The Frequency Distribution of Gene Family Sizes in Complete Genomes

Martijn A. Huynen†‡ and Erik van Nimwegen**

*Santa Fe Institute, Santa Fe, New Mexico; †European Molecular Biology Laboratory, Heidelberg, Germany; and
‡Max-Delbrück-Centrum for Molecular Medicine, Berlin-Buch, Germany

We compare the frequency distribution of gene family sizes in the complete genomes of six bacteria (*Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Mycoplasma pneumoniae,* and *Synechocystis* sp. PCC6803), two Archaea (*Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*), one eukaryote (*Saccharomyces cerevisiae*), the vaccinia virus, and the bacteriophage T4. The sizes of the gene families versus their frequencies show power-law distributions that tend to become flatter (have a larger exponent) as the number of genes in the genome increases. Power-law distributions generally occur as the limit distribution of a multiplicative stochastic process with a boundary constraint. We discuss various models that can account for a multiplicative process determining the sizes of gene families in the genome. In particular, we argue that, in order to explain the observed distributions, gene families have to behave in a coherent fashion within the genome; i.e., the probabilities of duplications of genes within a gene family are not independent of each other. Likewise, the probabilities of deletions of genes within a gene family are not independent of each other.

Introduction

One of the main challenges in the interpretation of sequence data of complete genomes is to go from the analysis of the evolutionary process at the level of single genes to that at the level of gene families and of complete genomes. Genes that have a significant similarity to each other are presumed to have evolved from a single ancestral gene and are part of the same gene family. The genomes of *Methanococcus jannaschii*, *Haemophilus influenzae*, and *Escherichia coli* have been shown to contain gene families of various sizes (Brenner et al. 1995; Koonin, Tatusov, and Rudd 1995; Bult et al. 1996). Although a comparative analysis has been done of the gene family sizes in *E. coli* and *H. influenzae* (Tatusov et al. 1996), the frequency distributions of gene family sizes have not been characterized, and the criteria for assignment to a gene family vary among the various classifications published. Here, we do a systematic, comparative analysis of the sizes of gene families versus the frequency per ‘‘size class’’ for the above-mentioned organisms and for *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Methanobacterium thermoautotrophicum*, *Synechocystis* sp. PCC6803, *Saccharomyces cerevisiae*, the vaccinia virus, and the bacteriophage T4. We use a single rigorous method, the Smith-Waterman algorithm, to determine whether two genes belong to the same family. Our analysis of the gene family size distributions of the different genomes shows them to be power-law distributions. We propose models that can explain such distributions. It is not our goal to analyze the evolutionary fate of specific gene families, but, rather, to find possible patterns in the size distributions of the gene families in genomes and to present a general model that could account for such distributions.

Key words: gene family, comparative genome analysis, power-law distribution.

Address for correspondence and reprints: Martijn Huynen, Biocomputing, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany. E-mail: huynen@embl-heidelberg.de.

Mol. Biol. Evol. 15(5):583–589. 1998

Materials and Methods

The Smith-Waterman algorithm (Smith and Waterman 1981), as implemented in the FASTA package (Pearson 1991), was used to compare the protein-coding regions within genomes. A previous analysis of this algorithm, in which its predictions were compared to a similarity analysis based on the 3D structure of proteins, showed that the algorithm produced no false positives for $E < 0.001$ (the E value is the theoretical fraction of false positives within the total set of positives) in comparisons of 320 proteins with less than 40% sequence similarity, and for $E \leq 0.01$, the evolutionary relationships derived from sequence comparisons differed from those derived from 3D structure comparisons by less than 1 case in 100 (Brenner et al. 1995). Analyses done on larger data sets confirm that the E parameter is a reliable parameter for the fraction of false positives the Smith-Waterman algorithm produces (Brenner 1996). The parameters used were the default settings of ‘‘ssearch’’ in the FASTA package, and the same as in Brenner’s analysis: the similarity matrix is BLOSUM50, and the gap penalties were -12 and -2 for the creation and extension of the gap, respectively. We tested whether the results depended on the E value. Reducing the E value from 0.01 to 0.001 generally led to only a small change (<0.1) in the exponents of the distributions and did not change their shape. For the results presented here, we used an E value cutoff of 0.01.

Cluster sizes were determined directly from the output of the Smith-Waterman algorithm; i.e., every sequence represents the core of a potential cluster, and all of the sequences with which it has an E score that is less than the E value threshold are considered part of the same cluster. In a perfect world, if one compares all of the sequences within a genome with each other, one expects to find N occurrences of each cluster of size N; i.e., one occurrence of a cluster for each time one of its members is taken to be the core. In practice, the similarity scores are neither symmetric nor transitive, and the number of sequences that belong to one cluster varies depending on the sequence which is assumed to be at its core (the sequence for which the similarities to the other sequences are calculated). We display the results in two ways: (1) The ''raw'' data, uncorrected for the fact that the large clusters are overrepresented by a factor of their size. This is the nonnormalized probability distribution that a protein sequence that is randomly chosen from the genome is part of a cluster of a specific size. (2) The cluster frequencies are divided by their size to obtain an approximation of the true distribution of the cluster sizes of the genome and are binned exponentially. The exponential binning allows us to get exponents for the distributions that include the frequencies of the larger clusters.

As we show below, the frequency distribution of the cluster sizes follows a power law. Such a distribution features very long tails relative to other distributions. We want to prevent such a distribution from occurring as a side effect of our methodology. Algorithms that include single-linkage clustering have been proposed to determine cluster sizes. Since it is intrinsic to single-linkage clustering that the probability that a sequence is added to a cluster grows with the size of the cluster, such methods tend to increase the frequency of large clusters relative to small clusters and, hence, will generate a bias toward a distribution with a long tail. Our method provides a conservative estimate of the frequency of large clusters; i.e., the long tails of the distributions we observe do not result from a bias in our methodology, and there is no systematic bias in our results. If, due to false negatives in our method, the ''true'' distribution of gene family sizes has longer tails than the one presented here, it would hardly affect our main conclusion because (1) the shape of the distribution depends mainly on the sizes of the relatively small gene families (<16), for which we have the most data points, and (2) our main conclusion is based on the observation that our distribution is significantly more like a power law than an exponential distribution (see *Results* and *Discussion*). Distributions with longer tails than the ones we observed here are even less like an exponential distribution.

The goal of our analysis is not to elucidate all the different clusters and the functions of their proteins, but to get a quantitative insight into the frequency distributions of cluster sizes. The low error rate that was reported for the Smith-Waterman algorithm might cause small changes in the exponents of the distributions, but one does not expect it to change the shapes of the distributions. The analyses of cluster size versus frequency that have been published so far for *H. influenzae* (Brenner et al. 1995), for the (then) partially sequenced genome of *E. coli* (Koonin, Tatusov, and Rudd 1996), and for *M. jannaschii* on http://www.tigr.org/tdb/mdb/mjdb/MJfamilies.html do show a power-law distribution of the cluster sizes versus their frequencies (data not shown).

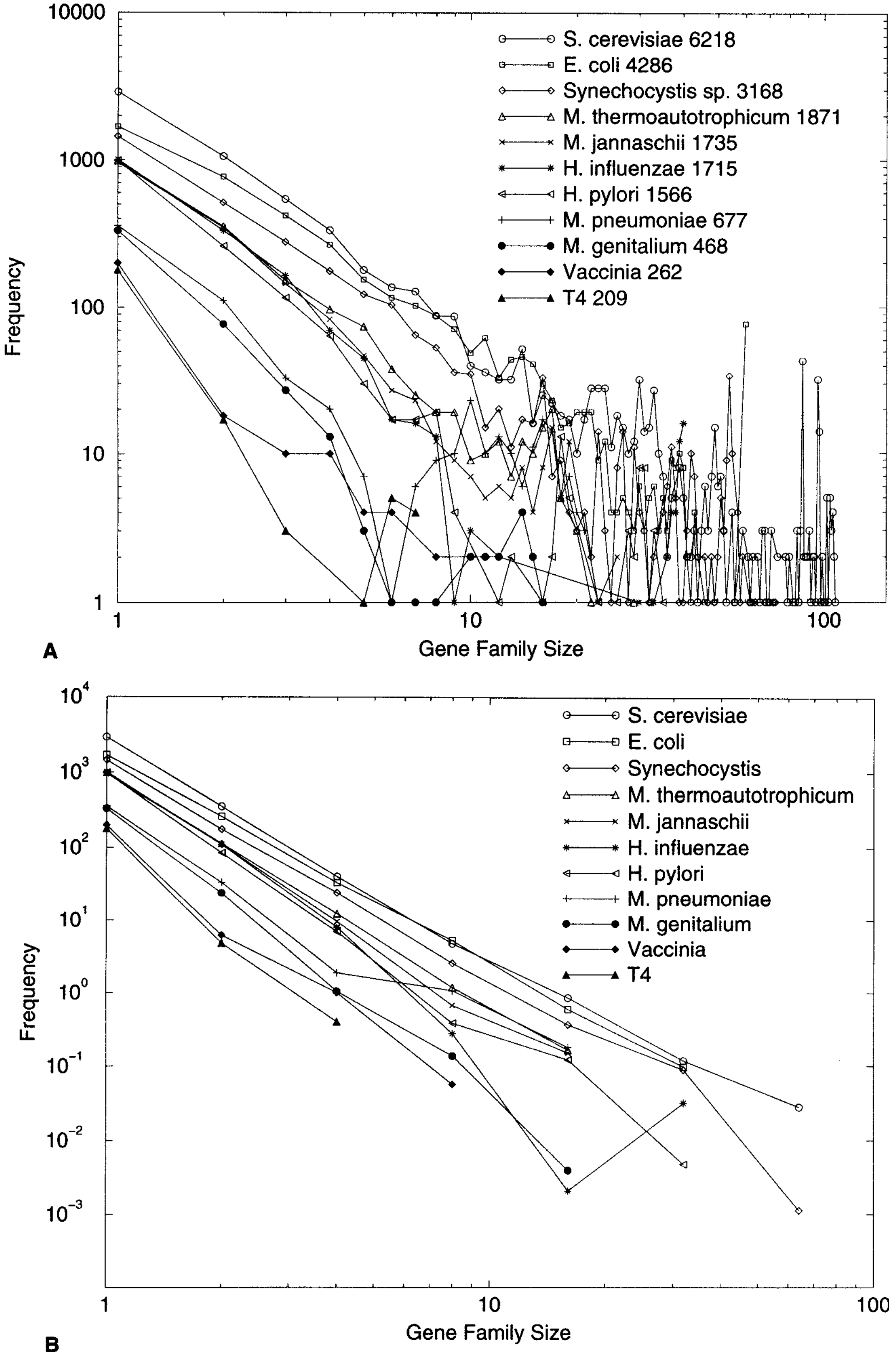
The protein sequence databases we used, with the dates of their last versions, follow: *H. influenzae* (Fleishmann et al. 1995) (August 21, 1996), *M. jannaschii* (Bult et al. 1996) (August 25, 1996), *M. genitalium* (Fraser et al. 1995) (August 16, 1996), and *H. pylori* (Tomb et al. 1997) (August 9, 1997) are from the TIGR database (http://www.tigr.org). The *E. coli* (Blattner et al. 1997) sequences are from the genetics department, University of Wisconsin (ftp://ftp.genetics.wisc.edu/pub/sequence/ecoli.seq) (January 24, 1997). The *Synechocystis* sp. (Kaneko et al. 1996) sequences are from the Cyanobase (http://www.kazusa.or. jp/cyano/cyano.html) (September 9, 1996). The *M. pneumoniae* sequences (Himmelreich et al. 1996) (November 25, 1996) are from NCBI (ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/). The *M. thermoautotrophicum* sequences (Smith et al. 1997) (September 24, 1997) are from Genome Therapeutics Corporation (http://www.cric.com/). The yeast sequences are from the Saccharomyces Genome database (ftp://genome-ftp.stanford.edu/pub/yeast/) (January 28, 1997). The Vaccinia virus sequences (Goebel et al. 1990) (August 1990) are from GenBank. The T4 sequences (Kutter et al. 1994) are from ftp://ncbi.nlm.nih.gov/repository/t4phage (May 23, 1997).

Results

The results of the clustering (fig. 1*A* and *B*) reveal remarkable similarities in the frequency distributions of gene family sizes over the wide variety of genomes analyzed here. All distributions are compatible with a power law, with the possible exception of that of *H. influenzae,* which has relatively few gene families of intermediate size, as was reported in Brenner et al. (1995). Taking the logarithm of both axes and doing a linear regression yields correlation coefficients larger than 0.99 for all species except *H. influenzae* (0.94) and *M. pneumoniae* (0.98) (fig. 1*B*). All the distributions had a significance of fit $P < 0.01$; the genomes with more than 1,500 genes, except for *H. influenzae,* had a significance $P < 5E - 5$. We compared the results of fitting the distribution to a power law to those of fitting it to an exponential distribution. In all cases, the P values were

→

FIG. 1.—*A,* The nonnormalized probability distribution (see *Methods*) for genes in a genome to be members of a gene family of specific size. Shown are the distributions for *S. cerevisiae, E. coli, Synechocystis* sp. PCC6803, *M. thermoautotrophicum, H. influenzae, M. jannaschii, H. pylori, M. pneumoniae, M. genitalium,* vaccinia virus, and the T4 bacteriophage. In the legend of the figure, the species name is followed by the (predicted) number of protein-coding regions in that species. *B,* The frequency–size distribution of gene families with an exponential binning of the gene family sizes. The binning is done logarithmically: family size 1 falls in class 1, family sizes 2 and 3 fall in class 2, family sizes 4, 5, 6, and 7 fall in class 3, etc. The x-coordinates of the binned classes are 1, 2, 4, 8, etc. Taking the logarithm of both axes and doing linear regression yields slopes (correlation coefficients, significances) of −2.81 (0.998, $P < 5E - 8$), −2.84 (1.0, $P < 5E - 9$), −3.17 (0.991, $P < 5E - 6$), −3.17 (1.0, $P < 5E - 7$), −3.27 (0.997, $P < 5E - 5$), −3.62 (0.94, $P < 0.005$), −3.45 (0.996, $P < 5E - 6$), −2.69 (0.979, $P < 0.005$), −4.02 (0.997, $P < 5E - 5$), and −3.8 (0.994, $P < 0.005$) for *S. cerevisiae, E. coli, Synechocystis* sp., *M. thermoautotrophicum, M. jannaschii, H. influenzae, H. pylori, M. pneumoniae, M. genitalium,* and vaccinia, respectively. No regression was done for T4, which, after exponential binning, has only three data points.


10000
1000
100
10
1
Frequency
1
10
100
Gene Family Size
A
S. cerevisiae 6218
E. coli 4286
Synechocystis sp. 3168
M. thermoautotrophicum 1871
M. jannaschii 1735
H. influenzae 1715
H. pylori 1566
M. pneumoniae 677
M. genitalium 468
Vaccinia 262
T4 209
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Frequency
1
10
100
Gene Family Size
B
S. cerevisiae
E. coli
Synechocystis
M. thermoautotrophicum
M. jannaschii
H. influenzae
H. pylori
M. pneumoniae
M. genitalium
Vaccinia
T4

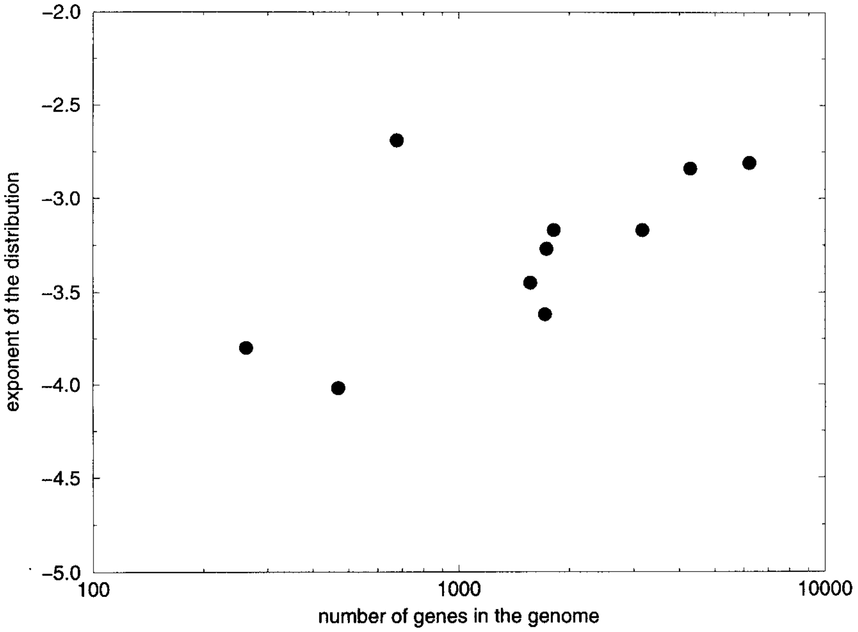


FIG. 2.—The exponent of the power-law distributions from figure 1*B* versus the number of genes in the genome. As the number of genes in the genome increases, the exponent of the distribution becomes larger (the slope become less steep). The correlation coefficient of the logarithm of the number of genes versus the exponent of the distribution is 0.63 ($P < 0.05$). If *M. pneumoniae* is not included, the correlation coefficient becomes 0.91 ($P < 0.01$).

at least a 10-fold smaller for the power-law fit than for the exponential fit. In the genomes with more than 1,500 genes, except for *H. influenzae,* the *P* values for the power-law fit are three to six orders of magnitude smaller than those for the exponential fit.

The results show a trend where, as the number of genes in the genome increases, the exponent of the distribution becomes larger (fig. 2). Thus, an increase in the number of genes leads not only to an increase in the frequency of clusters of all sizes (fig. 1), but also to a relative increase of the number of large clusters over the number of small clusters. A clear exception to this pattern is *M. pneumoniae,* which shows a relatively high frequency of large gene families. As more genomes become available, it will be possible to analyze how general the observed trend is. It has been argued that the large majority of the proteins in life on earth come from only a limited number of families, e.g., 1,000 in Chothia (1992). In such a scenario, one does of course expect a relative increase in the number of large gene families versus the number of small gene families as the number of genes in a genome becomes larger.

Although power-law distributions have been linked to complex processes that show self-organized criticality (Casti 1995), there are several quite simple processes that can lead to these distributions. One of the simplest stochastic dynamical processes that produces power-law distributions is a process with random multiplicative noise repelled away from zero (Kesten 1973; Sornette 1997; Sornette and Cont 1997). We will consider a stochastic process for the evolution of the genes in a gene family that acts coherently on all the genes within one family. At $t = 0$, a gene family is founded by a single ancestor, and through duplications and deletions, the size of this family will fluctuate over time, with the possibility of the family eventually going extinct and disappearing from the genome. The essential feature of our model is that the fluctuations that lead to duplications and deletions are coherent with respect to the genes within one gene family. That is, if a certain gene is likely to duplicate, then all genes of its family are likely to duplicate. The same holds for deletions; i.e., if one gene is likely to be deleted from the genome, then all genes in the family of that gene are as likely (or at least more likely than random genes) to be deleted. The size S_t of the gene family at time *t* will thus fluctuate in the following manner:

$$S_t = \alpha_t S_{t-1} \tag{1}$$

where α_t is a random multiplication factor drawn independently at each time step from some distribution $P(\alpha)$ that is likely to be peaked around $\alpha = 1$ for realistic scenarios. The multiplication factor α_t is determined, for example, by some randomly fluctuating environment. The unit of time is the timescale at which we observe duplication and deletion of genes. For example, the fact that in the *E. coli* genome there are 32 genes that are identical at the amino acid level to at least one other gene in *E. coli* (data not shown) indicates that duplications occur at relatively small evolutionary timescales. We can therefore safely assume that many time steps (eq. 1) have occurred over the history of the genomes that we analyzed. We thus focus in our analysis on the asymptotic behavior of equation (1) for large times *t.* The key point of the model is that all the genes within a family are affected in the same (or at least a similar) way by the environment. The model thus assumes that in consecutive time periods, each gene family tends to expand or shrink as a whole with a random factor α_t. This is in contrast to models in which each gene is independently affected by a fluctuating environment. In such models, the fluctuations in gene family size are on the order of the square root of the number of genes in the gene family.

The distribution of gene family sizes in whole genomes is then the result of many processes (eq. 1) occurring in parallel, for large times t, together with the occasional introduction into the genome of a new gene family of size one. This can be shown under very general conditions to lead to a power-law distribution $P(S)$ of the sizes of the surviving gene families in the genome (see Kesten 1973; Sornette 1997; Sornette and Cont 1997)

$$P(S) = cS^{\gamma}, \tag{2}$$

where c is a normalization constant, and the exponent γ of the power law is given by

$$\gamma = -\left(1 - \frac{\mu_\alpha}{\sigma_\alpha^2}\right), \tag{3}$$

and μ_α and σ_α^2 are the mean and variance of the logarithms of the multiplicative noise factor α

$$\mu_\alpha = \langle \log(\alpha) \rangle \tag{4}$$

and

$$\sigma_\alpha^2 = \langle \log^2(\alpha) \rangle - \mu_\alpha^2. \tag{5}$$

A heuristic derivation of this result can be given in the following way. The size of the gene family at time t is given by a product of t random factors α

$$S_t = \alpha_t \alpha_{t-1} \alpha_{t-2} \ldots \alpha_1, \tag{6}$$

or for the logarithm of the size of the gene family,

$$\log(S_t) = \sum_{i=1}^{t} \log(\alpha_i). \tag{7}$$

By virtue of the central-limit theorem, the distribution of $\log(S_t)$ becomes a normal distribution with average

$$\mu_t = \mu_\alpha t, \tag{8}$$

and variance

$$\sigma_t^2 = \sigma_\alpha^2 t. \tag{9}$$

If we define $y \equiv \log(S_t)$, we thus find for the distribution $P_t(y)$ at large t:

$$P_t(y)dy = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{1}{2}\left(\frac{y - \mu_t}{\sigma_t}\right)^2\right] dy, \tag{10}$$

and by a change of variables we find that S_t is lognormally distributed:

$$P(S_t)dS_t = \frac{1}{\sqrt{2\pi}\sigma_t S_t} \exp\left[-\frac{1}{2}\left(\frac{\log(S_t) - \mu_t}{\sigma_t}\right)^2\right] dS_t. \tag{11}$$

We therefore obtain for the logarithm of the probability density:

$$\log(P(S_t)) = -\log(\sqrt{2\pi}\sigma_t) - \frac{\mu_t^2}{2\sigma_t^2} - \left[1 - \frac{\mu_t}{\sigma_t^2}\right]\log(S_t) - \frac{\log^2(S_t)}{2\sigma_t^2}. \tag{12}$$

Substituting the expressions for μ_t and σ_t, we find:

$$\log(P(S_t)) = -\log(\sqrt{2\pi t}\sigma_\alpha) - \frac{\mu_\alpha^2 t}{2\sigma_\alpha^2} - \left[1 - \frac{\mu_\alpha}{\sigma_\alpha^2}\right]\log(S_t) - \frac{1}{2t\sigma_\alpha^2}\log^2(S_t). \tag{13}$$

The first two terms are independent of S_t and just take care of the normalization of the distribution. For large t, the last term becomes negligible, leaving only the second term, which is linear in $\log(S_t)$, such that the distribution effectively becomes a power-law distribution with exponent $-(1 - \mu_\alpha/\sigma_\alpha^2)$, as expected. For a more rigorous analysis of process (eq. 1) and its resulting distribution, the reader is again referred to Kesten (1973), Sornette (1997), and Sornette and Cont (1997). It is important for the derivation that $\mu_\alpha \leq 0$, since otherwise gene families tend to become infinitely large in the limit of time going to infinity, leaving the process (eq. 1) without a stable limit distribution. Our data are in accordance with this condition. The slopes we observe are smaller than -1.5, which implies that $\mu_\alpha/\sigma_\alpha^2 < -½$. This, in turn, implies that although the size of a gene family may fluctuate for a very large time, eventually each gene family tends to become extinct in a genome. That is, the slopes of the distributions we obtain, together with the stochastic model we propose, suggest that no gene family lives forever in any particular genome, unless other mechanisms prevent certain gene families from going extinct. The latter is actually an essential condition for the derivation of the power-law distribution. If gene families can go extinct, there should be the possibility of occasional introduction of a gene from a new family into the genome, e.g., by horizontal gene transfer. It can be shown (Kesten 1973; Sornette 1997) that the asymptotic distribution of gene family sizes is independent of the form of the influx of new genes. Alternatively, there could be mechanisms that prevent certain gene families from going extinct; i.e., gene family size one acts as a reflecting boundary for those gene families. One can imagine selection acting particularly strongly against the deletion of the last gene in a gene family. In either case, any combination of influx and/or reflective boundary conditions will lead to the same power-law distribution of gene family sizes in the genome.

We observe that the exponent of the distribution becomes larger as the number of genes increases. The variation in the exponent between the various organisms can be explained as a variation in the variance σ_α^2. Keeping the mean μ_α fixed, a smaller variance gives a smaller (steeper) angle. A biological interpretation of this is that as the size of the genome decreases, there will be less room for variation in the sizes of the multiplication events. A similar interpretation is that the competition between gene families for space in the genome is effectively bigger in a smaller genome, which leads to a smaller value of μ_α. That is, the larger competition in smaller genomes leads to shorter average lifetimes of gene families in these small genomes.

Discussion

It may seem contradictory that, in our model, gene families tend to decrease in size on average over all gene families over time, whereas, in evolution, many gene families have started with one member and have grown over time. Note, however, that our model does not prevent any particular gene family from growing. The situation is analogous to the game of roulette where, although on average people will lose, at the end of the evening there will be plenty of winners. Our model does not predict which families will grow, as it predicts the distribution of gene family sizes, rather than the size of any specific family. It has been our goal to define the minimum requirement for a dynamical process that explains the shape of the frequency distribution of gene family sizes. The essential feature of our model is that it describes the dynamics of gene duplication and deletion at the level of the gene family; i.e., the genes within one family behave alike. The most obvious explanation for such coherent behavior is that genes within one family have related functions. As the requirements of this function vary over time, so does the presence of the gene family in the genome. Another explanation for the fact that genes within one gene family behave alike is that they lie clustered on the genome and, as parts of the genome are duplicated or deleted, the genes within one gene family are affected in a similar way. A statistically significant clustering of related genes within the genome has been observed in *H. influenzae* and *E. coli* (Tamames et al. 1997); the overall trend for spatial clustering of genes within the same gene family is, however, only very slight (unpublished data) and cannot explain their coherent behavior.

We believe that it is unlikely that the observed power laws in the distribution of gene family sizes can be explained by a model which does not contain dynamical coherence at the level of gene families. Power-law distributions are distributions with long tails. We studied a number of birth and death processes, containing random duplication and deletion effects along with a random influx of genes. These models, which treat individual genes as independent, all lead to distributions with exponential tails (see Bell [1996] for a birth-and-death model to explain the size distribution of repetitive DNA elements). As shown above, the frequency distribution of gene family sizes is significantly more in accordance with a power-law distribution than with an exponential distribution. As noted earlier, processes that treat the genes as independent lead to fluctuations in gene family that are on the order of the square root of the size of the family. This fact contradicts the power-law shapes of the curves we obtained. Power laws are scale-invariant distributions, which implies that the fluctuations are on the order of the size of the gene family. An alternative model that has been used to explain power-law distributions is so-called self-organized criticality (see Casti 1995 and references therein). The power laws that we observed are very steep, some having exponents smaller than -4. Power laws with this kind of steepness generally cannot be accounted for by self-organized criticality, which typically produces power laws with slopes larger than -2. We therefore argue that the model presented here represents the most general setting by which the observed distributions can be explained. The main conclusion of this paper, then, is that the frequency distribution of gene family sizes can only be explained by a model that explicitly takes the relatedness of the genes within a gene family into account. This coherence of the genes within a gene family supports a shift in analyzing a genome in terms of the presence of gene families rather than of single genes.

Acknowledgments

M.A.H. thanks Richard Palmer, Maya Paczuski, and Kai Nagel for useful discussions. The authors thank the referees for useful suggestions. This work was supported at the Santa Fe Institute by NSF IRI-9320200, DOE DE-FG03-94ER25231, and ONR N00014-95-1-0975.

LITERATURE CITED

BELL, G. I. 1996. Evolution of simple sequence repeats. Comput. Chem. **20**:41–48.

BLATTNER, F. E., G. PLUNKETT III, C. A. BLOCH et al. (14 coauthors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1462.

BRENNER, S. E. 1996. Molecular propinquity: evolutionary and structural relationships of proteins. Ph.D. thesis, Cambridge University.

BRENNER, S. E., T. HUBBARD, A. MURZIN, and C. CHOTHIA. 1995. Gene duplications in *H. influenzae.* Nature **378**:140.

BULT, C. J., O. WHITE, G. J. OLSEN et al. (37 co-authors). 1996. Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. Science **273**:1058–1072.

CASTI, J. L. 1995. Bell curves and monkey languages. Complexity **1**:12–15.

CHOTHIA, C. 1992. One thousand families for the molecular biologist. Nature **357**:543–544.

FLEISHMANN, R., M. ADAMS, O. WHITE et al. (37 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae.* Science **269**:496–512.

FRASER, C. M., J. D. GOCAYNE, O. WHITE et al. (26 co-authors). 1995. The minimal gene complement of *Mycoplasma genitalium.* Science **270**:397–403.

GOEBEL, S. J., G. P. JOHNSON, M. E. PERKUS, S. W. DAVIS, J. P. WINSLOW, and E. PAOLETTI. 1990. The complete DNA sequence of vaccinia virus. Virology **179**:247–266.

HIMMELREICH, R., H. HILBERT, H. PLAGENS, E. PIRKL, B. LI, and R. HERRMANN. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res. **24**:4420–4449.

KANEKO, T., S. SATO, H. KOTANI et al. (24 co-authors). 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. ii. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3**:109–136.

Kesten, H. 1973. Random difference equations, and renewal theory for products of random matrices. Acta Math. **131**: 207–248.

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

———. 1996. *Escherichia coli*—functional and evolutionary implications of genome scale computer-aided sequence analysis. Pp. 177–210 *in* J. P. Gustafson and R. B. Flavell, eds. Genomes of plants and animals. Twenty-first Stadler Genetics Symposium. Plenum Press, New York.

Kutter, E., T. Stidham, B. Guttman et al. (11 co-authors). 1994. Genomic map of bacteriophage T4: the sequence-based map. Pp. 491–529 *in* J. Aram, ed. Molecular biology of bacteriophage T4. ASM Press, Washington, D.C.

Pearson, W. R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics **11**:635–650.

Smith, D. R., L. A. Doucette-Stamm, C. Deloughery et al. (34 co-authors). 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* h: functional analysis and comparative genomics. J. Bacteriol. **17**:7135–7155.

Smith, T., and M. S. Waterman. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

Sornette, D. 1997. Linear stochastic dynamics with nonlinear fractal properties. LANL preprint: cond-mat 9709101.

Sornette, D., and R. Cont. 1997. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. J. Physique I **7**:431–444.

Tamames, J., G. Casari, C. Ouzounis, and A. Valencia. 1997. Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. **44**:66–73.

Tatusov, R. L., A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. Rudd, and E. V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli.* Curr. Biol. **6**:279–291.

Tomb, J.-F., O. White, A. R. Kervalage et al. (39 co-authors). 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature **388**:539–547.

Claudia Kappen, reviewing editor

Accepted January 21, 1998