DISCRETE AND CONTINUOUS
DYNAMICAL SYSTEMS–SERIES B
Volume **00**, Number **0**, Xxxx **XXXX**

Website: http://AIMsciences.org

pp. **000–000**

A MODEL EXPLAINING THE SIZE DISTRIBUTION OF GENE AND PROTEIN FAMILIES

William J. Reed

Department of Mathematics and Statistics
University of Victoria
Victoria, British Columbia V8W 3P4, Canada

Barry D. Hughes

Department of Mathematics and Statistics
University of Melbourne
Victoria 3010, Australia

(Communicated by ???)

Abstract. This article deals with the theoretical size distribution of gene and protein families in complete genomes. A simple evolutionary model for the development of such families in which genes in a family are formed or selected against independently and at random, and in which new families are formed by the random splitting of existing families, is used to derive the resulting size distribution. Mathematically this turns out to be the distribution of the state of a homogeneous birth-and-death process after an exponentially distributed time, which it is shown will under certain conditions exhibit the power-law behaviour observed for gene and protein family sizes.

1. **Introduction.** Gene families comprise genes with a high degree of similarity in structure and function, which are presumed to have evolved from a single ancestral gene. Protein families similarly comprise proteins sharing sequence similarity and function. It has been observed that the size distributions of both gene families [4] and protein families [1] exhibit power-law behaviour over a wide range. While such distributions are characteristic of complex processes that exhibit self-organized criticality [3], they can also result from simpler mechanisms (see [6]). In this article we show the observed power-law behaviour in the distribution of gene and protein family sizes can be explained using a simple birth-and-death process model for the evolution of families. The model can be fitted by maximum likelihood to family size data, and in an example is shown to provide a very good fit. The model assumes that new genes in a family arise from mutations of existing genes (occurring independently and at random at a fixed probability rate) and that individual genes in a family can be eliminated (again independently and at random at a fixed probability rate). Furthermore new families arise from the random splitting of existing families (again at a fixed probability rate). The model is very similar to that used in a recent paper [5] to explain genus size distributions, and is a development of the model of Yule [7], proposed almost eighty years ago.

1991 *Mathematics Subject Classification.* 92D10.

Key words and phrases. Birth-and-death process, power-law distribution, asymptotic methods, Yule model.

2. **The model.** Consider a gene family which begins with one gene (at time $t = 0$). Suppose that in time $(t, t+h)$ there is a probability $\lambda h + o(h)$ that any given gene may mutate and create in addition to replicates of itself, a new gene in the family (a speciation); and a probability $\mu h + o(h)$ that the individual gene alone is selected out of the genome (an extinction). Suppose further that all speciations and extinctions are independent. Under these assumptions, N_t, the number of genes in the family in existence at time t follows a homogeneous birth-and death process (see *e.g.* [2]) for which the probability mass function (p.m.f.)

$$p_n(t) = \Pr\{N_t = n\} \tag{1}$$

satisfies the differential-difference equation

$$\frac{d}{dt} p_n(t) = -(\lambda + \mu) n p_n(t) + \lambda(n-1) p_{n-1}(t) + \mu(n+1) p_{n+1}(t), \tag{2}$$

with initial condition

$$p_n(0) = 1 \text{ if } n = 1; \quad p_n(0) = 0 \text{ otherwise.}$$

Now let

$$\phi(z;t) = \sum_{n=0}^{\infty} p_n(t) z^n \tag{3}$$

be the generating function for N_t. Multiplying both sides of (2) by z^n and summing over $n = 0, \ldots, \infty$ yields the partial differential equation

$$\phi_t = (\lambda z - \mu)(z-1)\phi_z, \tag{4}$$

with initial condition

$$\phi(z,0) = z. \tag{5}$$

This can readily be solved by the method of characteristics (see *e.g.* [2]) to yield

$$\phi(z,t) = \begin{cases} \dfrac{\mu(1-z) - (\mu - \lambda z)\exp[-t(\lambda-\mu)]}{\lambda(1-z) - (\mu - \lambda z)\exp[-t(\lambda-\mu)]} & \text{if } \lambda \neq \mu, \\ \\ 1 - (1-z)/[1 + \lambda t(1-z)]^{-1} & \text{if } \lambda = \mu. \end{cases} \tag{6}$$

From this the well-known formulas for the p.m.f. of $N(t)$ can be derived. In the case $\lambda \neq \mu$

$$p_0(t) = \frac{\mu - \mu e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}; \tag{7}$$

$$p_n(t) = \frac{(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda - \mu e^{-t(\lambda-\mu)}]^2} \left\{ \frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}} \right\}^{n-1}, \quad n \geq 1. \tag{8}$$

Since different gene families will have will have originated at different times, in order to obtain the p.m.f. (or generating function) of the unconditional distribution of family size $\bar{N}$ say, the p.m.f. $\{p_n(t)\}$ or the generating function $\phi(z,t)$ must be integrated with respect to the distribution of t over gene families. It seems reasonable to assume that any gene family originated when an individual gene in an existing family mutated to a form so different from others in the family that it could no longer be considered a member of that family. Let us suppose that such radical mutations can occur in any existing family in a time interval of length h with probability $\rho h + o(h)$. This implies that the number of families follows a Yule process [7], and that the time in existence of any family will follow a truncated exponential distribution (the truncation time being the time since the establishment of the first gene family). Since evolution has been happening for a very long time, this truncation can essentially be ignored, so that the p.m.f. of the distribution of current gene family size, $\bar{N}$, is

$$q_n = \Pr(\bar{N} = n) = \int_0^\infty p_n(t)\rho e^{-\rho t} dt \text{ for } n = 0, 1, \ldots \tag{9}$$

and its generating function is

$$\bar{\phi}(z) = \int_0^\infty \phi(z,t)\rho e^{-\rho t} dt. \tag{10}$$

Neither of the integrals in (9) or (10) have simple closed-form expressions. However by returning to the partial differential equation (4), multiplying throughout by $\rho e^{-\rho t}$ and integrating with respect to t between 0 and ∞, one arrives at the following ordinary differential equation for $\bar{\phi}(z)$:

$$(\lambda z - \mu)(z-1)\frac{d\bar{\phi}}{dz} - \rho\bar{\phi}(z) = -\rho z. \tag{11}$$

While this can be solved in terms of Lerch's phi function (see [5]), the form is not particulary useful. However a series solution can be obtained by equating coefficients of powers of z on both sides of (11), yielding the following recursion for the p.m.f. $\{q_n\}$

$$(n-1)\lambda q_{n-1} - [n(\lambda+\mu)+\rho]q_n + (n+1)\mu q_{n+1} = 0, \text{ for } n \geq 2 \tag{12}$$

$$-(\lambda+\mu+\rho)q_1 + 2\mu q_2 = -\rho \tag{13}$$

$$-\rho q_0 + \mu q_1 = 0. \tag{14}$$

The p.m.f. $\{\tilde{q}_n\}$ of the size of a non-extinct gene family is obtained as

$$\tilde{q}_n = \frac{q_n}{1-q_0}, \; n = 1, 2, \ldots \tag{15}$$

3. **Power-law behaviour in the family size distribution.** It is shown in this section that in the case $\lambda > \mu$, the distribution of family size exhibits power-law behaviour in the upper tail. Consider (from (9) and (8))

$$q_{n+1} = \int_0^\infty \frac{\rho e^{-\rho t}(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda - \mu e^{-t(\lambda-\mu)}]^2} \left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n dt. \tag{16}$$

The factor in braces is strictly increasing in t from zero at $t = 0$, approaching unity as $t \to \infty$. Effecting a change of variable from t to τ with

$$t = (\lambda-\mu)^{-1}\log[n(1-\mu/\lambda)/\tau], \tag{17}$$

gives for $n \to \infty$,

$$\left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n \sim \left\{1 - \frac{\tau}{n}\right\}^n \to e^{-\tau}, \tag{18}$$

so that in (16)

$$q_{n+1} \sim \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)} n^{-1-\rho/(\lambda-\mu)} \int_0^\infty e^{-\tau}\tau^{\rho/(\lambda-\mu)} d\tau. \tag{19}$$

The integral can be evaluated as a gamma function. The important conclusion is that for large n, the p.m.f. exhibits power-law behaviour *i.e.*

$$q_n \sim c_1 n^{-(\rho/(\lambda-\mu)+1)}, \tag{20}$$

where

$$c_1 = \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)}\Gamma(1+\rho/(\lambda-\mu)) \tag{21}$$

It follows that $\tilde{q}_n$ will also exhibit asymptotic power-law behaviour with exponent $-(\rho/(\lambda-\mu)+1)$.

For $\lambda \leq \mu$, it can be shown that q_n does not exhibit exact power-law behaviour, but rather behaves in a 'stretched exponential' form.

4. **A special case - no extinctions.** If the extinction rate parameter μ is set to zero, the birth-and-death process reduces to the Yule process and the resulting size distribution is the eponymous Yule distribution [7] with p.m.f.

$$q_n = \Big(\frac{\rho}{\lambda}\Big)\frac{\Gamma(\rho/\lambda+1)\Gamma(n)}{\Gamma(\rho/\lambda+n+1)}, \text{ for } n = 1, 2, \ldots. \tag{22}$$

This exhibits power-law behaviour in the upper-tail, and indeed as Yule showed, exhibits almost linear behaviour in the log-log plot, over the whole range.

5. **Fitting the model to data.** The model can be fitted to gene family size data by estimating model parameters by maximum likelihood (ML). The log-likelihood is

$$\ell = \sum_{i=1}^{N} f_i \log \tilde{q}_i = \sum_{i=1}^{N} f_i \log q_i - \log(1-q_0)\sum_{i=1}^{N} f_i \tag{23}$$

Although there are three parameters in the model, the p.m.f. $\{q_i\}$ depends only on the two ratios $\lambda/\rho = \Lambda$, say, and $\mu/\rho = \mathrm{M}$, say. To calculate numerically the log-likelihood for given values of Λ and M, the recursion (12) can be iterated backwards from the asymptotic value (20) for a suitably large n. ML estimates of Λ and M can be found by numerically maximizing the log-likelihood.

This has been done for data on 71 gene families of cytochrome P450[1]. The ML estimates are:

$$\widehat{\Lambda} = 3.134;\ \widehat{\mathrm{M}} = 2.759.$$

It should be noted that these estimates are highly correlated and confidence intervals for individual parameters very wide (95% confidence intervals using the profile log-likelihoods are: 0.72 – 81.4 for Λ and 0 – 90.0 for M). Thus although the ML estimates suggests that the rates of mutation and extinction of individual genes are about three times the rate at which new families are formed, it is quite plausible, based on the data alone, that they are up to a factor of 80 or so larger—or even smaller. Fig. 1 shows the observed frequencies (natural and binned) of gene families of various sizes along with the expected frequencies using the above ML estimates (joined by a solid line for illustration purposes). The chi-squared goodness of fit test (on binned data) suggests an extremely good fit (1.459 on 6 degrees of freedom).

It should be noted that the data are consistent (P-value = 0.23, for a likelihood ratio test) with the extinction parameter μ being zero, *i.e.* with the Yule model. For the Yule model, the MLE of Λ is 0.907 (chi-square statistic = 1.890 on 7 degrees of freedom). While the Yule model thus seems to fit the data well statistically, the MLE for the model suggests that new gene families emerge at a slower probabilistic rate than new genes within a family, which from a biological point of view, seems

[1] Data from http://arabidopsis.org/info/genefamily/p450.html, a website of the Arabidopsis Information Resource (TAIR).

implausible. The expected frequencies using the MLE of Λ under the Yule model are illustrated by the dotted line in Fig. 1.

6. **Conclusions.** Earlier authors have shown empirically that the size distributions of gene and protein families exhibit power-law behaviour. This paper offers a simple evolutionary model which predicts a theoretical size distribution which exhibits such behaviour and which can be fitted to family-size data by maximum likelihood. The model is one of neutral evolution in the sense that mutations creating new genes (or new families) and extinctions of genes occur independently and at random. It can be thought of as a null model, in that it does not include interactions in the evolution of genes, temporal heterogeneity or other complex aspects, but can nonetheless explain the observed size distributions of gene and protein families.

REFERENCES

[1] J. S. Bader, Evolutionary implications of a power-law distribution of protein family sizes, http://arxiv.org/abs/physics/9908032.

[2] N. T. J. Bailey, "The Elements of Stochastic Processes", John Wiley and Sons, New York (1964).

[3] P. Bak, C. Tang and K. Wiesenfeld, Self-organized criticality: an explanation of $1/f$ noise, Phys. Rev. Lett., 59 (1987), 381–384.

[4] M. A. Huynen and E. van Nimwegen, The frequency distribution of gene family sizes in complete genomes, Mol. Biol. Evolution, 15 (1998), 583–589.

[5] W. J. Reed and B. D. Hughes, On the size distribution of live genera, J. Theoret. Biol., 217 (2002), 125–135.

[6] W. J. Reed and B. D. Hughes, From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature, Phys. Rev. E, in press.

[7] G. U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S., Phil. Trans. Roy. Soc. Lond., Series B, 213 (1924), 21–87.

Received October 2002.

E-mail address: reed@math.uvic.ca
E-mail address: hughes@ms.unimelb.edu.au

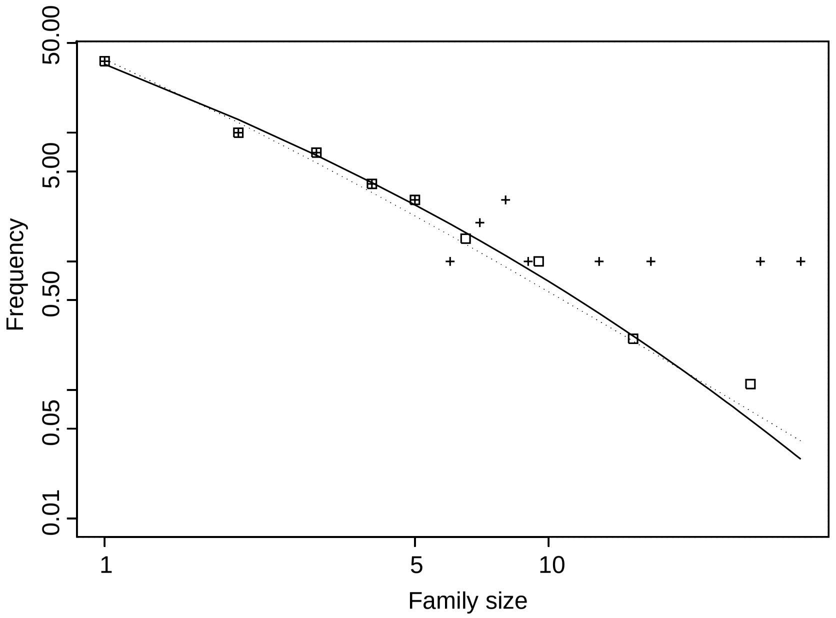

FIGURE 1. The size distribution of cytochrome P450 gene families. The crosses show the observed frequencies. Note that larger sizes occur once only. The boxes show the same data binned for larger less frequent sizes. The solid line shows the expected frequencies using the ML estimates of model parameters Λ and M, while the dotted line uses the ML estimate of Λ with M set at zero.