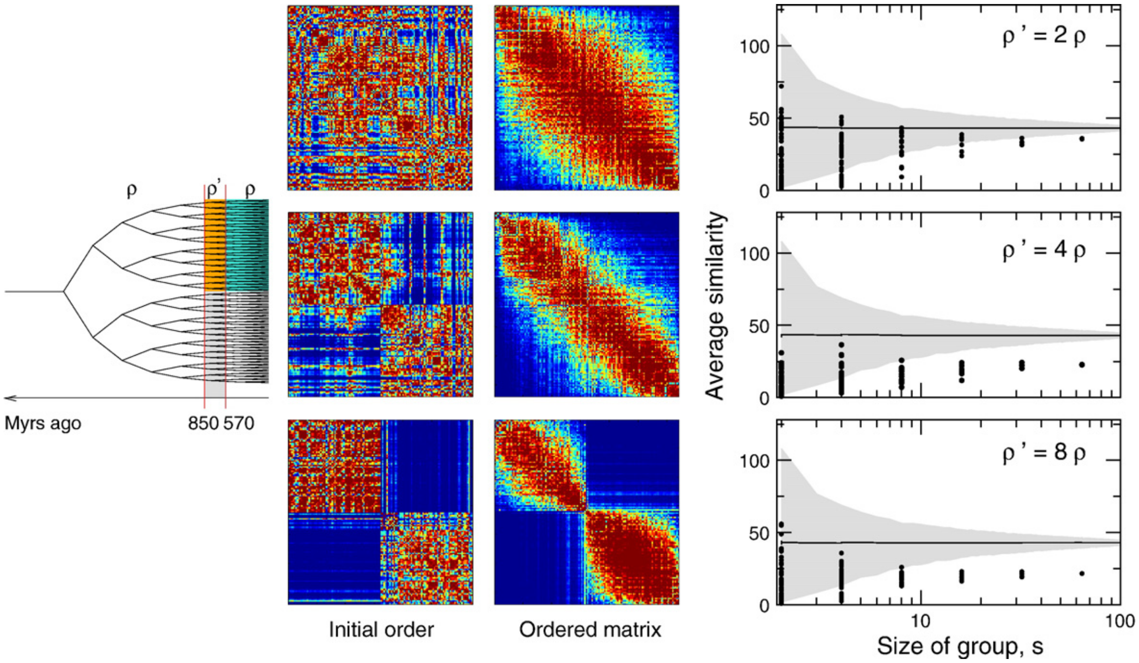

Fig. 7. Testing the Neomuran hypothesis using the BDI model. We perform the following numerical experiment: Starting form a single species with rates $\lambda = 1$, $\mu = 0.1$, and $\rho = 1.405$, we generate a pool of 128 species following the speciation scheme depicted on the left panel. The total simulation time is $t = 5$ time units, which corresponds roughly to 3.9 billion years (see Materials and methods and Fig. S-2 in Supplementary Material for choice of rates and simulation time). We select half of the species (top half highlighted in the left panel) to evolve with a different innovation rate $\rho' > \rho$ during a period of time between 850 Myrs and 570 Myrs ago. We show the results for $\rho' = 2\rho$ (top row), $\rho' = 4\rho$ (central row), and $\rho' = 8\rho$ (bottom row). For each case, we show the **R** matrix ordered following the tree (left panel) and ordered so that species with more similar distributions are close to each other (Materials and methods). Finally, the rightmost column of panels shows, as in Fig. 5, the average distance (according to the best ordering found) for groups of species in the same branch of the tree of sizes 2,…,64. The shadowed area shows the 95% confidence interval for the random expectation. The average distance shows that species ordering divides species into two groups of 64 species, one corresponding to the species that evolve with the same rates and one for species that changed the innovation rate ρ.

be able to detect changes in the rates of evolution of protein families as small as a two-fold increase. Indeed, for really drastic changes in innovation rates ($\rho' = 8\rho$) large differences in genome evolutionary rates should be detectable, even if they happened during a period of time of only 100 Myrs (Fig. S-7 in Supplementary Material).

Acknowledgments

We thank A. Díaz-Guilera, S.M.D. Seaver, and D.B. Stouffer for useful comments and discussions. M. S.-P. and R.G. thank the Fulbright Program and the Spanish Ministry of Education, Culture & Sports. L.A.N.A. gratefully acknowledges the support of the Keck Foundation and the J. S. McDonnell Foundation, and of a NIH/NIGMS K-25 award.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.07.029.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

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, 3389–3402.

Andersson, S.G., Kurland, C.G., 1998. Reductive evolution of resident genomes. Trends Microbiol. 6, 263–268.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2006. GenBank. Nucleic Acids Res. 34, D16.

Brenner, S.E., Hubbard, T., Murzin, A., Chothia, C., 1995. Gene duplications in *H. influenzae*. Nature 378, 140.

Cavalier-Smith, T., 2002a. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int. J. Syst. Evol. Microbiol. 52, 7–76.

Cavalier-Smith, T., 2002b. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int. J. Syst. Evol. Microbiol. 52, 297–354.

Dagan, T., Martin, W., 2006. The tree of one percent. Genome Biol. 7.

Dembo, A., Karlin, S., Zeitouni, O., 1994. Limit distribution of maximal non-aligned two sequence segmental score. Ann. Probab. 22, 2022–2039.

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, 307–311.

Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284, 2124–2128.

Embley, T.M., Martin, W., 2006. Eukaryotic evolution, changes and challenges. Nature 440, 623–630 (URL http://dx.doi.org).

Enright, A.J., Dongen, S.V., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.

Enright, A.J., Kunin, V., Ouzounis, C.A., 2003. Protein families and tribes in genome sequence space. Nucleic Acids Res. 31, 4632–4638.

Fares, M.A., Byrne, K.P., Wolfe, K.H., 2006. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. Mol. Biol. Evol. 23, 245253.

Fitz-Gibbon, S.T., House, C.H., 1999. Whole genome-based phylogenetic analysis of freeliving microorganisms. Nucleic Acids Res. 27, 4218–4222.

Gattiker, A., et al., 2003. Automated annotation of microbial proteomes in SWISS-PROT. Comput. Biol. Chem. 27, 49–58.

Harrison, P.M., Gerstein, M., 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J. Mol. Biol. 318, 1155–1174.

House, C.H., Fitz-Gibbon, S.T., 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J. Mol. Evol. 54, 539–547.

Huelsenbeck, J.P., 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276, 227–232.

Huynen, M.A., van Nimwegen, E., 1998. The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. 15, 583–589.

Ijiri, Y., Simon, H.A., 1977. Skew Distributions and the Sizes of Business Firms, Volume 24 of *Studies in mathematical and managerial economics*. North-Holland, Amsterdam, Netherlands.

Karev, G.P., Wolf, Y.I., Berezovskaya, F.S., Koonin, E.V., 2004. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth–death–innovation models. BMC Evol. Biol. 4, 32.

Karev, G.P., Wolf, Y.I., Koonin, E.V., 2003. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics 19, 1889–1900.

Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, E., 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol. Biol. 2, 18.

Karlin, S., Altschul, S., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U. S. A. 87, 2264–2268.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Koonin, E.V., Tatusov, R.L., Rudd, K.E., 1995. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. Evolution 92, 11921–11925.

Koonin, E.V., Wolf, Y.I., Karev, G.P., 2002. The structure of the protein universe and genome evolution. Nature 420, 218–223.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.

Kurland, C., Collins, L., Penny, D., 2007. The evolution of eukaryotes—response. Science 316, 543–543.

Kurland, C.G., Collins, L.J., Penny, D., 2006. Genomics and the irreducible nature of eukaryote cells. Science 312, 1011.

Liò, P., Goldman, N., 1998. Models of molecular evolution and phylogeny. Genome Res. 8, 1233–1244.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

Martin, W., Dagan, T., Koonin, E., Dipippo, J.L., Gogarten, J., Lake, J., 2007. The evolution of eukaryotes. Science 316, 542–543.

Martin, W., Müller, M., 1998. The hydrogen hypothesis for the first eukaryote. Nature 392, 37–41.

Ochman, H., Wilson, A.C., 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. 26, 74–86.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2002. Numerical Recipes in C: the Art of Scientific Computing, 2 edition. Cambridge University Press, New York.

Reed, W.J., Hughes, B.D., 2004. A model explaining the size distribution of gene and protein families. Math. Biosci. 189, 97–102.

Sales-Pardo, M., Guimerà, R., Moreira, A.A., Amaral, L.A.N., in press. Extracting the hierarchical organization of complex systems. Proc. Natl. Acad. Sci. USA. ArXiv. 0705.1679.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York, NY.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22–28.

Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D.A., Domany, E., 2005. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. Bioinformatics 21, 2301–2308.

Unger, R., Uliel, S., Havlin, S., 2003. Scaling law in sizes of protein sequence families: from super-families to orphan genes. Proteins 51, 569–576.

van Nimwegen, E., 2003. Scaling laws in the functional content of genomes. Trends Genet. 19, 479–484.

Welch, J.J., Bromham, L., 2005. Molecular dating when rates vary. Trends Ecol. Evol. 20, 320–327.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., nad, G.D.S., Tatusova, T.A., Rapp, B.A., 2000. Database resources of the national center for biotechnology information. Nucleic Acids Res. 28, 10–14.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 87, 4576–4579.

Yanai, I., Camacho, C.J., DeLisi, C., 2000. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys. Rev. Lett. 85, 2641–2644.

Zhaxybayeva, O., Lapierre, P., Gogarten, J.P., 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma 227, 53–64.