

## Predictions of Gene Family Distributions in Microbial Genomes: Evolution by Gene Duplication and Modification

Itai Yanai, Carlos J. Camacho,\* and Charles DeLisi

Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215

(Received 2 March 2000)

A universal property of microbial genomes is the considerable fraction of genes that are homologous to other genes within the same genome. The process by which these homologues are generated is not well understood, but sequence analysis of 20 microbial genomes unveils a recurrent distribution of gene family sizes. We show that a simple evolutionary model based on *random* gene duplication and point mutations fully accounts for these distributions and permits predictions for the number of gene families in genomes not yet complete. Our findings are consistent with the notion that a genome evolves from a set of precursor genes to a mature size by gene duplications and increasing modifications.

PACS numbers: 87.23.Kg, 05.40.-a, 87.10.+e

Knowledge of the complete genomic sequences of organisms provides an invaluable starting point towards an understanding of the organization and evolution of genomes. To date, the public databases contain the genomes of over 20 microbial organisms, with many more in the making. This information provides the raw data that enables the study of genomes as an interrelated self-organized collective [1–3], as opposed to a mere set of individual genes.

Genes of common ancestry are known as homologues. Intragenomic homologues, or paralogues, account for nearly half the number of genes of most genomes, including, for example, those of *E. coli* and *B. subtilis*. Based on a sequence similarity analysis, paralogues may be clustered into families. The number of gene families rapidly decreases as a function of family size, which ranges from 1 to more than 70 [4–6]. About half the genes have no detectable homologues (singlets), with about  $\frac{1}{8}$ th as many families of only two genes (doublets). Some large families constitute functionally related proteins, such as the ubiquitous ABC (ATP-binding cassette) transporters [7]. A remarkable observation is that the distribution of gene family sizes has a recurrent shape across all known microbial genomes [8,9]. That such a regular pattern can categorically characterize the organization of genes into families across the kingdoms suggests that a common underlying process dominates genome evolution, notwithstanding the many and varied mechanisms for generating diversity.

The process by which paralogues are generated can be understood only in terms of gene duplication [10–12]. In a recent study, Huynen and van Nimwegen [9] postulated that only a stringent mechanism where gene duplication and deletion behave coherently within a gene family could explain the observed distribution of gene families. Indeed, based on this assumption, these authors predicted power-law distributions which are in qualitative agreement with the genomic data. Here we propose a simpler dynamical model for the evolution of gene families which embodies the basic key features of genome evolution: accumulation

of viable (or adapted) but otherwise randomly added point mutations within genes and a random mechanism for gene duplication. Together with a mutation threshold for the detection of homology and a single genome-size dependent parameter  $N_0$  mimicking an initial set of genes prone to duplication, the model recovers the observed distribution of gene family sizes for all genomes.

*Model of genome evolution.*—Genes are identified by a single index  $i = 1, \dots, N$ , where  $N$  is the total number of genes. The number of mutations that gene  $i$  has accumulated at time  $t$  is denoted by  $G(i, t)$ . We say that genes  $i$  and  $j$  belong to the same family if and only if they are separated by less than  $\epsilon$  random mutations.

At some initial condition  $t = t_0$  we assume there are  $N = N_0$  independent viable genes with  $|G(i, 0) - G(j, 0)| > \epsilon$ , for  $i \neq j = 1, \dots, N_0$ . This condition implies that these genes are distinguishable either because they were original *ab initio* genes or shared a common ancestry but have since diverged beyond similarity.

The genome evolves according to the following rules:

(i) Gene mutations: At every time step a gene, say  $k$ , is randomly chosen from the current pool of genes, and a single selected or neutral mutation is added to it such that  $G(k, t + \Delta t) = G(k, t) + 1$ . Similarly, at each time step, we update a counter for the amount of mutations to be added to a new gene upon duplication  $\eta(t + \Delta t) = \eta(t) + 1.5/(N_0 N)^{1/2}$ , with  $\eta(0) = 0$  (see below).

(ii) Gene duplication: Every  $M$  time steps a gene  $k$  ( $\leq N$ ) is picked at random and a viable mutated (or adapted) copy is inserted into the genome, thereby increasing  $N$  by one. The new gene is denoted by index  $i = N$ , and differs from its predecessor  $k$  by  $G(i, t) = \eta(t)$  mutations.

Steps (i) and (ii) are repeated until  $N$  reaches the total number of genes,  $N_{\text{GEN}}$ , in the genome of each specific microbe.

The genome size increases by gene duplication such that  $N$  increases by one every  $M$  point mutations. In the spirit of Darwin's hypothesis of "descent with change," each

duplicate has an inherent modification with respect to its ancestor, specified in terms of a given number of random mutations. By following the relationships between genes and their ancestors we build the patterns of branching descent still discernible as gene families. Since the observed paralogue distribution in microbial genomes is based upon sequence similarity among genes, we cluster gene families by computing the exact number of random mutations separating a gene from its paralogues. We say that gene  $i$  belongs to a family of size  $l$ , if  $i$  has  $l - 1$  genes from which it is separated by some number of mutations less than a given threshold  $\epsilon$ .

We compare our simulations with an all-against-all pairwise sequence alignment of all genes in the available genomes using the program BLAST (Basic Local Alignment Search Tool) [13]. For a given gene, the number of alignments having an  $E$  value of  $10^{-10}$  or lower within its own genome determines the family size of that gene. The  $E$  value is related to the probability that the similarity score occurs by chance. The paralogue distribution is similar for  $E$  values as large as  $10^{-3}$  [4].

The only fitting parameter for each genome is the number of genes  $N_0$ . All other parameters are fitted once and then fixed across all genomes: gene duplication occurs every  $M = 49$  point mutations; and, the homology threshold is fixed to  $\epsilon = 720$ . This value corresponds to the number of random mutations that a protein of roughly 400 residues can withstand before its similarity (BLAST  $E$  value) score, when aligned to the native sequence, becomes greater than  $10^{-10}$ . The results are quite robust with respect to changes in these parameters (see below).

We summarize our results by comparing the predicted and observed number of gene families as a function of family size  $l$  for 20 microbial genomes. As shown in Fig. 1, the simulations agree well with the overall shape of the observed distributions, including the varying slopes of different size genomes [9]. Moreover, the predicted and observed number of singlets ( $l = 1$  intercept of the distributions), which range from 391 for *M. genitalium* (MG) to 3263 for *S. cerevisiae* (YE), agree to within 3%. Given the limited data currently available, we cannot reliably differentiate a power law [9] from a sum of two exponential distributions.

The simple model also accounts for specific features of the paralogue distribution. In particular, the model mimics quite well the higher level of fluctuations observed for smaller genomes, while converging to a limiting shape for larger genomes. Short of the large family clusters of ABC transporters (see below), the largest family size predicted by the model (which increases with genome size) coincides with those observed. In accord with the data, we find that the tails of the distributions have gaps where no gene families of certain sizes are found (these gaps are not seen on the scale of Fig. 1). The positions of the gaps as well as the deviations from the average distribution depend upon the given realization of the random process—i.e., each evolutionary path yields a unique set of gene families.

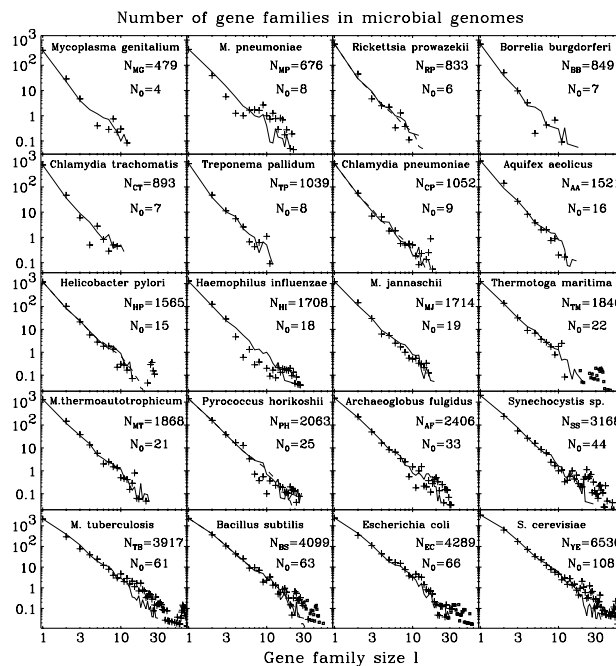


FIG. 1. The number of gene families for 20 microbial genomes [22]. These numbers correspond to the total number of genes belonging to a family of size  $l$  divided by  $l$ . Each plot indicates the name of the genome, the genome size denoted by  $N_{GEN}$  (where GEN in the figure corresponds to a two letter acronym for each genome), and the number of initial genes  $N_0$ . The + symbols correspond to the observed data. The solid lines correspond to a typical realization of the model. We emphasize that we do not show the best realization that fit the data, but to assure an unbiased sampling all the simulations shown invoke consecutive seeds for the sequence of random numbers. Also, for comparison we show five gene family distributions averaged over 50 realizations (dashed lines in RP, CP, HP, PH, and BS). Note that, due to the gaps in the tail of the distribution, for large  $l$  the average and typical realization differ significantly. Families of multidomain ABC transporter genes for *T. maritima*, *B. subtilis*, and *E. coli* are indicated with a square symbol.

Strikingly, the free parameter  $N_0$  scales almost linearly with the corresponding present size of the genome  $N_{GEN}$  (see Fig. 2). This dependence allows us to predict the paralogue distribution for any genome by extrapolating  $N_0$  from the estimated number of genes in the genome in question. For instance, Fig. 2 indicates  $N_0$  for two genomes soon to be released, as well as a predicted number of families. Values of  $N_0$  range between  $N_0 \approx 4-6$ , comparable to the size of a small virus, to  $N_0 \approx 100$  for yeast. Although we do not have an *a priori* mechanism for the apparent relationship between  $N_0$  and  $N_{GEN}$ , the notion that the definitive feature differentiating organisms is the presence of a series of new genes is not new [14]. Thus, it is not unreasonable to expect a parameter like  $N_0$  to be the only determining factor for the size of gene families.

The model is robust enough to allow different randomly chosen divergence rates for each gene. The latter was confirmed by having each gene accumulate a preassigned number of mutations drawn from a Poisson distribution.

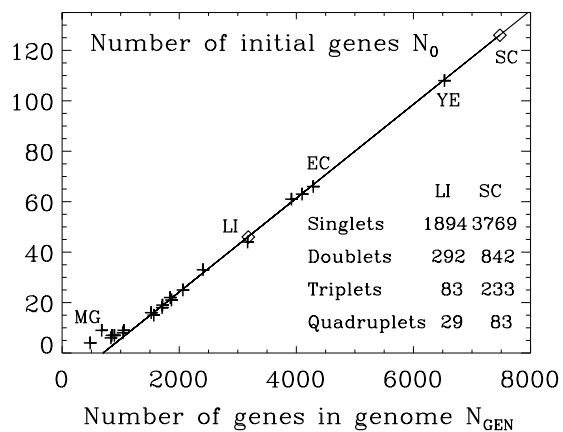


FIG. 2. The number of initial genes  $N_0$  as a function of the genome size  $N_{\text{GEN}}$ . The solid line indicates a linear interpolation of the data  $N_0 = 0.0186N_{\text{GEN}} - 13$ . The  $\times$  symbols indicate the genomes shown in Fig. 1. The diamond symbols are predictions for *L. innocua* (LI) and *S. coelicolor* (SC). Also shown are some predictions for the expected number of gene families based on an estimate of the total number of genes  $N_{\text{LI}} = 3180$  ( $N_0 = 46$ ) and  $N_{\text{SC}} = 7476$  ( $N_0 = 126$ ). Error bars are around 10%.

This assignment of varying divergence rates did not change the predictions of the model. Similarly, we have checked that our results remain the same if instead of a constant  $M$  the rate of gene duplication follows a Poisson distribution of mean equal to  $M$ .

Our results are also robust with respect to changes in the parameters of the model. As expected the choice of BLAST  $E$  value and homology threshold  $\epsilon$  are highly correlated; e.g., a decrease in the similarity threshold to an  $E$  value of  $10^{-50}$  can easily be accounted for by a decrease in  $\epsilon$  to 540. The value of  $M$  is also closely related to  $\epsilon$ . Indeed, equally good fits to the experimental data are obtained for ratios of  $\epsilon/M \approx 15$ . On the other hand, as much as a 30% decrease (increase) in  $N_0$  can be compensated by a 20% decrease (increase) in  $M$  though the size of the largest family is also increased (decreased) by a similar amount. In summary, for any given  $\epsilon$  there will be a more or less optimal value for  $M$  and  $N_0$ .

The model allows only for duplications of single genes. Thus, the predicted distributions do not account for multidomain homologues. This explains in part some of the deviations from the data that we observed at the tail of some distributions. For example, the small cluster of large families indicated for TM, BS, and EC in Fig. 1, which are often separated from the main distribution, are the triple-domain ABC transporters [7,15,16]. The family size of these multidomain genes, which are precisely two- to threefold the size of the predicted single domain family, are overestimated due to the integration into one single large family of both the multidomain paralogues as well as the paralogues of their single domain units (to be published elsewhere).

In this paper we do not attempt to identify precise biological mechanisms for gene duplication. Indeed, our

working assumptions are quite general and could be rationalized in various ways. Here, we assume that (a) genomes evolve from a set of initial genes; (b) the duplication rate is proportional to the number of point mutations; (c) a duplicated gene inherits some mutations with respect to its ancestor; and, (d) the effective number of mutations upon duplication increases with time.

(a) How the parameter  $N_0$  relates to the universal ancestor is not clear. If, as suggested by Woese [17], this ancestor consists of a diverse community of cells which over time refined into a smaller number of increasingly complex cell types,  $N_0$  could be rationalized as some initial set of genes prone to duplicate. We note that varying degrees of common ancestry could be rationalized by having species evolve from a similar set of precursor genes. Apart from these ‘‘precursor’’ genes however, genomes might also include some extra set of housekeeping genes less susceptible to viable duplications, e.g., ribosomal proteins and *t*RNA synthetases. Without loss of generality, the model could set aside some of these ubiquitous genes with either a very low duplication rate or mutation rate (such that highly similar duplicated genes might not be advantageous). As long as these genes do not generate new families these additional precursors will not change the predicted paralogue distributions.

(b) The number of point mutations is for the most part proportional to time, and we expect the same to hold true for gene duplication. (c) Processes by which microbial organisms exchange genes, or gene domains, and undergo genetic recombination are known [2,18]. Thus, the notion that a duplicated gene might inherit some modifications, which we model as a given number of random mutations, is well founded.

(d) Early in the evolutionary process the duplicated gene and its ancestor differ by a relatively small number of random mutations. We assume that the increase in complexity (genome size) leads to a higher loss of fidelity upon duplication. Indeed, one can argue that since the duplication of genetic information is generally not advantageous, a duplicated gene identical to the parent gene must rapidly adapt to its own functional niche in order to survive [10]. With time, the increased competition for space in the genome should lead to the faster adaptation of the duplicated gene. Other possible rationalizations are that new genes are imported from another colony, or specie [2], which has been mutated independently. We envisage that eventually the effective mutations inherited by the duplicated genes (finite in size) should increase more slowly. However, late in evolution, most new genes will be mutated beyond the threshold of homology detection leading to the emergence of singlets. This implies that for very long simulations (i.e.,  $N > N_{\text{GEN}}$ ) the distributions do not change much from those in Fig. 1, except for  $l = 1$ , since the loss of similarity by point mutations alone is very small. The increasing modifications upon gene duplication plotted in Fig. 3 are, perhaps, the simplest functional form that satisfies the aforementioned general principles.

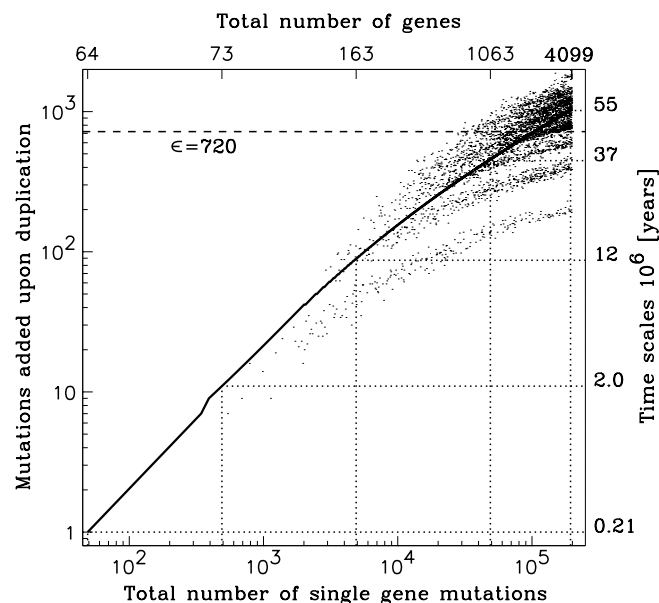


FIG. 3. The number of mutations added upon duplication  $\eta(t)$  in the simulation of the evolution of *B. subtilis* is indicated with a solid line ( $N_0 = 63$ ). The value of  $\eta(t)$  was initially developed as a stochastic term (shown as dots) which scaled with the number of homologues of the duplicated gene, and whose goal was to limit the size of gene families. Both functions  $\eta(t)$  yield the same results. The effective dependence on  $N^{-1/2}$  for the rate increase of  $\eta(t)$  in the first step of the model is reminiscent of the suggestion by Kimura [11] and Ohta [12] that mutations per generation should scale as  $1/N^\beta$ , with  $\beta \approx 0.5$  [19]. To give the readers some feeling for the relative time scales involved in our minimal evolutionary process, dotted lines indicate time scales for the 1st, 10th, 100th, 1000th, and (last) 4036th duplication. These times are estimated rather arbitrarily by assuming  $\Delta t = 1/(NK)$ , where  $K = 10^{-8}/T$  [18] is the mutation rate per gene and  $T = 1$  day is the generation time.

The simple model presented here does not include a method for gene deletion, a recognized major force in genome evolution [20]. Within the framework of our model one could test different mechanisms for gene deletion improving the accuracy of the predictions. It is important to stress, however, that adding genome-specific selective pressures into the model would also lead to new free parameters, which would significantly undermine the validation of the model based on the data.

Selection plays a fundamental role in the evolution of genomes. Instances of the positive selection of gene duplications are well known [21]. Despite selection's powerful role, however, the recurrent distribution of gene family sizes shown in Fig. 1 suggests a general organizing mechanism distinct from selective pressures. Here, we have focused our attention on the random component of gene duplication and point mutations. In doing this, we have assumed that at the genome level, selective pressures do not present a drastic effect to the shape of the distribution

of gene family sizes. Nevertheless, a selective mechanism could still account for the size of some gene families. Our results are consistent with the notion that the optimization provided by natural selection is not pervasive enough as to destroy the seemingly randomlike nature of the shape of the overall distribution of gene family sizes.

We argue that a genome evolves from a set of precursor genes to a mature size by random gene duplications and increasing modifications. Based on these minimal assumptions, we find that an essentially random process can describe and predict the recurrent paralogue distribution observed across kingdoms.

We thank L. Levitin, I. Grosse, J. Valentine, S.R. Kimura, and Z. Weng for stimulating conversations. This work was supported by Grants from the DOE (No. DE-F602-96ER62263) and NSF (No. DBI-9904834). I. Y. received support from NSF.

\*To whom all correspondence should be addressed.

Email address: ccamacho@bu.edu

- [1] R. F. Doolittle, *Nature (London)* **392**, 339 (1998).
- [2] W. F. Doolittle, *Science* **284**, 2124 (1999).
- [3] A. Danchin, *Curr. Opin. Struct. Biol.* **9**, 363 (1999).
- [4] S. E. Brenner *et al.*, *Nature (London)* **378**, 140 (1995).
- [5] F. Kunst *et al.*, *Nature (London)* **390**, 249 (1997).
- [6] F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
- [7] C. F. Higgins, *Annu. Rev. Cell Biol.* **8**, 67 (1992).
- [8] P. P. Slonimski *et al.*, in *Proceedings of Microbial Genomes II*, Hilton Head, South Carolina, 1998.
- [9] M. A. Huynen and E. van Nimwegen, *Mol. Biol. Evol.* **15**, 585 (1998).
- [10] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Heidelberg, 1970).
- [11] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, England, 1983).
- [12] T. Ohta, in *Molecular Evolution and Polymorphism*, edited by M. Kimura (National Institute of Genetics, Mishima, Japan, 1977).
- [13] S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
- [14] J. M. W. Slack, P. W. H. Holland, and C. F. Graham, *Nature (London)* **361**, 490 (1993).
- [15] K. Tomii and M. A. Kanehisa, *Genome Res.* **8**, 1048 (1998).
- [16] Y. Quentin, G. Fichant, and F. Denizot, *J. Mol. Biol.* **287**, 467 (1999).
- [17] C. Woese, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854 (1998).
- [18] T. D. Brock, M. T. Madigan, J. M. Martinko, and J. Parker, *Biology of Microorganisms* (Prentice-Hall, Englewood Cliffs, New Jersey, 1997).
- [19] L. Chao and D. E. Carr, *Evolution* **47**, 688 (1993).
- [20] S. G. E. Andersson and C. Kurland, *Trends Microbiol.* **6**, 263 (1998).
- [21] L. Patthy, *Protein Evolution* (Blackwell Science, London, 1999).
- [22] Listed at the TIGR Microbial Database, [www.tigr.org](http://www.tigr.org)