ON A CONTINUOUS MODEL OF THE EVOLUTION OF PARALOG FAMILIES IN GENOMES

RYSZARD RUDNICKI, JERZY TIURYN, AND DAMIAN WÓJTOWICZ

Abstract. We introduce and analyse a simple probabilistic model of genome evolution. It is based on three fundamental evolutionary events: gene loss, duplication and accumulated change. 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 system of infinite number ‘of linear equations. We show that this system generates a semigroup of linear operators on the space l^1. We prove that size distribution of paralogous gene families in a genome converges to the equilibrium as time goes to infinity. Moreover we show that when probabilities of gene removal and duplication are close to each other, then the resulting distribution is close to logarithmic distribution. Some empirical results for yeast genomes are presented.

1. Introduction

Gene and genome duplication is a fundamental feature of evolution. Since the seminal work of Ohno [12] it is considered as one of the main mechanisms of sequence divergence, functional innovation, speciation and constitutes a substratum for the Darwinian selection for adaptive fitness. These duplication processes lead to the appearence of paralogous genes. Two genes present in the same genome are said to be *paralogs* if they are homologous and have evolved through a duplication from a single ancestor gene. It should be born in mind that the term homology as defined above is an abstraction in that it is a relationship which can only be inferred with more or less certainty (see Fitch [4] for an in depth discussion). Operationally, in the study of molecular evolution, one infers homology not by the study of descent, which is in most

Key words and phrases. genome evolution, paralogous genes, Markov semigroups, asymptotic stability.

This research was partially supported by the State Committee for Scientific Research (Poland) Grants: 2 P03A 031 25, 3 T11F 021 28, and 3 TF11 016 28.

Praca wstepnie zaakceptowana do J. Math. Biology.

cases not accessible, but by the comparison of extant genes/proteins. If the degree of similarity/identity is so great (using a treshhold or cut-off value which is reasonable albeit arbitrary) as compared to a random explanation of similarity, then one infers homology.

Obviously a genome is not simply a set of genes, but rather a dynamic collection of genes which changes in time. Various molecular events (e.g. gene duplication and loss, point mutation, recombination, gene conversion, rearrangement, DNA repair, translocation, horizontal transfer) constantly act on genomes and drive them to evolve dynamically. In this paper we propose to study a simple model of genome evolution in the spirit of Kimura [9], i.e. in the total absence of selective pressures. We are aware that such a purely neutralistic model cannot be truly realistic. We belive, however, that it can be useful as a basis for further discussions. The model addresses three evolutionary events: gene duplication, gene loss or removal and gene change. Although it may seem trivial we prefer to define these notions: (i) by gene duplication we understand an event in which one gene gives rise to two genes which cannot be operationally distinguished between themeselves, which remain in the same genome and are therefore paralogs; (ii) by gene loss we understand an event which leads to a removal of the gene from the genome; (iii) by gene change we understand an event (or a cumulative series of events like mutations, rearrangements, recombinations, ...) which lead 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 a genome.

A motivation for the present work comes from the study of the size distribution of paralog families in several microbial genomes which was undetaken in late 90's. In 1998 P. Slonimski et al. [14, 15] and independently M.A. Huynen and E. van Nimwegen [5] have counted the numbers of i-element clusters of paralogous genes (for $i = 2, 3, \ldots$, etc.) in several genomes which have been sequenced till then and came with two different claims concerning the shape of the observed distribution: [14] states that the distribution is *logarithmic* (i.e. a probability of being an i-element cluster is $C \cdot \theta^i/i$, where $0 < \theta < 1$ and C is a normalizing constant), while [5] claims that this distribution is *power law* (i.e. the probability is $D \cdot i^{-\gamma}$, where $\gamma > 1$ and D is a suitable normalizing constant). In 2001 Jordan et al. [6] have analysed 21 completely sequenced bacterial genomes and concluded that the logarithmic approximation fits the distribution slightly better than the power law approximation. The above cited papers did not propose any probabilistic model which explained the observed distributions. In 2000 Yanai et al. [17] designed a simple model of genome evolution 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 paralog distribution of the genome. Mathematical analysis of the model was not given in that paper.

To our knowledge the first paper which proposed a model of genome evolution together with complete mathematical analysis of the equilibrium frequences of domain families is Karev et al. [7] published in 2002 (see also [8]). 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 serves the purpose of stabilizing the asymptotic behaviour of the model. Karev et al. 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 two important respects. (1) 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. (2) 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.

Two models in the spirit of the present paper but without the mechanism of accumulated change was analysed in Tiuryn at al. [16]. In that paper we proposed both a discrete and a comtinous time model for this phenomena. It is shown there that the asymptotic distribution in this 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).

The organization of our paper is the following. First we present our model. The size distribution of paralogous gene families in a genome is described by a system of infinite number of linear equations. We show that this system generates a continuous semigroup of linear operators on the space l^1. Then we check that after suitable substitution we obtain a semigroup of Markov operators on that space. Using "the lower-bound function" theorem by Lasota and Yorke [10] we prove that this semigroup is asymptotically stable. It implies that the size distribution of paralogous gene families in a genome converges to the equilibrium as time goes to infinity. Moreover we show that when probabilities of gene removal and duplication are close to each other, then the resulting distribution is close to logarithmic distribution. The paper contains also a presentation of some experimental results for five yeasts.

2. 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 all colored genes. A *gene family* in a genome is a 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 continuous time. States of the Markov chain at time t are infinite sequences $(s_i(t))_{i\geq 1}$ of non-negative integers. A state $(s_i(t))_{i\geq 1}$ represents a genome in which for every $i \geq 1$, the number of i-element gene families is $s_i(t)$.

The model is parameterized by three positive reals: d, r and m. A transition from a genome $\mathcal{G}$ at time t to $\mathcal{G}'$ at time $t + \Delta t$ is based on the following process of *evolution* which is performed independently for each gene of $\mathcal{G}$. A gene, which is subject to the process of evolution during time interval of length Δt is:

- *duplicated* with probability $d \cdot \Delta t + o(\Delta t)$. 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 $\mathcal{C}_{i+1}$,
- *removed* from the genome with probability $r \cdot \Delta t + o(\Delta t)$. 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.
- *changed* with probability $m \cdot \Delta t + o(\Delta t)$. 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.

It is assumed that $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$. Moreover, we assume that all elementary events are independent of each other.

Let $s_i(t)$ be the number of i-element families in our model at the time t and let $\Delta s_i = s_i(t+\Delta t) - s_i(t)$. It follows from the description of our model that

$$\Delta s_1 = -(d+r)s_1(t)\Delta t + 2(m+r)s_2(t)\Delta t + \sum_{k=2}^{\infty} mks_k(t)\Delta t + o(\Delta t)$$

and

$$\begin{aligned}\Delta s_i =& - (d+r+m)is_i(t)\Delta t + d(i-1)s_{i-1}(t)\Delta t\\ &+ (r+m)(i+1)s_{i+1}(t)\Delta t + o(\Delta t)\end{aligned}$$

for $i \geq 2$. From these equations it follows that

$$s_1'(t) = -(d+r)s_1(t) + 2(m+r)s_2(t) + m\sum_{k=2}^{\infty} ks_k(t), \tag{1}$$

$$s_i'(t) = d(i-1)s_{i-1}(t) - (d+r+m)is_i(t) + (r+m)(i+1)s_{i+1}(t) \tag{2}$$

for $i \geq 2$. Let $s(t) = \sum_{i=1}^{\infty} s_i(t)$ be the total number of families. Then the sequence $(p_i(t))$, where $p_i(t) = s_i(t)/s(t)$ is the size distribution of paralogous gene families in a genome at time t.

3. Markov semigroups approach

In this section we prove a theorem on the existence and uniqueness of solutions of the system (1) and (2).

Assume that $r \geq 0$, $m \geq 0$ and $d \geq 0$. Let $S(t) = \sum_{i=1}^{\infty} is_i(t)$ be the total number of genes in the genome. We show that for each non-negative sequence $(s_i(0))_{i\geq 1}$ such that $S(0) < \infty$ the system (1) and (2) has a unique solution such that $s_i(t) \geq 0$ for all $t > 0$ and all positive integers i. The idea is the following. First, we change variables to obtain a new system which will be easier to study. We check that this new system generates a continuous Markov semigroup on the space l^1. Consequently, both systems have unique solutions in properly chosen spaces.

Let

$$y_i(t) = e^{(r-d)t} is_i(t).$$

Then

$$y_1' = -2dy_1 + (2m+r)y_2 + \sum_{k=3}^{\infty} my_k, \tag{3}$$

$$y_i' = -(d+r+m+\tfrac{d-r}{i})iy_i + diy_{i-1} + (r+m)iy_{i+1} \tag{4}$$

for $i \geq 2$.

Let l^1 denote the space of absolutely summable sequences. We check that the system (3) and (4) generates a Markov semigroup on l^1. We recall some notions concerning Markov operators and Markov semigroups.

A linear mapping $P : l^1 \to l^1$ is called a *Markov* or *stochastic operator* if $P(D) \subset D$, where

$$D = \{x \in l^1 : x_i \geq 0 \text{ for all } i \geq 1 \text{ and } \sum_{i=1}^{\infty} x_i = 1\}.$$

A family $\{P(t)\}_{t\geq 0}$ of Markov operators which satisfies conditions:

(a) $P(0) = \mathrm{Id}$,

(b) $P(t+s) = P(t)P(s) \quad \text{for} \quad s,\, t \geq 0$,

(c) for each $x \in l^1$ the function $t \mapsto P(t)x$ is continuous with respect to the l^1 norm

is called a *Markov* or *stochastic semigroup.*

Let $q_{1,1} = -2d$, $q_{1,2} = 2m+r$, $q_{1,j} = m$ for $j \geq 3$, and $q_{i,i} = -i(d+r+m+\frac{d-r}{i})$, $q_{i,i-1} = di$, $q_{i,i+1} = (r+m)i$, for $i \geq 2$, and $q_{i,j} = 0$ in other cases. The system (3) and (4) can be written in the following way:

$$y_i'(t) = \sum_{j=1}^{\infty} q_{i,j} y_j(t), \text{ for } i \geq 1 \tag{5}$$

and in the abstract form:

$$y'(t) = Qy(t), \tag{6}$$

where $Q = (q_{i,j})_{i,j\geq 1}$. The matrix Q has the following properties:

(i) $q_{i,j} \geq 0$ for $i \neq j$,

(ii) $\sum_{i=1}^{\infty} q_{i,j} = 0$ for $j \geq 1$.

We need the following result

Theorem 1. *Let the matrix Q satisfies conditions (i) and (ii). Let $Q^* = (q^*_{i,j})_{i,j\ge 1}$, where $q^*_{i,j} = q_{j,i}$ for $i,j \ge 1$ and let θ be a positive constant. Then the operator Q generates a Markov semigroup on l^1 if and only if there is no non-zero solution of the equation $Q^*x = \theta x$, where $x \in l^\infty$.*

Recall that l^∞ is the space of bounded sequence. The proof of Theorem 1 can be found in Norris [11].

The direct proof that the operator Q corresponding to system (3) and (4) by means of Theorem 1 is not easy and we do it in a little different way.

Assume additionaly that $d + m > 0$. The case $d = m = 0$ is simple and we omit it here. Define matrices $A = (a_{i,j})_{i,j\ge 1}$ and $K = (k_{i,j})_{i,j\ge 1}$. Let $a_{i,i} = -(i-1)(d+r+m)$, $a_{i+1,i} = (i-1)d$, $a_{i-1,i} = (i-1)(r+m)$ for $i \ge 2$, and $a_{i,j} = 0$ in other cases, and let $k_{1,i} = \frac{m}{2d+m}$ and $k_{i+1,i} = \frac{2d}{2d+m}$ for $i \ge 1$, and $k_{i,j} = 0$ in other cases. Then

$$Q = A - (2d+m)I + (2d+m)K. \tag{7}$$

Since $k_{i,j} \ge 0$ for all i,j and $\sum_{i=1}^\infty k_{i,j} = k_{1,j} + k_{j+1,j} = 1$ for all j, the matrix K is a Markov operator on l^1.

We check that A is an infinitesimal generator of a Markov semigroup. First, we observe that the matrix A satisfies conditions (i) and (ii). Condition (i): $a_{i,j} \ge 0$ is obvious. Since $\sum_{i=1}^\infty a_{i,j} = a_{j-1,j} + a_{j,j} + a_{j+1,j} = 0$ we also have (ii). If $A^*x = \theta x$, then $x_1 = 0$ and $|x_{n+1}| \ge (1 + \frac{\theta}{n-1})|x_n|$ for $n \ge 2$, which imples that $\lim_{n\to\infty} |x_n| = \infty$, and, consequently, $x \notin l^\infty$. According to Theorem 1 the operator A is an infinitesimal generator of a Markov semigroup, which we denote by $\{S(t)\}_{t\ge 0}$.

¿From the Phillips perturbation theorem [2], equation (6) generates a Markov semigroup $\{P(t)\}_{t\ge 0}$ on l^1 given by

$$P(t)x = e^{-(2d+m)t} \sum_{n=0}^\infty (2d+m)^n S_n(t)x, \tag{8}$$

where $S_0(t) = S(t)$ and

$$S_{n+1}(t)x = \int_0^t S(t-s)KS_n(s)x\,ds, \quad n \ge 0.$$

Thus we have proved the following result.

Theorem 2. *For each $y_0 \in l^1$ equation (6) has a unique solution $y(t)$ such that $y(0) = y_0$. If we denote by $P(t)y_0$ the solution $y(t)$ of with the initial condition $y(0) = y_0$, then $\{P(t)\}_{t\geq 0}$ is a Markov semigroup on l^1.*

Now we can return to the orginal system (1) and (2).

Corollary 1. *For each sequence (s_i) such that $\sum_{i=1}^{\infty} i|s_i| < \infty$ system (1), (2) has a unique solution $s(t)$ such that $s_i(0) = s_i$ for $i = 1, 2, \dots$. We have $\sum_{i=1}^{\infty} is_i(t) = \sum_{i=1}^{\infty} e^{(d-r)t} is_i$ for all $t \geq 0$. Moreover, if $s_i \geq 0$ for all i then $s_i(t) \geq 0$ for all i and t.*

4. Asymptotic behaviour

Next result characterize asymptotic behaviour of the solutions of equation (6).

Theorem 3. *If $m > 0$ then there exists a sequence $y^* = (y_n^*)_{n\geq 1}$ such that $y^* \in D$ and for each $y \in D$ we have $\lim_{t\to\infty}(P(t)y)_n = y_n^*$ for each $n \geq 1$.*

The proof of Theorem 3 is based on the following Lasota-Yorke lower bound function theorem [10]:

Theorem 4. *Let $\{P(t)\}_{t\geq 0}$ be a Markov semigroup on $L^1(X, \Sigma, \mu)$. If there exists $0 \leq h \in L^1(X, \Sigma, \mu)$ such that $\lim_{t\to\infty} \|(P(t)f - h)^-\|_1 = 0$ for every $f \in D$, then there exists $f^* \in D$ with $Pf^* = f^*$ such that for every $f \in D$ $\lim_{t\to\infty} \|P(t)f - f^*\| = 0$.*

Remark 1. A continuous semigroup $\{P(t)\}_{t\geq 0}$ of linear operators on $L^1(X, \Sigma, \mu)$ is called a *Markov semigroup* if $P(t)(D) \subset D$ for all $t \geq 0$, where

$$D = \{f \in L^1(X, \Sigma, \mu) : f \geq 0, \int f(x)\mu(dx) = 1\}.$$

In our case $l^1 = L^1(\mathbb{N}, 2^{\mathbb{N}}, \mu)$ and $\mu(A)$ is the number of elements of the set A. We use the notation $f^-(x) = -f(x)$ if $f(x) < 0$ and $f^-(x) = 0$ if $f(x) \geq 0$.

Proof of Theorem 3. Let $y(0) \in D$ and $m > 0$. Since

$$y_1'(t) \geq -2ry_1(t) + m(1 - y_1(t))$$

we have

$$\liminf_{t\to\infty} y_1(t) \geq \frac{m}{2r+m}.$$

Let $h = (\frac{m}{2r+m}, 0, 0, \dots)$. Then $h \in l^1$, $h \geq 0$, $h \neq 0$ and

$$\lim_{t\to\infty} (P(t)y - h)^- = 0$$

for each $y \in D$. ¿From Theorem 4 it follows that there exists $y^* \in D$ such that $\lim_{t\to\infty} P(t)y = y^*$ for $y \in D$. □

Now we are looking for a stationary solution of the system (1) and (2). Then $d = r$ and $y_n^* = \frac{m}{r}(\frac{r}{r+m})^n$ is an element of D such that $P(t)y^* = y^*$.

Corollary 2. *If $d = r$ then for every non-negative and non-zero solution of the system (1) and (2) there exists a positive constant C such that* $\lim_{t\to\infty} x_n(t) = Cn^{-1}(\frac{r}{r+m})^n$.

5. 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 genomes of yeast species: *Saccharomyces cerevisiae*, *Candida glabrata*, *Klyveromyces lactis*, *Debaromyces hansenii*, and *Yarrowia lipolytica*, whose genomes (with the exception of *Saccharomyces cerevuisiae*) have been recently sequenced [1]. The paralogous families were taken from `http://cbi.labri.fr/Genolevures/raw/fam/family-20040327-byfamily.txt` which provides a suplementary material to [1]. As it was observed by many researchers [5, 7, 14] 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 [14] is that large families are subjected to Darwinian selection of adaptive functions, a feature which is not present in our model. For this reason some researchers [14, 15] restrict analysis of families to small classes (cluster size 2 through 6), while others

[5, 7] group families into bins, each containing a certain prespecified 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 reject all single–member families because their number is very uneven: there are many pseudogenes, inactive genes, *etc.* Thus, it seems that the number of detected singletons is too big in this data set. The observed data was fitted to the logarithmic distribution and the distribution parameter $\theta = \frac{1}{1+m/r}$ was choosen to minimize the value of Pearson's χ^2 test. For each genome, before we evaluated the χ^2 test we grouped the expected paralog family frequencies into bins, each containing at least 10 genes. For all analysed genomes, with the exception of *Y. lipolytica*, $P(\chi^2)$ 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 θ, the fraction m/r and the goodness-of-fit $P(\chi^2)$ are given in Table 1. It should be noticed that the maximal likelihood method gives us almost the same θ for each genome.

Yeast genome	**# of families**	θ	**m/r**	$\mathbf{P}(\chi^2)$
Candida glabrata	576	0.475	1.105	0.229
Debaryomyces hansenii	755	0.564	0.773	0.101
Kluyveromyces lactis	465	0.517	0.934	0.060
Saccharomyces cerevisiae	723	0.496	1.016	0.167
Yarrowia lipolytica	632	0.536	0.866	0.005

Table 1. Paralogous families in yeast genomes [1] and the parameter of the best-fit model.

It appears from studying Table 1 that constant θ for yeasts is around 0.5. This is consistent with Slonimski's First Law of Genomics, because the Slonimski's group claims that this value is 1/2 for all genomes [14, 15]. The fit of the model reveals also the existance of a simple relationship between probabilities of accumulated change and gene duplication/loss. On average, the probability of gene dupliaction/loss appears to be approximately equals the *per gene* probability of accumulated change (see the fraction m/r in Table 1).

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 was not essential. However, it seems that the power law distribution has better goodness of fit than the logarithmic distribution when we cosider bigger family sizes. Some explanations of this situation can be found in [14, 15].

The only organism for which the difference between the observed distribution of paralogous gene families and the predicted value is statistically significant, is *Y. lipolytica.* Its genome differs from the other four genomes in several respects [1], the most pronounced being that its size (20.5 Mb) is almost twice as big as any other genome (10-12 Mb). The only class which causes this discrepancy is the class of three element families, which is much smaller in the genome that the predicted value. This suggests that during the evolution of this genome some kind of an accident has happend which disrupted the distribution of paralogs. It has been also observed in [1] that this genome *"shows a strong tendency for map dispersion"*, and *"by contrast the other yeast species show significant constraints on genome size, possibly associated with their ability to duplicated blocks and tandem gene repeats in their genomes"*. Finding an explanation of this mystery sounds like a good research topic and we plan to work on this in the future.

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 mathematical analysis of the asymptotic distribution of gene families. Genome evolution is a very complicated stochastic process which involves many additional events than 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 analyse it, and yet the theory behind it is quite involved mathematically. It would be interesting to see how other evolutionary events (like gene invention proposed in [7]), when introduced into the model affect the asymptotic distribution. For example removing gene change from the model results in geometric distribution [16].

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 be considered too.

Although our theoretical results are not fully consistient with empirical data, the model has some advantages. For example, it is practically impossible to check experimentally the relationship between probabilities of accumulated change and gene duplication/loss. Our model allow to find this relationship (maybe not precisely) only studying the distribution of paralogous families.

References

1. B. Dujon *et al.*, *Genome evolution in yeasts*, Nature **430** (2004), 35-44.
2. N. Dunford and J.T. Schwartz, *Linear Operators, Part I*, Interscience Publ., New York, 1968.
3. A.J. Enright, S. Van Dongen, C.A. Ouzounis, *An efficient algorithm for large-scale detection of protein families*, Nucleic Acids Research **30(7)** (2002), 1575-1584.
4. W.M. Fitch, *Homology, a personal view on some of the problems,* Trends in Genetics **16**/5 (2000), 227–321,
5. M.A. Huynen, E. van Nimwegen, *The frequency distribution of gene family size in complete genomes*, Molecular Biology Evolution **15**/5 (1998), 583–589.
6. K. Jordan, K.S. Makarova, J.L. Spouge,Y.I. Wolf, Y.I., E.V. Koonin, *Lineage-specific gene expansions in bacterial and archeal genomes,* Genome Research **11** (2001), 555–565.
7. G.P. Karev, Y.I. Wolf, A.Y. Rzhetsky, F.S. Berezovskaya, E.V. Koonin, *Birth and death of protein domains: A simple model of evolution explains power law behavior,* BMC Evolutionary Biology **2**/18 (2002).
8. G.P. Karev, Y.I. Wolf, E.V. Koonin, *Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?,* Bioinformatics **19**:15 (2003), 1889–1900.
9. M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, 1983.

10. A. Lasota, J.A. Yorke, *Exact dynamical systems and the Frobenius-Perron operator,* Trans. AMS **2**73(1982), 375–384.
11. J. Norris, *Markov Chains*, Cambridge Series on Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 1997.
12. S. Ohno, *Evolution by Gene Duplication*, Springer Verlag, Berlin, 1970.
13. R. Rudnicki, *On asymptotic stability and sweeping for Markov operators*, Bull. Pol. Ac.: Math. **43** (1995), 245–262.
14. 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, J.J. Codani, *The first laws of genomics,* Microbial and Comparative Genomics **3:46** (1998).
15. P.P. Slonimski, *Comparision of complete genomes: Organization and evolution,* Proceedings of the Third Annual Conference on Computational Molecular Biology, RECOMB'99, Stanislaw Ulam Memorial Lecture, pp. 310, ACM Press, 1999.
16. J. Tiuryn, R. Rudnicki, D. Wójtowicz, *A case study of genome evolution: from continuous to discrete time model,* in: Proceedings of Mathematical Foundations of Computer Science 2004 (J. Fiala, V. Koubek, and J. Kratochvíl eds.), LNCS 3153, 1–24, Springer, 2004.
17. I. Yanai, C.J. Camacho, C. DeLisi, *Predictions of Gene Family Distributions in Microbial Genomes: Evolution by Gene Duplication and Modification,* Physical Review Letters, 85(12) (2000), 2641–2644.

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

Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warszawa, Poland
E-mail address: tiuryn@mimuw.edu.pl

Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warszawa, Poland
E-mail address: dami@mimuw.edu.pl