

## A DISCRETE MODEL OF EVOLUTION OF SMALL PARALOG FAMILIES

JERZY TIURYN\* and DAMIAN WÓJTOWICZ†

*Institute of Informatics, Warsaw University,  
Banacha 2, 02-097 Warsaw, Poland*  
\*tiuryn@mimuw.edu.pl  
†dami@mimuw.edu.pl

RYSZARD RUDNICKI

*Institute of Mathematics, Polish Academy of Sciences  
and Institute of Mathematics, Silesian University,  
Bankowa 14, 40-007 Katowice, Poland*  
rudnicki@us.edu.pl

Received 10 March 2006  
Revised 22 November 2006  
Communicated by Z. Agur

We introduce and analyze a simple probabilistic model of genome evolution. It is based on three fundamental evolutionary events: gene loss, duplication and accumulated change. We are mainly interested in asymptotic size distribution of small paralogous gene families in a genome. This is motivated by previous works which consisted in fitting the available genomic data into, what is called, *paralog distributions*. This formalism is described as a discrete-time Markov chain. The formulas for equilibrium paralog family sizes are derived. Moreover, we show that when probabilities of gene removal and duplication are small and close to each other, then the resulting distribution is close to logarithmic distribution. Some empirical results for microbial genomes are presented.

*Keywords:* Genome evolution; paralogous genes; gene family size distribution; Markov chain.

### 1. Introduction

The seminal Ohno's work *Evolution by Gene Duplication*<sup>18</sup> shows that gene duplication is a fundamental feature of evolution. It creates the redundancy necessary to free one copy of a gene to evolve a new function via accumulation of gene changes/mutations. Any two genes in a genome, that have evolved through a duplication from a single ancestral gene, are called *paralogs*. We do not discuss here this important issue of deciding which genes are paralogs. An in-depth discussion of this matter can be found in Ref. 7. Here we assume that all genes have already been clustered into groups of pairwise paralogous genes. It should be mentioned that

many authors make this clustering in different ways, see for example Refs. 4, 5, 10, 21 and 25.

A genome is a dynamic collection of genes which change in time, due to diverse biochemical processes that constantly act on it. These processes include gene duplication and loss, point mutation, recombination, gene conversion, rearrangement, DNA repair, translocation and horizontal transfer. However, the role of duplication remains unquestioned in the whole genome evolution process. Families of paralogs constitute a significant part of all genes in a genome, about half of the genes have detectable homologous gene.<sup>10–12,21,25</sup> In this work we propose a simple model of genome evolution in the spirit of Kimura,<sup>15</sup> i.e. in the total absence of selective pressure. We are fully aware that such a purely neutralistic model cannot be truly realistic. However, it can be very useful as a basis for further discussion. Our model is based on three fundamental evolutionary events: gene duplication, loss and accumulated change. We define here these notions to avoid possible misunderstandings:

- *gene duplication* — an event in which one gene gives rise to two genes which cannot be operationally distinguished between themselves, which remain in the same genome and are therefore paralogs;
- *gene loss* — an event which leads to a removal of the gene from the genome;
- *accumulated change* — an event (or cumulative series of events like mutations, rearrangements, recombinations,...) which leads to such a modification of a sequence that the resulting gene is no longer similar to its parental ancestor and therefore is no longer classified as a paralog.

Mathematical analysis of the model allows us to rigorously study the problem of size distribution of paralogous gene families in genome. This problem was studied in a comparative genomics in late '90s.<sup>10,21</sup>

### 1.1. *Related work*

In 1998, Slonimski *et al.*<sup>21</sup> and independently Huynen and Nimwegen<sup>10</sup> have compared the frequency distribution of gene families in the complete genomes of several species whose genomes have already been sequenced. Both papers came up with two different claims. Slonimski *et al.*<sup>21</sup> propose that the sizes of the gene families versus their frequencies follow logarithmic distributions (i.e. the probability of being an  $i$ -element cluster is  $C \cdot \theta^i / i$ , where  $0 < \theta < 1$  and  $C$  is a normalizing constant), whereas Huynen and van Nimwegen<sup>10</sup> state power law distribution (i.e. the probability is  $D \cdot i^{-\gamma}$ , where  $\gamma > 1$  and  $D$  is a suitable normalizing constant). In 2001 Jordan *et al.*<sup>11</sup> have analyzed 21 completely sequenced prokaryotic genomes and concluded that the logarithmic approximation fits the observed distributions slightly better than the power law approximation. However, they have also noticed that it is difficult to reach any meaningful biological conclusion concerning the shape of the gene family size distribution by merely fitting the data. None of the above cited papers proposes a model which could explain the observations. The first such

model was designed in 2000 by Yanai *et al.*<sup>25</sup> This genome evolution model is based on random gene duplication and point mutations. The main result of the paper consisted in showing that it is possible for each of the 20 microbial genomes to tune the parameters of the model so that the obtained distribution matches closely the observed paralog distribution of the genome. However, any mathematical analysis of the model was not given in that paper.

To our knowledge the first paper which propose a model of genome evolution together with complete mathematical analysis of the equilibrium frequencies of domain families is Karev *et al.*<sup>12,13</sup> The model in that paper is based on three elementary processes: domain birth (duplication), domain death (deletion), and domain innovation (acquisition via horizontal transfer, or emergence from a non-coding sequence), the so-called *BDIM model*. The external source of new genes (innovation) serves the purpose of stabilizing the asymptotic behavior of the model. Karev *et al.*<sup>12</sup> show in their paper that depending on relative rates of duplication and death of domains in families (these rates depend on the size of the family and are constant in time) one obtains various equilibrium distributions, including logarithmic and power law.

The BDIM model and the model presented in this paper differ in three important respects: (a) BDIM model is a continuous time process described by a finite system of differential equations, while our model is a discrete-time Markov chain with infinitely many states. The advantage of working with differential equations is that there are available strong analytical tools which make the analysis of the model easier. On the other hand, the advantage of working with discrete time Markov chains is that they allow computer simulations which are helpful in establishing working hypothesis which can be further verified in a strict manner. To our knowledge this paper gives the first mathematical analysis of the asymptotic distribution for a discrete-time model of evolution of gene families. (b) BDIM model sets a fixed upper bound on the maximal size of a family, while our model allows families of arbitrary unbounded size. It is not clear what are the consequences for the resulting distribution if one bounds the maximal size of the family. In technical terms bounding the size results in a finite system of differential equations (as this is the case for the BDIM model), while without the bound the system becomes infinite. (c) Finally, in the BDIM model there is an external source of new genes (invention), while our model is a “closed system” in the sense that there are no new genes coming from outside. New gene families are being created in our model via accumulated change. It would be interesting to see what happens when both features are present in the model: innovation and change.

A model in the spirit of this paper but without the mechanism of accumulated change was analyzed in Ref. 23. It is shown there that the asymptotic distribution in that model is *geometric* (i.e. the probability of being an  $i$ -element family is  $G \cdot \theta^i$ , where  $0 < \theta < 1$  and  $G$  is a normalization constant). However, the geometric distribution does not fit the genomic data. It will follow therefore, from the results

of this paper, that the accumulated change event may be an important mechanism in genome evolution.

It should be noticed that there are several other general approaches to the whole genome evolution<sup>1-3,9,14,17,20,24</sup> as well as analyses of evolution restricted to a group of genes.<sup>8,16</sup>

## 1.2. The duplication, loss and change (DLC) model

Now we describe more formally our model of duplication, loss and change (DLC) of genes. In order to express the concept of gene homology, we will assume that all genes we are working with are colored. The convention is that genes with the same color are *homologous* and genes of different colors are not homologous in the operational sense of the term (*vide supra*). We will assume that an unlimited supply of colors is given. A *genome* is a finite set of colored genes. A *gene family* in a genome is the set of all genes of that genome which have the same color. We group families according to their size. For any  $i > 0$ , let  $\mathcal{C}_i$  denote the class/cluster of all  $i$ -element families of the genome.

Evolution of genome is modeled by a Markov chain with discrete time. States of the Markov chain are infinite sequences  $(s_i)_{i \geq 1}$  of non-negative integers such that all but finitely many  $s_i$ 's are zero. A state  $(s_i)_{i \geq 1}$  represents a genome in which for every  $i \geq 1$ , the number of  $i$ -element gene families is  $s_i$ . The initial state is  $(1, 0, 0, \dots)$ , i.e. a genome with only one gene.

The model is parametrized by three positive reals:  $p$ ,  $a$  and  $b$ , subjected to the condition that  $p + a \cdot p + b \cdot p \leq 1$ . A transition from a genome  $\mathcal{G}$  to  $\mathcal{G}'$  in one step is based on the following process of *evolution* which is performed simultaneously and independently for each gene of  $\mathcal{G}$ . A gene, which is subjected to the process of evolution is:

- *removed (lost)* from the genome with probability  $p_L = p$ . For  $i > 1$ , removal of a gene from a family of class  $\mathcal{C}_i$  moves this family to class  $\mathcal{C}_{i-1}$ ; removal of a gene from one-element family results in elimination of this family from the genome. A removed gene is eliminated permanently from the pool of all genes.
- *duplicated* with probability  $p_D = a \cdot p$ . A new gene is created in the genome and this gene inherits the color of its parent, i.e. duplication of a gene in a family of class  $\mathcal{C}_i$  moves this family to the class  $\mathcal{C}_{i+1}$ .
- *changed* with probability  $p_C = b \cdot p$ . It changes its color to a new one, not present in the genome, i.e. the gene starts a new one-element family and is removed from the family to which it belonged.
- *unchanged* with the remaining probability  $p_U = 1 - p_L - p_D - p_C$ .

It is natural to assume that  $p_U \gg p_L + p_C + p_D$ , i.e. that  $p$  is very small. We view the parameters  $a$  and  $b$  as constants, which define a class of processes with varying  $p$ . It is also reasonable to assume that  $a$  is close to 1 since otherwise the

model becomes very unstable: with  $a \ll 1$  all genes are removed very fast, while  $a \gg 1$  implies a fast exponential blow up of the genome.

**Remark.** Our model is based on the assumption that each gene can independently be subjected to elementary evolutionary events. For small probability parameter  $p$  one could consider a simplified model in which in one evolutionary step only one gene can be lost, duplicated or changed. In this case the probability of two or more simultaneous events is negligibly low.

**1.3. Main contributions of the paper**

We are interested in the asymptotic distribution of ratios  $\frac{\#C_i}{\sum_j \#C_j}$  ( $i = 1, 2, \dots$ ), as the number of evolution steps tends to infinity. We use the method of generating functions to find it. Recall that if  $S = (s_i)_{i \in I}$  is a sequence of reals and  $I$  is a subset of non-negative integers, then the *generating function* for  $S$  is a function  $f$  defined by the power series  $f(x) = \sum_{i \in I} s_i x^i$ . When  $S$  is a probability distribution, the generating function  $f$  is called *probability generating function* (see Ref. 6).

We give, in Theorem 2.1, a functional characterization of the probability generating function  $f_{p,a,b}$  for the asymptotic distribution of family class size ratios. This theorem works under the assumption that  $a < 1 + b$ , which is quite general as it covers the cases of  $a$  being equal, or slightly larger, or slightly smaller than 1, provided  $b > 0$ . Notice that *a priori* it is not obvious at all that such a distribution exists. The function  $f_{p,a,b}$  is non-elementary, i.e. cannot be defined by a closed expression.

The second result of the paper, Theorem 2.2, shows that the function  $f_{p,a,b}$  has a limit  $f_{a,b}$ , as the probability parameter  $p$  tends to  $0^+$  (the parameters  $a$  and  $b$  are fixed). Moreover,  $f_{a,b}$  is uniquely described by a linear differential equation. The probability generating function  $f_{a,b}$  is also non-elementary, but the differential equation gives rise to recurrence equations which define the asymptotic distribution of gene family sizes (see Corollary 2.1).

Moreover, if parameter  $p$  is sufficiently small and  $a$  is close to 1 and  $a < 1 + b$ , then the asymptotic distribution of ratio of class sizes is close to the logarithmic

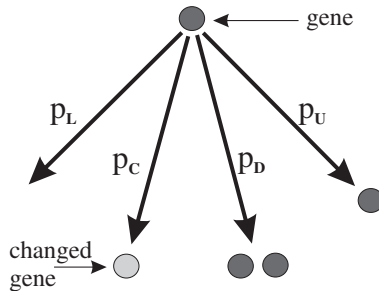


Fig. 1. One step of gene evolution.

distribution  $P(i) = C \cdot \theta^i / i$ , where  $\theta = 1/(1 + b)$  and  $C = (-\log(1 - \theta))^{-1}$  is the normalizing constant. A precise formulation of this result is given in Corollary 2.2.

The paper is organized as follows. Section 2 presents the main results of the paper. Due to space limitations and readability most of the proofs are moved to an appendix — see Appendices A and B (for the case  $a = 1$ ). Section 3 contains the experimental results for five bacterial genomes. Concluding remarks are presented in Sec. 4.

**2. Results and Discussion**

Let  $S_i^{(n)}$  be an integer-valued random variable which represents the number of  $i$ -element gene families after  $n$  steps of evolution process ( $i \geq 1, n \geq 0$ ). For  $n = 0$ , we have  $(S_1^{(0)}, S_2^{(0)}, S_3^{(0)}, \dots) = (1, 0, 0, \dots)$ . Now, for each  $n \geq 0$  and  $i \geq 1$ , we label all  $i$ -element gene families at step  $n$  with different integer values  $l = 1, 2, \dots, S_i^{(n)}$  (order of the labeling is insignificant). We also define auxiliary random variables  $X_{i,l}^{(n)}$  and  $Y_{i,l}^{(n)}$ , where  $1 \leq l \leq S_i^{(n)}$ .

Variable  $X_{i,l}^{(n)}$  takes value  $j \in \{0, 1, 2, \dots\}$  if the  $l$ th family passes from  $i$ th to  $j$ th class in the transition from step  $n$  to step  $(n + 1)$ . Let  $q_{i,j}$  be the probability that an  $i$ -element gene family gets size  $j$  as a result of duplications, removals and changes applied independently to each gene of this family with probabilities described in Introduction. It can be easily shown that

$$q_{i,j} = \sum_{d=0}^{\lfloor j/2 \rfloor} \binom{i}{d, j - 2d} p_U^{j-2d} p_D^d (p_L + p_C)^{i-j+d},$$

where  $\lfloor j/2 \rfloor$  is the largest integer less than or equal to  $j/2$ , and  $\binom{i}{i_1, i_2}$  equals  $\frac{i!}{i_1! i_2! (i - i_1 - i_2)!}$  if  $i_1, i_2 \geq 0$  and  $i_1 + i_2 \leq i$ , or 0 otherwise. It follows from the Markov property of the sequence  $(S_i^{(0)})_{i \geq 1}, (S_i^{(1)})_{i \geq 1}, \dots$  that  $\mathbb{P}(X_{i,l}^{(n)} = j | (S_i^{(0)})_{i \geq 1}, \dots, (S_i^{(n)})_{i \geq 1}) = q_{i,j}$ , which is independent of  $n$  and  $l$ .

The other variable  $Y_{i,l}^{(n)}$  takes value  $k \in \{0, 1, 2, \dots\}$  if exactly  $k$  genes of the  $l$ th family change color in the transition from step  $n$  to step  $(n + 1)$ . It follows that  $\mathbb{P}(Y_{i,l}^{(n)} = k | (S_i^{(0)})_{i \geq 1}, \dots, (S_i^{(n)})_{i \geq 1}) = \binom{i}{k} p_C^k (1 - p_C)^{i-k}$ , where  $\binom{i}{k}$  equals  $\frac{i!}{k!(i-k)!}$  if  $0 \leq k \leq i$ , or 0 otherwise. Again, this probability is independent of  $n$  and  $l$ .

Now, for  $n \geq 0$  and  $j \geq 1$ , we can derive the number of  $j$ -element gene families at step  $(n + 1)$  from all families at step  $n$ :

$$S_1^{(n+1)} = \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} \mathbb{I}_{\{1\}} \left( X_{i,l}^{(n)} \right) + \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} Y_{i,l}^{(n)}, \quad \text{for } j = 1$$

and

$$S_j^{(n+1)} = \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} \mathbb{I}_{\{j\}} \left( X_{i,l}^{(n)} \right), \quad \text{for } j > 1,$$

where  $\mathbb{I}_A$  is an indicator of set  $A$ . Notice that the case  $j = 1$  is slightly different from the remaining cases because of singletons which, in addition to the possibility of arriving from other classes by adjusting their size, could have also been created by accumulated change.

The conditional expectation of  $S_j^{(n+1)}$  given  $(S_1^{(n)}, S_2^{(n)}, \dots)$  is obtained by the following computation for  $j = 1$ :

$$\begin{aligned} \mathbb{E}\left(S_1^{(n+1)} \mid S_1^{(n)}, S_2^{(n)}, \dots\right) &= \sum_{i \geq 1} \mathbb{E}\left(\sum_{l=1}^{S_i^{(n)}} \mathbb{I}_{\{1\}}(X_{i,l}^{(n)}) + \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} Y_{i,l}^{(n)} \mid S_1^{(n)}, S_2^{(n)}, \dots\right) \\ &= \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} \mathbb{P}\left(X_{i,l}^{(n)} = 1 \mid S_1^{(n)}, S_2^{(n)}, \dots\right) \\ &\quad + \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} \sum_{k \geq 0} k \cdot \mathbb{P}\left(Y_{i,l}^{(n)} = k \mid S_1^{(n)}, S_2^{(n)}, \dots\right) \\ &= \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} q_{i,1} + \sum_{i \geq 1} \sum_{l=1}^{S_i^{(n)}} p_C i \\ &= \sum_{i \geq 1} S_i^{(n)} q_{i,1} + p_C \sum_{i \geq 1} S_i^{(n)} i \end{aligned}$$

and analogously for  $j > 1$ :

$$\mathbb{E}\left(S_j^{(n+1)} \mid S_1^{(n)}, S_2^{(n)}, \dots\right) = \sum_{i \geq 1} S_i^{(n)} q_{i,j}.$$

Now, for  $n \geq 0$  and  $j \geq 1$ , we can derive the expected number  $E_j^{(n+1)}$  of  $j$ -element gene families in the genome after  $(n + 1)$  steps, assuming the initial state being a one-element genome, i.e.  $E_j^{(n+1)} = \mathbb{E}S_j^{(n+1)}$ . It follows from the equation  $\mathbb{E}S_j^{(n+1)} = \mathbb{E}(\mathbb{E}(S_j^{(n+1)} \mid S_1^{(n)}, S_2^{(n)}, \dots))$  that

$$E_1^{(n+1)} = \sum_{i \geq 1} E_i^{(n)} q_{i,1} + p_C \sum_{i \geq 1} E_i^{(n)} i, \quad \text{for } j = 1$$

and

$$E_j^{(n+1)} = \sum_{i \geq 1} E_i^{(n)} q_{i,j}, \quad \text{for } j > 1.$$

Therefore, we have obtained an infinite system of equations for  $(E_j^{(n)})_{j \geq 1}$  with  $(E_1^{(0)}, E_2^{(0)}, E_3^{(0)}, \dots) = (1, 0, 0, \dots)$ . Let  $Q = (q_{i,j})_{i,j \geq 1}$  and  $N = (n_{i,j})_{i,j \geq 1}$ , where  $n_{i,1} = i$  and  $n_{i,j} = 0$  for  $i \geq 1, j > 1$ . Thus, we can rewrite the above system as  $(E_1^{(n+1)}, E_2^{(n+1)}, E_3^{(n+1)}, \dots) = (E_1^{(n)}, E_2^{(n)}, E_3^{(n)}, \dots)(Q + p_C \cdot N)$ . It follows further that, for all  $n \geq 0$ ,  $(E_j^{(n)})_{j \geq 1}$  can be represented as a product of matrices:

$$(E_1^{(n)}, E_2^{(n)}, E_3^{(n)}, \dots) = (1, 0, 0, \dots)(Q + p_C \cdot N)^n. \tag{2.1}$$

Notice that starting with  $K > 0$  one element gene families results in family classes which are  $K$  times bigger uniformly across all sizes. It follows then that the resulting distribution of ratios is not affected.

For  $n \geq 0$  and  $i \geq 1$ , we define the distribution of ratio of expected values by

$$p_{p,a,b,i}^{(n)} = \frac{E_i^{(n)}}{\sum_{j=1}^{\infty} E_j^{(n)}}.$$

We have already mentioned that we want to find the asymptotic distribution of ratios  $(p_{p,a,b,i}^{(n)})_{i \geq 1}$ , as  $n$  tends to infinity. Notice that *a priori* it is not clear that such a distribution always exists since  $\lim_{n \rightarrow \infty} E_i^{(n)} = 0$  for all  $i \geq 1$ .

Let  $f_{n,p,a,b}(x)$  be the probability generating function for the distribution  $(p_{p,a,b,i}^{(n)})_{i \geq 1}$ , i.e.  $f_{n,p,a,b}(x) = \sum_{i=1}^{\infty} p_{p,a,b,i}^{(n)} x^i$ . If a sequence of probability generating functions  $\{f_{n,p,a,b}\}_{n \geq 0}$  converges, as  $n \rightarrow \infty$ , then the limit is a probability generating function  $f_{p,a,b}$  for the asymptotic distribution. We have the following result.

**Theorem 2.1.** *Let  $p, a, b > 0$  be such that  $p + ap + bp \leq 1$  and  $1 + b > a$ . The limit  $f_{p,a,b}$  of  $f_{n,p,a,b}$ , as  $n \rightarrow \infty$ , exists and is an analytic function with the radius of convergence  $\frac{1+b}{a}$ . Moreover, it satisfies the following equation:*

$$f_{p,a,b}(\varphi(x)) = \gamma \cdot f_{p,a,b}(x) + \frac{bp}{c_{p,a,b}(0)} \cdot (1 - x) + 1 - \gamma, \tag{2.2}$$

where  $\varphi(x) = p + bp + (1 - p - ap - bp)x + apx^2$ ,  $\gamma = 1 - p + ap$  and  $c_{p,a,b}$  is a function defined in Lemma 2.2.

**Proof.** Observe that  $\varphi(x) - \varphi(0)$  is a generating function of the first row of  $Q$ . We have to investigate the generating function  $\phi_{n,p,a,b}$  of the first row of matrix  $(Q + p_C \cdot N)^n$ . We have the following result:

**Lemma 2.1.** *The generating function  $\phi_{n,p,a,b}$  for the first row of the matrix  $(Q + p_C \cdot N)^n$  satisfies*

$$\phi_{n,p,a,b}(x) = (\varphi^{(n)}(x) - \varphi^{(n)}(0)) + bp\gamma^{n-1} \sum_{i=0}^{n-1} (\varphi^{(i)}(x) - \varphi^{(i)}(0)) / \gamma^i, \tag{2.3}$$

where  $\varphi^{(i)}$  is  $i$ -fold composition of  $\varphi$  with itself (with  $\varphi^{(0)}$  being the identity function).

Now we can define the generating function  $f_{n,p,a,b}$  in a different way:

$$f_{n,p,a,b}(x) = \frac{\phi_{n,p,a,b}(x)}{\phi_{n,p,a,b}(1)}. \tag{2.4}$$

Existence of the limit function  $f_{p,a,b}(x) = \lim_{n \rightarrow \infty} f_{n,p,a,b}(x)$  follows from the following equality:

$$f_{n,p,a,b}(x) = 1 - \frac{(\phi_{n,p,a,b}(1) - \phi_{n,p,a,b}(x)) / \gamma^n}{\phi_{n,p,a,b}(1) / \gamma^n}$$

and the following lemma:

**Lemma 2.2.** *Let  $a, b > 0$  be fixed, where  $1 + b > a$ , and  $p > 0$ . Then for every  $x \in (-\frac{1-ap}{ap}, \frac{1+b}{a})$  there exists a limit*

$$\lim_{n \rightarrow \infty} \frac{\phi_{n,p,a,b}(1) - \phi_{n,p,a,b}(x)}{\gamma^n} = c_{p,a,b}(x)$$

and  $|c_{p,a,b}(x)| < \infty$ . Moreover, for  $x \geq 0$  we have:  $c_{p,a,b}(x) = 0$  iff  $x = 1$ .

To show that the convergence is almost uniform in the open disk  $|z| < \frac{1+b}{a}$  in complex numbers we take any  $0 < r < \frac{1+b}{a}$ . We have for all  $|z| < r$  the following inequalities

$$|f_{n,p,a,b}(z)| \leq f_{n,p,a,b}(|z|) \leq f_{n,p,a,b}(r). \tag{2.5}$$

The first inequality follows from non-negative coefficients of  $f_{n,p,a,b}$  and the second from the fact that  $f_{n,p,a,b}$  is increasing function in the interval  $[0, \frac{1+b}{a})$ . Since the sequence  $f_{n,p,a,b}(r)$  is converging, the sequence of functions  $f_{n,p,a,b}(z)$  is uniformly bounded on the disk  $|z| < r$ . According to the Vitali's Theorem a uniformly bounded sequence of analytic functions which is converging in some set with a limit point in this disk, covers uniformly to an analytic function  $f_{p,a,b}(z)$  on each compact subset of this disk. This implies that the sequence  $f_{n,p,a,b}(z)$  covers uniformly to an analytic function  $f_{p,a,b}(z)$  on each compact subset of the disk  $|z| < \frac{1+b}{a}$ .

Finally, condition (2.2) follows from Lemmas 2.1 and 2.2, but the actual proof which is a little complicated, is moved to Appendix A.3. □

It follows from the theory of analytic functions that  $f_{p,a,b}$  is the unique analytic function which satisfies (2.2) and the two constraints:  $f_{p,a,b}(0) = 0$  and  $f_{p,a,b}(1) = 1$ . In this sense the above theorem gives a complete characterization of the generating function for the distribution of interest. Unfortunately, it cannot be expressed by elementary functions.

However, we are interested in the behavior of DLC model for small values of  $p$ . This is a reasonable assumption since the probabilities of removal, change and duplication should indeed be small, because these events are really rare for short time intervals. The next result shows that for small values of  $p$ , the behavior of pgf  $f_{p,a,b}$  can be described by a linear ordinary differential equation. This equation will be further used to derive the main result of this paper.

**Theorem 2.2.** *Let  $p, a, b > 0$  and assume that  $1 + b > a$ . The limit  $f_{a,b}$  of  $f_{p,a,b}(x)$ , as  $p \rightarrow 0^+$ , exists and is differentiable in the open interval  $(0, 1)$ . Moreover, it satisfies the following differential equation*

$$f'_{a,b}(x) = \frac{(1-a)(1-f_{a,b}(x))}{(1-x)(1+b-ax)} + \frac{C_{a,b}}{1+b-ax}, \tag{2.6}$$

where  $C_{a,b} = (\int_0^1 \frac{z^{(a-1)/(1-a+b)}}{1+b-az} dz)^{-1}$ .

The proof of Theorem 2.2 needs some auxiliary results, so it is moved to Appendix A.4.

We know that  $f_{a,b}(0) = 0$ , so from the above theorem we have a unique characterization of the generating function  $f_{a,b}$ . In addition, another constraint should also be satisfied:  $f_{a,b}(1) = 1$ , whenever function  $f_{a,b}$  is a generating function of probability distribution.

Theorem 2.2 is a generalization of the results presented in our previous paper.<sup>23</sup> When  $b = 0$  (i.e. there is no gene change), we have  $C_{a,0} = 0$ , and Eq. (2.6) reduces to Eq. (31) from that paper. Equation (31) in Ref. 23 was the main step towards the principal result (Theorem 7) presented there, i.e. the statement that the size distribution of paralog families in a model without accumulated change asymptotically approaches geometric distribution. In a similar way Theorem 2.1 in this paper generalizes the corresponding functional equation in Ref. 23. From the mathematical point of view, the introduction of gene change will result in the distribution having a less substantial tail than geometric distribution, because accumulated change creates one-element gene families from the bigger families.

We would like to solve the differential equation (2.6) with the constraint  $f_{a,b}(0) = 0$ . Unfortunately, the general solution contains non-elementary functions, thus we omit it here. Instead we derive a system of recurrence equations defining the distribution corresponding to  $f_{a,b}$ . Let  $(p_{a,b,i})_{i \geq 1}$  be the probability distribution determined by  $f_{a,b}$ , i.e.

$$f_{a,b}(x) = \sum_{i=1}^{\infty} p_{a,b,i} \cdot x^i.$$

Thus we obtain  $p_{a,b,i}$  by taking the  $i$ th derivative of  $f_{a,b}$  defined by (2.6) in zero. This leads to the following result.

**Corollary 2.1.** *Let  $a, b \geq 0$  and assume that  $1 + b > a$ . Then the distribution  $(p_{a,b,i})_{i \geq 1}$  is defined by the following system of recurrence equations:*

$$p_{a,b,1} = \frac{C_{a,b}}{1+b} + \frac{1-a}{1+b},$$

$$p_{a,b,2} = \frac{1}{2} \frac{a}{1+b} p_{a,b,1} + \frac{1}{2} \frac{1-a}{1+b} (1 - p_{a,b,1}), \tag{2.7}$$

$$p_{a,b,i+1} = \frac{((1+b+a)i - (1-a))p_{a,b,i} - a(i-1)p_{a,b,i-1}}{(i+1)(1+b)}, \quad \text{for } i \geq 2. \tag{2.8}$$

Notice that if  $b = 0$  and  $a < 1$ , then the above system defines a geometric distribution:  $p_{a,b,i} = (1-a)a^{i-1}$  for  $i \geq 1$ , as it is expected due to Ref. 23. On the other hand, if  $b > 0$  and  $a = 1$ , then this system defines the logarithmic distribution:  $p_{a,b,i} = C_{1,b} \left(\frac{1}{1+b}\right)^i / i$  for  $i \geq 1$ , where  $C_{1,b} = \left(-\log\left(1 - \frac{1}{1+b}\right)\right)^{-1}$ . We are mainly interested in the latter situation which is biologically significant. Thus we can rewrite the above system (now for any  $a$  and  $b$ ) in a different way

to see better its logarithmic behavior (proof of the latter equation can be found in Appendix A.5).

$$p_{a,b,1} = \frac{C_{a,b}}{1+b} + \frac{1-a}{1+b}, \tag{2.9}$$

$$p_{a,b,i} = \frac{a}{1+b} \cdot \frac{i-1}{i} \cdot p_{a,b,i-1} + \frac{1}{i} \cdot \frac{1-a}{1+b} \left( 1 - \sum_{j=1}^{i-1} p_{a,b,j} \right) \quad \text{for } i > 1. \tag{2.10}$$

We can see that the first sum component of (2.10) behaves logarithmically, i.e. when we omit the second component ( $a = 1$ ), then the system defines logarithmic distribution. Thus we have that the limit distribution is close to the logarithmic distribution in the presence of accumulated gene change, when the probability of gene loss is close to the probability of gene duplication. It is easy to notice that for large  $i$ 's, the obtained size distribution of paralog families is indistinguishable from a logarithmic distribution, but for small  $i$ 's, it deviates significantly, with the magnitude of the deviation depending on  $|a - 1|$ .

Now we can state the final result of the present paper.

**Corollary 2.2.** *Let  $(p_{a,b,i})_{i \geq 1}$  be the probability distribution determined by  $f_{a,b}$ , i.e.  $f_{a,b}(x) = \sum_{i=1}^{\infty} p_{a,b,i} \cdot x^i$ . Then for every  $i \geq 1$  and for every  $b > 0$*

$$p_{1,b,i} = \frac{1}{-\log(1-\theta)} \cdot \frac{\theta^i}{i},$$

where  $\theta = \frac{1}{1+b}$ .

The case  $a = 1$  is important from biological point of view because it describes the situation when the total number of genes is constant in time on average (when  $a \neq 1$ , then the total number of genes increases or decreases exponentially fast in time). Corollary 2.2 suggests that the size distribution of paralogous gene families in a genome should be described by a logarithmic distribution  $C \cdot \theta^i / i$ , with  $\theta$  depending only on the ratio of the probability of accumulated change and removal (= duplication). An alternative direct proof of Corollary 2.2 was suggested to us by one of the referees. It is given in Appendix B.

**Remark.** When  $p_L > p_D$ , the size of the genome becomes zero for large  $n$  with probability one. Indeed, let  $Z_n = \sum_{i=1}^{\infty} i S_i^{(n)}$  be the total number of genes after  $n$  steps of evolution process. Then  $\mathbb{E}Z_n = (1 - p_L + p_D)^{n-1} \mathbb{E}Z_1$  and since  $1 - p_L + p_D < 1$  we get  $\lim_{n \rightarrow \infty} \mathbb{E}Z_n = 0$ . Since the values of the process  $Z_n$  are non-negative integers we obtain  $\lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) = 1$ .

### 3. Experimental Results

In order to compare the observed families of paralogous genes which occur in species with the values predicted by our model we have examined five bacterial genomes: *Burkholderia mallei* ATCC:23344, *Geobacter sulfurreducens* PCA, *Pseudomonas putida* KT2440, *Bacillus anthracis* Ames, *Shewanella oneidensis* MR-1.

The bacterial paralogous families were taken from TIGR-CMR<sup>19</sup> web service at [http://www.tigr.org/tigr-scripts/CMR2/paralog.info\\_form.spl](http://www.tigr.org/tigr-scripts/CMR2/paralog.info_form.spl). As it was observed by many researchers<sup>10,12,21</sup> the distribution of large families of paralogous genes in organisms is very uneven: large families may span hundreds of classes, most of them empty. An explanation proposed by Ref. 21 is that large families are subject to natural selection, a feature which is not present in our model. For this reason some researchers<sup>21,22</sup> restrict analysis of families to small classes (cluster size 2 through 6), while others<sup>10,12</sup> group families into bins, each containing a certain pre-specified minimal number of families. In our analysis we choose the former method, i.e. we consider only families which have between 2 and 6 members. The number of such families for each organism considered is given in Table 1. We omit all single-member families from considerations because the information about singletons is incomplete in this data set (the number of detected singletons is too small). The observed data was fitted to the logarithmic distribution and the distribution parameter  $\theta$  was chosen to minimize the value of Pearson's  $\chi^2$  test. For each genome, before we evaluated the  $\chi^2$  test we grouped the expected paralog family frequencies into bins, each containing at least 10 genes. For all analyzed genomes  $P(\chi^2)$  for this model was larger than 0.05, i.e. no significant difference between the observed and predicted values was detected. The values of parameter  $\theta$  and the goodness-of-fit  $P(\chi^2)$  are given in Table 1. Plots of the observed data versus predicted number of families are given in Fig. 2. It should be noticed that the maximal likelihood method gives us almost the same  $\theta$  for each genome.

It may appear from studying Table 1 that constant  $\theta$  for bacterial genomes is grouped around 0.7. We have also experimented (data not shown) with another method of clustering, TribeMCL,<sup>5</sup> which resulted in different clustering for which the best  $\theta$  was around 0.5. This experiment clearly indicates that the shape of the distribution of paralog families under study may critically depend on the method of clustering paralogous genes. This calls for further investigation, especially in the light of lack of "golden standards" in this area.

We have also performed a similar analysis for the power law distribution. The resulting  $P(\chi^2)$  was similar to the corresponding value for the logarithmic distribution (data not shown). Difference between the goodness-of-fit for both distributions

Table 1. Paralogous families in bacterial genomes and the parameter of the best-fit model.

Bacterial genome	Number of families	$\theta$	$P(\chi^2)$
<i>Bacillus anthracis Ames</i>	815	0.673	0.288
<i>Burkholderia mallei</i>	703	0.734	0.710
<i>Geobacter sulfurreducens</i>	581	0.702	0.297
<i>Pseudomonas putida</i>	745	0.764	0.464
<i>Shewanella oneidensis</i>	586	0.687	0.170

Note: Bacterial genome families are taken from TIGR Comprehensive Microbial Resource.<sup>19</sup>

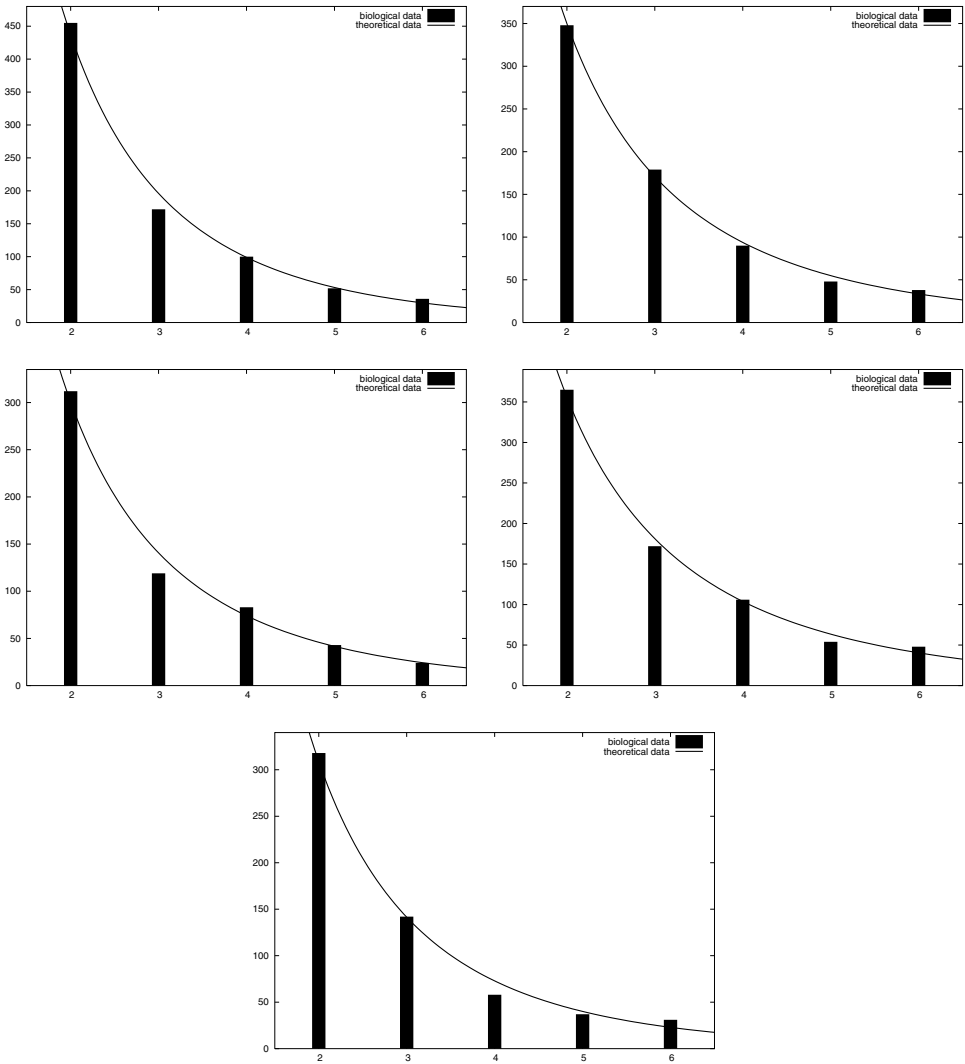


Fig. 2. Fit of empirical gene family size distribution to the logarithmic distribution ( $x$ -axis: paralog family size,  $y$ -axis: number of gene families).

was not essential. However, it seems that the power law distribution has better goodness-of-fit than the logarithmic distribution when we consider larger family sizes.

#### 4. Conclusions

Here we present a mathematical description of the size distribution of paralog families encoded in genomes for a simple but a very natural model of evolution, which includes three types of events: gene removal, duplication and accumulated change.

The paper presents the first, to our knowledge, mathematical analysis of the asymptotic distribution of gene families in the discrete time approach. Genome evolution is a very complicated stochastic process which involves many events in addition to the ones considered in this paper. We do not claim that our model is the most accurate description of this process. It is the simplicity of our model which allows us to mathematically analyze it, and yet the theory behind it is quite involved mathematically. It would be interesting to see how other evolutionary events (like gene innovation proposed in Ref. 12), when introduced into the model affect the asymptotic distribution. For example removing gene change from the model results in geometric distribution.<sup>23</sup>

Another interesting topic of research is to investigate the role of changing the rate of shrinking and expanding the size of a family, as a function of the size. In our model this rate depends on the size, but other dependencies may also be considered. We also ignore in our approach the possibility of dependence of rates of elementary events on such geometrical properties of the genome as gene length, location of a gene in the genome, or genome size. However, the most important advantage of this model is its simplicity that allows us to find the relationship between probabilities of accumulated change and gene duplication/loss by only studying the parameter of distribution of paralogous families. It is practically impossible to check experimentally this relationship.

In order to obtain a better goodness-of-fit for gene family size distributions for all family sizes which are present in a genome, one may consider a mixture of logarithmic distributions. This can be justified in our model by assuming that genes of a genome are divided into disjoint groups, each governed by a different evolutionary process, i.e. DLC process with different probabilities. These groups reflect the functional importance of gene families. For example, families that are responsible for life processes of an organism possibly do not show propensity to high change over time. We plan to investigate such a model in the future.

## Appendix A. Proofs of Some Statements

It is useful to extend the concept of a state  $(s_i)_{i \geq 1}$  of Markov chain by introducing the quantity  $s_0$  which represents the number of genes which have been removed. Thus we expand the matrices  $Q$  and  $N$  by adding a row and a column indexed by 0. From now on let  $Q$  and  $N$  represent the expanded matrices. We define  $q_{i,0} = (p_L + p_C)^i$ ,  $q_{0,j} = 0$  and  $n_{i,0} = n_{0,i} = 0$  for  $i \geq 0$  and  $j \geq 1$ . Now Eq. (2.1) is replaced by

$$(E_0^{(n)}, E_1^{(n)}, E_2^{(n)}, \dots) = (0, 1, 0, \dots)(Q + p_C \cdot N)^n, \quad (\text{A.1})$$

where  $Q$  and  $N$  represent the expanded matrices.

We also introduce some useful notation for matrices. For  $k \geq 0$ , let  $w_k(x) = \sum_{j=0}^{\infty} w_{k,j} x^j$  be the generating function for the  $k$ th row of a matrix  $W = (w_{i,j})_{i,j \geq 0}$ . Then the matrix  $W$  can be written as  $(w_k)_{k \geq 0}$ .

**A.1. Proof of Lemma 2.1**

It is well known from the theory of branching processes (see Chap. XII in Ref. 6 and Theorem 4 in Ref. 23) that for all  $n, i \geq 0$ , the  $i$ th row in matrix  $Q^n$  has generating function  $[\varphi^{(n)}(x)]^i$ , where  $\varphi^{(n)}(x)$  is  $n$ -fold composition of  $\varphi$  with itself. We have to find a generating function of the row indexed by 1 in  $(Q + p_C N)^n$ . Here, we show that this generating function can be expressed in a simple way by composition of the polynomial  $\varphi^{(n)}(x)$ . However, we first need the theorem about multiplication of matrix  $Q + p_C N$ .

**Theorem A.1.** (Matrix multiplication theorem) *For every  $n \in \mathbb{N}$  we have*

$$(Q + p_C N)^n = Q^n + p_C \gamma^{n-1} \sum_{i=0}^{n-1} R_i / \gamma^i, \tag{A.2}$$

where  $\gamma = \varphi'(1) + p_C = 1 + p_D - p_L$  and  $R_i = NQ^i$ .

**Proof.** Before we start to prove the theorem, we need the following two lemmas.

**Lemma A.1.** *Let  $W = (w_i)_{i \geq 0}$  be a matrix, where  $w_i$ 's are generating functions. We have*

$$WN = (w'_i(1) \cdot x)_{i \geq 0}, \tag{A.3}$$

$$NW = (i \cdot w_1)_{i \geq 0}. \tag{A.4}$$

**Proof.** Let  $WN = (v_i)_{i \geq 0} = (v_{i,j})_{i,j \geq 0}$ , where  $v_i$ 's are generating functions. Then  $v_{i,j} = 0$  for all  $i \geq 0$  and  $j \neq 1$ . For  $i \geq 0$  and  $j = 1$  we have  $v_{i,1} = \sum_{k=0}^{\infty} w_{i,k} k = w'_i(1)$ . Hence  $v_i = w'_i(1)x$ .

Now, let  $NW = (v_i)_{i \geq 0} = (v_{i,j})_{i,j \geq 0}$ . Then  $v_{i,j} = iw_{1,j}$ . Hence  $v_i = iw_1$ . □

By (A.4) we have  $R_k = (i\varphi^{(k)})_{i \geq 0}$ . From Lemma A.1 we obtain the following corollary for multiplications of matrices  $Q$  and  $N$ .

**Corollary A.1.** *Let  $k \geq 0$ . Then we have*

$$Q^k N = \mu^k N, \tag{A.5}$$

$$R_k Q = R_{k+1}, \tag{A.6}$$

$$R_k N = \mu^k N, \tag{A.7}$$

where  $\mu = \varphi'(1) = p_U + 2p_D$ .

**Proof.** Formula (A.5) follows from applying  $k$  times the formula (A.3) and the fact that  $\varphi'(1) = \mu$ . Formula (A.6) follows from definition of  $R_k$ . The last formula (A.7) follows from definition of  $R_k$ , formula (A.5) and the fact  $N^2 = N$  (from (A.3) or (A.4)). □

**Lemma A.2.** *For every  $x, y \in \mathbb{R}$  and  $n \in \mathbb{N}$  we have<sup>a</sup>*

$$(x + y)^n = x^n + y \sum_{i=0}^{n-1} (x + y)^{n-1-i} x^i . \tag{A.8}$$

An easy proof by induction on  $n$  is left for the reader.

Now we are ready to prove (A.2) by induction on  $n$ . Obviously, the formula (A.2) is satisfied for  $n = 0$ . Now, we assume that this formula is satisfied for some  $n$  and we prove it for  $(n + 1)$ . We have the following equalities:

$$\begin{aligned} (Q + p_C N)^{n+1} &= (Q + p_C N)^n (Q + p_C N) \\ &\stackrel{\text{ind.}}{=} \left( Q^n + p_C \gamma^{n-1} \sum_{i=0}^{n-1} R_i / \gamma^i \right) (Q + p_C N) \\ &= Q^{n+1} + p_C \gamma^{n-1} \sum_{i=0}^{n-1} \underbrace{R_i Q / \gamma^i}_{(A.6)} + p_C \underbrace{Q^n N}_{(A.5)} + p_C^2 \gamma^{n-1} \sum_{i=0}^{n-1} \underbrace{R_i N / \gamma^i}_{(A.7)} \\ &= Q^{n+1} + p_C \gamma^{n-1} \sum_{i=0}^{n-1} R_{i+1} / \gamma^i + p_C \mu^n N + p_C^2 \gamma^{n-1} \sum_{i=0}^{n-1} \mu^i N / \gamma^i \\ &= Q^{n+1} + p_C \gamma^{n-1} \sum_{i=1}^n R_i / \gamma^{i-1} + p_C \underbrace{\left( \mu^n + p_C \sum_{i=0}^{n-1} (p_C + \mu)^{n-1-i} \mu^i \right)}_{(A.8)} N \\ &= Q^{n+1} + p_C \gamma^n \sum_{i=1}^n R_i / \gamma^i + p_C (p_C + \mu)^n N \\ &= Q^{n+1} + p_C \gamma^n \sum_{i=0}^n R_i / \gamma^i . \end{aligned}$$

Now, it is obvious from Theorem A.1 that the generating function  $\delta_{n,p_L,p_D,p_C}$  for the row indexed by 1 of the expanded matrix  $(Q + p_C N)^n$  can be expressed by the functions  $\varphi^{(i)}(x)$  ( $i = 0, 1, \dots, n$ ):

$$\delta_{n,p_L,p_D,p_C}(x) = \varphi^{(n)}(x) + p_C \gamma^{n-1} \sum_{i=0}^{n-1} \varphi^{(i)}(x) / \gamma^i . \tag{A.9}$$

Thus we have that the generating function  $\phi_{n,p,a,b}$  for the row indexed by 1 of the unexpanded matrix  $(Q + p_C N)^n$  is

$$\phi_{n,p,a,b}(x) = \delta_{n,p_L,p_D,p_C}(x) - \delta_{n,p_L,p_D,p_C}(0) . \tag{A.10}$$

This completes the proof of Lemma 2.1.

<sup>a</sup>We adopt the convention that  $0^0 = 1$ .

**A.2. Proof of Lemma 2.2**

We proved in Ref. 23 (see Lemmas 1 and 2) that the limit  $\lim_{n \rightarrow \infty} \frac{1 - \varphi^{(n)}(x)}{\mu^n}$  exists and it is finite for every  $x \in (-\frac{1-ap}{ap}, \frac{1+b}{a})$ . Moreover, if  $x$  is non-negative, then this limit equals zero iff  $x = 1$ . Thus there exists a finite function  $M_{p,a,b}$  such that for every  $n \in \mathbb{N}$  we have  $|\frac{1 - \varphi^{(n)}(x)}{\mu^n}| \leq M_{p,a,b}(x)$ .

It follows from Lemma 2.1 that

$$\frac{\phi_{n,p,a,b}(1) - \phi_{n,p,a,b}(x)}{\gamma^n} = \frac{1 - \varphi^{(n)}(x)}{\gamma^n} + \frac{bp}{\gamma} \sum_{i=0}^{n-1} \frac{1 - \varphi^{(i)}(x)}{\gamma^i}. \tag{A.11}$$

Let  $\xi = \frac{\mu}{\gamma}$ . It is clear that  $\xi < 1$ . Hence

$$\left| \frac{1 - \varphi^{(n)}(x)}{\gamma^n} \right| = \left| \frac{1 - \varphi^{(n)}(x)}{\mu^n} \right| \left| \frac{\mu}{\gamma} \right|^n \leq M_{p,a,b}(x) \xi^n \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

and

$$\left| \sum_{i=0}^{\infty} \frac{1 - \varphi^{(i)}(x)}{\gamma^i} \right| \leq \sum_{i=0}^{\infty} \left| \frac{1 - \varphi^{(i)}(x)}{\mu^i} \right| \left| \frac{\mu}{\gamma} \right|^i \leq M_{p,a,b}(x) \sum_{i=0}^{\infty} \xi^i < \infty.$$

Thus we have proved that  $|c_{p,a,b}(x)| < \infty$ . Moreover, it follows from (A.11) that if  $x$  is non-negative then  $c_{p,a,b}(x) = 0$  iff  $x = 1$ .

**A.3. Proof of condition (2.2)**

It follows from (A.9) that for all  $n \in \mathbb{N}$  we have:

$$\delta_{n+1,pL,pD,pC}(x) = \delta_{n,pL,pD,pC}(\varphi(x)) + p_C \gamma^n x. \tag{A.12}$$

It follows from (A.10) and (A.12) that:

$$\begin{aligned} f_{n,p,a,b}(x) &= \frac{\delta_{n-1,pL,pD,pC}(\varphi(x)) + p_C \gamma^{n-1} x - \delta_{n,pL,pD,pC}(0)}{\phi_{n,p,a,b}(1)} \\ &= 1 + \frac{\phi_{n-1,p,a,b}(\varphi(x)) + bp\gamma^{n-1} x}{\phi_{n,p,a,b}(1)} - \frac{\delta_{n,pL,pD,pC}(1) - \delta_{n-1,pL,pD,pC}(0)}{\phi_{n,p,a,b}(1)} \\ &= 1 + \frac{\phi_{n-1,p,a,b}(\varphi(x))}{\phi_{n,p,a,b}(1)} + \frac{bp\gamma^{n-1}}{\phi_{n,p,a,b}(1)} x \\ &\quad - \frac{\delta_{n-1,pL,pD,pC}(\varphi(1)) + bp\gamma^{n-1} - \delta_{n-1,pL,pD,pC}(0)}{\phi_{n,p,a,b}(1)} \\ &= 1 + \frac{\phi_{n-1,p,a,b}(\varphi(x))}{\phi_{n,p,a,b}(1)} + \frac{bp\gamma^{n-1}}{\phi_{n,p,a,b}(1)} x - \frac{\phi_{n-1,p,a,b}(1)}{\phi_{n,p,a,b}(1)} - \frac{bp\gamma^{n-1}}{\phi_{n,p,a,b}(1)}. \end{aligned}$$

Now, it follows from Lemma 2.2 that

$$\lim_{n \rightarrow \infty} \frac{\phi_{n-1,p,a,b}(\varphi(x))}{\phi_{n,p,a,b}(1)} = \lim_{n \rightarrow \infty} \frac{\phi_{n-1,p,a,b}(\varphi(x))}{\phi_{n-1,p,a,b}(1)} \cdot \frac{\phi_{n-1,p,a,b}(1)}{\phi_{n,p,a,b}(1)} = \frac{f_{p,a,b}(\varphi(x))}{\gamma}$$

and

$$\lim_{n \rightarrow \infty} \frac{bp\gamma^{n-1}}{\phi_{n,p,a,b}(1)} = \frac{bp}{\gamma} \lim_{n \rightarrow \infty} \frac{\gamma^n}{\phi_{n,p,a,b}(1)} = \frac{bp}{\gamma c_{p,a,b}(0)}$$

and

$$\lim_{n \rightarrow \infty} \frac{\phi_{n-1,p,a,b}(1)}{\phi_{n,p,a,b}(1)} = \frac{1}{\gamma} \lim_{n \rightarrow \infty} \frac{\phi_{n-1,p,a,b}(1)}{\gamma^{n-1}} \lim_{n \rightarrow \infty} \frac{\gamma^n}{\phi_{n,p,a,b}(1)} = \frac{1}{\gamma}.$$

Hence, we have

$$f_{p,a,b}(x) = 1 + \frac{f_{p,a,b}(\varphi(x))}{\gamma} + \frac{bp}{\gamma \cdot c_{p,a,b}(0)}x - \frac{1}{\gamma} - \frac{bp}{\gamma \cdot c_{p,a,b}(0)}.$$

This is equivalent to (2.2). The proof of Theorem 2.1 is fully completed.

#### A.4. Proof of Theorem 2.2

We are interested in the behavior of DLC model for small values of  $p$  and fixed values of  $a$  and  $b$ . We will use here an explicit dependency on  $p$ . Thus, let  $\varphi_p(x) := \varphi(x)$  and  $\gamma_p := \gamma$ .

**Lemma A.3.** *Let  $1 + b > a$ . Let  $x_*$  be a fixed real such that  $0 < x_* < 1$  and let  $n(p)$  be an integer such that  $\varphi_p^{n(p)-1}(0) < x_* \leq \varphi_p^{n(p)}(0)$ . Then there exists the limit*

$$\lim_{p \rightarrow 0^+} n(p)p = \Lambda(x_*), \tag{A.13}$$

where  $\Lambda: (0, 1) \rightarrow (0, \infty)$  is a function such that  $\Lambda(x) = \frac{1}{1+b-a} \ln \frac{1+b-ax}{(1+b)(1-x)}$ .

**Proof.** We have  $\varphi_p(x) = x + p\eta(x)$ , where  $\eta(x) = (1-x)(1+b-ax)$ . It is sufficient to assume that  $p$  is sufficiently small, i.e.  $p \ll x_*$ . Let us divide the segment  $[0, x_*]$  into the segments of length  $h$ , where  $p \ll h \ll x_*$ . Now, let  $x_j$  and  $x_{j+1}$  be any successive points of this partition. Then the number  $n_j$  of all  $i$ 's such that  $\varphi_p^{(i)}(0) \in [x_j, x_{j+1}]$  satisfies  $n_j \approx \frac{h}{p\eta(x_j)}$ . Thus we have  $\lim_{p \rightarrow 0^+} n(p)p = \int_0^{x_*} \frac{dx}{\eta(x)} = \Lambda(x_*)$ . It is clear that the function  $\Lambda(x) = \int_0^x \frac{ds}{(1-s)(1+b-as)} = \frac{1}{1+b-a} \ln \frac{1+b-ax}{(1+b)(1-x)}$  is finite for every  $x \in (0, 1)$ . □

**Lemma A.4.** *Let  $1 + b > a$ . Then there exists the limit*

$$\lim_{p \rightarrow 0^+} c_{p,a,b}(0) = c_{a,b}, \tag{A.14}$$

where  $c_{a,b} = b \int_0^1 \frac{x^{\frac{a-1}{1+b-a}}}{1+b-ax} dx$  is positive (whenever  $b > 0$ ) and finite.

**Proof.** In order to prove that the limit A.4 exists, we will prove the existence of the limit  $\lim_{p \rightarrow 0^+} p \cdot s_{p,a,b}$ , where  $s_{p,a,b} = \sum_{n=0}^{\infty} \frac{1-\varphi_p^{(n)}(0)}{\gamma_p^n}$  and  $c_{p,a,b}(0) = \frac{bp}{\gamma_p} s_{p,a,b}$  (see the proof of Lemma 2.2).

Let  $H: (0, \infty) \rightarrow (0, 1)$  be a function such that  $H(x) = 1 - \Lambda^{-1}(x) = \frac{(1+b-a)e^{-(1+b-a)x}}{1+b-ae^{-(1+b-a)x}}$ . It is easy to check that  $H$  is decreasing and continuous.

It follows from Eq. (A.13) that

$$1 - \varphi_p^{(n)}(0) \approx H(pn).$$

We also have the fact

$$\gamma_p^n = (\varphi'_p(1) + bp)^n = (1 - (1 - a)p)^n \approx e^{-(1-a)pn}. \tag{A.15}$$

Thus we have

$$\lim_{p \rightarrow 0^+} p \cdot s_{\beta,a,b,p} = \lim_{p \rightarrow 0^+} p \sum_{n=0}^{\infty} \frac{1 - \varphi_p^{(n)}(0)}{\gamma_p^n} = \lim_{p \rightarrow 0^+} p \sum_{n=0}^{\infty} \frac{H(pn)}{e^{-(1-a)pn}}.$$

The function  $\psi(x) = \frac{H(x)}{e^{-(1-a)x}} = \frac{(1+b-a)e^{-bx}}{1+b-ae^{-(1+b-a)x}}$  is continuous. Thus we obtain

$$\lim_{p \rightarrow 0^+} p \cdot s_{p,a,b} = \lim_{p \rightarrow 0^+} p \sum_{n=0}^{\infty} \psi(pn) = \int_0^{\infty} \psi(x)dx.$$

The integral is finite and positive.

Hence, we obtain an existence of the limit  $c_{a,b} = \lim_{p \rightarrow 0^+} c_{p,a,b}(0) = \lim_{p \rightarrow 0^+} \frac{bp}{\gamma_p} s_{p,a,b} = b \int_0^{\infty} \psi(x)dx = b \int_0^1 \frac{z^{\frac{a-1}{1+b-a}}}{1+b-az} dz$ , which is positive iff  $b > 0$ . It is also easy to check that  $c_{a,b}$  is finite. □

**Lemma A.5.** *Let  $1 + b > a$  and let  $\chi(x)$  be a solution of the following differential equation:*

$$(1 - a)\chi_{a,b}(x) + C_{a,b}(1 - x) + \chi'_{a,b}(x)\eta(x) = 0, \tag{A.16}$$

where  $C_{a,b} = \left( \int_0^1 \frac{x^{\frac{a-1}{1+b-a}}}{1+b-ax} dx \right)^{-1}$  and  $\eta(x) = (1 - x)(1 + b - ax)$ . Then function  $h_p(x) = 1 - f_{p,a,b}(x) - \chi_{a,b}(x)$  satisfies the equation

$$h_p(\varphi(x)) = \gamma_p \cdot h_p(x) + o(p). \tag{A.17}$$

**Proof.** Let's recall Eq. (2.2), marking the explicit dependence on  $p$ ,

$$f_{p,a,b}(\varphi_p(x)) = \gamma_p f_{p,a,b}(x) + (1 - \gamma_p) + \frac{bp}{c_{p,a,b}(0)}(1 - x).$$

We have  $h_p(x) = 1 - f_{p,a,b}(x) - \chi_{a,b}(x)$ . Thus, it follows that

$$h_p(\varphi_p(x)) = \gamma_p h_p(x) + \gamma_p \chi_{a,b}(x) - \chi_{a,b}(\varphi_p(x)) - \frac{bp}{c_{p,a,b}(0)}(1 - x). \tag{A.18}$$

Let  $\omega_p(x) = \gamma_p \chi_{a,b}(x) - \chi_{a,b}(\varphi_p(x)) - \frac{bp}{c_{p,a,b}(0)}(1 - x)$ . Thus we have to prove that  $\omega_p(x) = o(p)$ .

It follows from  $\varphi_p(x) = x + p\eta(x)$  and Taylor's Theorem that  $\chi_{a,b}(\varphi_p(x)) = \chi_{a,b}(x) + \chi'_{a,b}(x)p\eta(x) + o(p)$ . Thus we obtain

$$\omega_p(x) = -(1 - \gamma_p)\chi_{a,b}(x) - \chi'_{a,b}(x)p\eta(x) - \frac{b}{c_{p,a,b}(0)}p(1 - x) - o(p).$$

We have  $1 - \gamma_p = (1 - a)p$  and  $C_{a,b} = \lim_{p \rightarrow 0^+} b/c_{p,a,b}(0) = b/c_{a,b}$  (see Lemma A.4). Now, notice that we have  $\chi_{a,b}$  satisfies (A.16). Then it follows that  $\lim_{p \rightarrow 0^+} \frac{\omega_p(x)}{p} = 0$ , so  $\omega_p(x) = o(p)$ . □

**Lemma A.6.** *Let  $1 + b > a$  and let  $h_p$  be a function defined on the interval  $[0, 1]$  such that  $h_p(0) = 0$  and (A.17) is satisfied. Then for every  $0 < x < 1$  we have*

$$\lim_{p \rightarrow 0^+} h_p(x) = 0. \tag{A.19}$$

In the following proof we will use the Iverson's notation  $[\mathcal{P}]$ , where  $\mathcal{P}$  is any property. If a property  $\mathcal{P}$  holds, then  $[\mathcal{P}]$  equals 1, otherwise it is 0.

**Proof.** Similarly to Lemma A.3, let  $0 < x < 1$  and let  $n(p)$  be an integer such that

$$\varphi_p^{(n(p)-1)}(0) < x \leq \varphi_p^{(n(p))}(0).$$

Now, it follows from (A.17), (A.15) and Lemma A.3 that

$$\begin{aligned} \lim_{p \rightarrow 0^+} h_p(x) &= \lim_{p \rightarrow 0^+} h_p(\varphi_p^{(n(p))}(0)) \\ &= \lim_{p \rightarrow 0^+} \left( \gamma_p^{n(p)} h_p(0) + o(p) \sum_{i=0}^{n(p)-1} \gamma_p^i \right) \\ &= \lim_{p \rightarrow 0^+} \gamma_p^{n(p)} h_p(0) + \lim_{p \rightarrow 0^+} o(p) \left( \frac{1 - \gamma_p^{n(p)}}{1 - \gamma_p} [\gamma_p \neq 1] + n(p) [\gamma_p = 1] \right) \\ &= 0 + \lim_{p \rightarrow 0^+} o(p) \left( \frac{1 - e^{-(1-a)pn(p)}}{(1-a)p} [\gamma_p \neq 1] + \frac{pn(p)}{p} [\gamma_p = 1] \right) \\ &= \lim_{p \rightarrow 0^+} \frac{o(p)}{p} \left( \frac{1 - e^{-(1-a)pn(p)}}{1-a} [\gamma_p \neq 1] + pn(p) [\gamma_p = 1] \right) \\ &= \lim_{p \rightarrow 0^+} o(1) = 0. \end{aligned} \tag{□}$$

Without loss of generality we may assume that function  $\chi_{a,b}$  additionally satisfies the condition  $\chi_{a,b}(0) = 1$ . Thus an existence of the limit function  $f_{a,b}$  follows from Lemma A.6 and the differential equation from Lemma A.5. Therefore the proof of Theorem 2.2 is completed.

**A.5. Proof of equation (2.10)**

For  $i = 2$ , Eq. (2.10) is just (2.7). Thus we have to prove it only for  $i \geq 3$ . Let  $q_{a,b,i} = (i + 1)p_{a,b,i+1} - \frac{a}{1+b}ip_{a,b,i}$  for  $i \geq 2$ . We have for  $i \geq 2$ :

$$q_{a,b,i+1} = q_{a,b,i} - \frac{1-a}{1+b}p_{a,b,i}. \tag{A.20}$$

The above equality follows from (2.8). Thus we obtain for  $i \geq 2$ :

$$\begin{aligned} q_{a,b,i+1} &= q_{a,b,i-1} - \frac{1-a}{1+b}(p_{a,b,i-1} + p_{a,b,i}) = \dots \\ &= q_{a,b,2} - \frac{1-a}{1+b}(p_{a,b,2} + \dots + p_{a,b,i}) \\ &= 2p_{a,b,2} - \frac{a}{1+b}p_{a,b,1} - \frac{1-a}{1+b}(p_{a,b,2} + \dots + p_{a,b,i}) \\ &= p_{a,b,1} - \frac{C_{a,b}}{1+b} - \frac{1-a}{1+b} \sum_{j=1}^i p_{a,b,j} \\ &= -\frac{1-a}{1+b} \left( 1 - \sum_{j=1}^i p_{a,b,j} \right). \end{aligned}$$

This completes the proof of (2.10).

**5. An Alternative Proof for the Case  $a = 1$**

One of the reviewers suggested to us a simple and elegant proof of Corollary 2.2 and Theorem 2.2 when  $a = 1$ . Here, we sketch an outline of the suggested reasoning.

For  $a = 1$ , we have  $\varphi(x) = x + p\eta(x)$ , where  $\eta(x) = 1 + b - (2 + b)x + x^2 = (1 - x)(1 + b - x)$ . Now, from Taylor’s Theorem, we get

$$f_{p,1,b}(\varphi(x)) = f_{p,1,b}(x) + p\eta(x)f'_{p,1,b}(x) + \frac{1}{2}(p\eta(x))^2 f''_{p,1,b}(\xi_{x,p}),$$

where  $\xi_{x,p}$  lies between  $x$  and  $\varphi(x)$ . Then Eq. (2.2) becomes

$$p\eta(x)f'_{p,1,b}(x) + \frac{1}{2}p^2\eta^2(x)f''_{p,1,b}(\xi_{x,p}) = \frac{bp}{c_{p,1,b}(0)}(1 - x).$$

Divide both sides by  $p$  and recall that the map  $p \mapsto c_{p,1,b}(0)$  is bounded and away from 0 on  $(0, \delta)$  for some  $\delta > 0$ . Then, any sequence  $(p_n)_{n \geq 1}$ , with  $p_n \in (0, \delta)$  for all  $n$  and  $p_n \rightarrow 0^+$ , contains a subsequence  $(p_{n'})_{n'}$  such that  $c_{p_{n'},1,b}(0) \rightarrow C'_{1,b}$  ( $n' \rightarrow \infty$ ). Hence, taking the limit along such a subsequence, one gets

$$\lim_{n' \rightarrow \infty} f'_{p_{n'},1,b}(x) = \frac{b}{C'_{1,b}} \frac{1-x}{\eta(x)} = \frac{b}{C'_{1,b}} \frac{1}{1+b-x}.$$

Therefore, by the dominated convergence theorem,

$$\lim_{n' \rightarrow \infty} \int_0^x f'_{p_{n'},1,b}(y)dy = \frac{b}{C'_{1,b}} \int_0^x \frac{1}{1+b-y}dy = \frac{b}{C'_{1,b}} \log \frac{1+b}{1+b-x}$$

which entails

$$\lim_{n' \rightarrow \infty} f_{p_{n'}, 1, b}(x) = \frac{b}{C'_{1, b}} \log \frac{1+b}{1+b-x}.$$

Let  $f_{0,1,b}(x) = \lim_{n' \rightarrow \infty} f_{p_{n'}, 1, b}(x)$ . Since  $b > 0$ ,  $x = 1$  is interior to the disc of convergence and, consequently,  $f_{0,1,b}(1) = 1$ , i.e.

$$C'_{1, b} = b \log \frac{1+b}{b}.$$

Let  $\theta = \frac{1}{1+b}$ . The limiting probability generating function can be written as

$$f_{0,1,b}(x) = \frac{1}{-\log(1-\theta)} \log \frac{1}{1-\theta x} = \frac{1}{-\log(1-\theta)} \sum_{i \geq 1} \frac{(\theta x)^i}{i}$$

which gives the limiting probability distribution

$$p_{1,b,i} = \frac{1}{-\log(1-\theta)} \frac{\theta^i}{i}, \quad i = 1, 2, \dots$$

Since this is independent of the subsequence  $(p_{n'})_{n'}$  and of the sequence  $(p_n)_n$ , the desired result follows from an obvious *reductio ad absurdum* argument.

## Acknowledgments

We wish to thank Piotr Slonimski for drawing our attention to the models of genome evolution and to the anonymous reviewers whose suggestions helped us to improve the paper. This work was partially supported by research grants from Polish Ministry of Science and Higher Education: 3 T11F 021 28, 3 T11F 016 28, and 2 P03A 031 25, and by the EC FP6 Marie Curie ToK programme SPADE2, MTKD-CT-2004-014508, hosted at IMPAN, accompanied by the Polish MNiI SPB-M.

## References

1. B. O. Bengtsson, Modelling the evolution of genomes with integrated external and internal functions, *J. Theor. Biol.* **231** (2004) 271–278.
2. R. N. Curnow, The use of Markov chain models in studying the evolution of the proteins, *J. Theor. Biol.* **134** (1988) 51–57.
3. M. Di Giulio and F. Caldararo, Absorbent Markov chains as a model for the study of the evolution of proteins, *J. Theor. Biol.* **124** (1987) 485–494.
4. B. Dujon *et al.*, Genome evolution in yeasts, *Nature* **430** (2004) 35–44.
5. A. J. Enright, S. Van Dongen and C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families, *Nucl. Acids Res.* **30** (2002) 1575–84.
6. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1 (John Wiley & Sons, 1961).
7. W. M. Fitch, Homology, a personal view on some of the problems, *Trends in Genetics* **16** (2000) 227–321.
8. R. A. Gatenby, T. L. Vincent and R. J. Gillies, Evolutionary dynamics in carcinogenesis, *Math. Mod. Meth. Appl. Sci.* **15** (2005) 1619–1638.

9. L. Hsieh, L. Luo, F. Ji and H. Lee, Minimal model for genome evolution and growth, *Phys. Rev. Lett.* **90** (2003) 101–104.
10. M. A. Huynen and E. van Nimwegen, The frequency distribution of gene family size in complete genomes, *Mol. Biol. Evolution* **15** (1998) 583–589.
11. K. Jordan, K. S. Makarova, J. L. Spouge, Y. I. Wolf and E. V. Koonin, Lineage-specific gene expansions in bacterial and archeal genomes, *Genome Res.* **11** (2001) 555–565.
12. G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya and E. V. Koonin, Birth and death of protein domains: A simple model of evolution explains power law behavior, *BMC Evolutionary Biol.* **2** (2002) 18.
13. G. P. Karev, Y. I. Wolf and E. V. Koonin, Simple stochastic birth and death models of genome evolution: Was there enough time for us to evolve?, *Bioinform.* **19** (2003) 1889–1900.
14. G. P. Karev, Y. I. Wolf, F. S. Berezovskaya and E. V. Koonin, Modeling genome evolution with a diffusion approximation of a birth-and-death process, *Bioinform.* **21** (2005) iii12–iii19.
15. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
16. R. Mikelsaar, Human mitochondrial genome and the evolution of methionine transfer ribonucleic acids, *J. Theor. Biol.* **105** (1983) 221–232.
17. T. Miura and P. Sonigo, A mathematical model for experimental gene evolution, *J. Theor. Biol.* **209** (2001) 497–502.
18. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, 1970).
19. J. D. Peterson, L. A. Umayam, T. M. Dickinson, E. K. Hickey and O. White The comprehensive microbial resource, *Nucl. Acids Res.* **29** (2001) 123–125.
20. W. J. Reed and B. D. Hughes, A model explaining the size distribution of gene and protein families, *Math. Biosci.* **189** (2004) 97–102.
21. P. P. Slonimski, M. O. Mosse, P. Golik, A. Henaût, Y. Diaz, J. L. Risler, J. P. Comet, J. C. Aude, A. Wozniak, E. Glemet and J. J. Codani, The first laws of genomics, *Microbial Comp. Genomics* **3** (1998) 46.
22. P. P. Slonimski, Comparison of complete genomes: Organization and evolution, *Proc. of the Third Annual Conference on Computational Molecular Biology*, RECOMB'99 Stanislaw Ulam Memorial Lecture (ACM Press, 1999), p. 310.
23. J. Tiuryn, R. Rudnicki and D. Wójtowicz, A case study of genome evolution: From continuous to discrete time model, in *Proc. of Mathematical Foundations of Computer Science 2004*, eds. J. Fiala, V. Koubek and J. Kratochvíl, Lecture Notes in Computer Science, Vol. 3153 (Springer, 2004), pp. 1–24.
24. Y. Wang, W. Li, T. Zhang, C. Ding, Z. Lu, N. Long, J. P. Rose, B. C. Wang and D. Lin, Reconstruction of ancient genome and gene order from complete microbial genome sequences, *J. Theor. Biol.* **239** (2006) 494–8.
25. I. Yanai, C. J. Camacho and C. DeLisi, Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification, *Phys. Rev. Lett.* **85** (2000) 2641–2644.