

Zipf Law and Distribution of Paralog Families

Lev B. Levitin¹
levitin@bu.edu

Dmitriy Leyfer²
dmitriyl@bu.edu

^{1,2} Boston University, Bioinformatics Graduate Program, Boston, MA 02215, USA

Keywords: Zipf law, power law, Pareto distribution, paralog families, gene families, COGs

1 Introduction

The distribution of the sizes of families of paralogous genes in prokaryotic genomes has been found to follow a remarkable empirical rule, namely a power law with an exponent between 0 and 1 (the so called Pareto distribution or Zipf law). A number of researchers have pointed this phenomenon and considered various explanations (e. g. [1-5]).

We suggest a mathematical model of the process of random duplications of genes belonging to an existing paralog family and introduction of new genes (as a result of mutations or horizontal transfer), which, with certain probabilities, may or may not give rise to a new paralog family.

2 The Mathematical Model

The suggested model is a Markov branching process. Denote by t the discrete time: $t = 0, 1, 2, \dots$; M - the size of the population of non-duplicating genes; N - the size of the population of duplicating genes; A - the number of families of duplicating genes; N_k - the size of k -th family of duplicating genes ($k = 1, 2, \dots, A$) in the order of its emergence. We assume that at any discrete moment of time exactly one individual gene emerges and, therefore, $M + N = t$. Obviously,

$$\sum_{k=1}^A N_k = N$$

The rules of time evolution are given below:

1. $\Pr\{M(t+1)=m+1 \mid M(t)=m\} = \alpha$
2. $\Pr\{A(t+1)=a+1 \mid A(t)=a\} = \beta$
3. $\Pr\{N_k(t+1)=n_k+1 \mid N_k(t) = n_k \neq 0\} = (1-\alpha-\beta) \frac{n_k}{N}$

Let initial conditions be $M(0) = 0$, $N(0) = 0$. Then the expected values are:

$$E(M) = \alpha t; E(N) = (1-\alpha)t; E(A) = \beta t$$

It follows from (1) (after a tedious calculation) that for $t \gg 1$, asymptotically,

$$E(N_k) = \left(\frac{\beta t}{k} \right)^{1-\frac{\beta}{1-\alpha}} \quad (2)$$

Expression (2) represents a power law with exponent $1 - \frac{\beta}{1-\alpha}$.

The model is based on just two simple and natural assumptions about the process and contains only two free parameters. Nevertheless, as shown below, the distribution given by the model turns out to be in amazingly good agreement (for appropriate values of parameters α and β) with the empirically observed power-law distribution of genes among families of paralogs.

3 Empirical data and results

Since prokaryotic genomes are relatively small, the empirical distributions of sizes of paralog families are affected by large random fluctuations, which make it difficult to discern the theoretical underlying rule. However, this problem can be alleviated by considering ortholog families as representing parallel realizations of the same process of developing paralog families in a number of species. We have used a well-known COG data set [6] to derive an empirical distribution of the average sizes of paralog families. The distribution of family sizes ranked in the decreasing order has been compared with the theoretical curve for the best fitting values of α and β (Fig.1). The empirical distribution follows power law predicted by the theoretical model with highly remarkable precision ($R^2 = 0.995$).

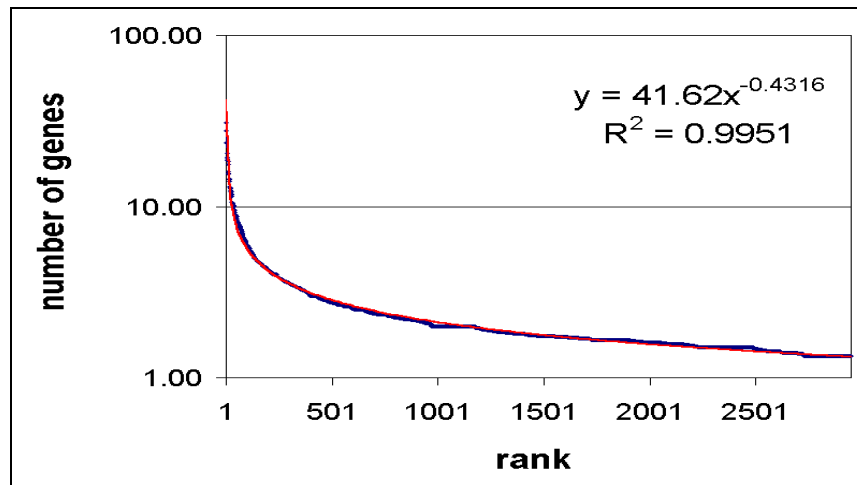


Figure 1: Distribution of the “average” paralog family sizes. Pearson correlation coefficient between the empirical data (blue line) and the theoretical curve (red line) is 0.9951. Here $x = k$, $y = E(N_k)$.

References

- [1] Slonimski P. et al., in *Proceedings of Microbial Genomes II*, Hilton Head, South Carolina, 1998
- [2] Yanai I, Camacho CJ, DeLisi C., Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys Rev Lett.* 2000 Sep 18; 85(12):2641-4.
- [3] Huynen MA, van Nimwegen E., The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol.* 1998 May;15(5):583-9.
- [4] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol.* 2002 Oct 14
- [5] Karev GP, Wolf YI, Koonin EV., Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics.* 2003 Oct 12;19(15):1889-1900.
- [6] Tatusov RL, Koonin EV, Lipman DJ., A genomic perspective on protein families. *Science.* 1997 Oct 24;278(5338):631-7.