CORNELL UNIVERSITY MATHEMATICS DEPARTMENT SENIOR THESIS

Power Law Distributions of Gene Family Sizes

A THESIS PRESENTED IN PARTIAL FULFILLMENT
OF CRITERIA FOR HONORS IN MATHEMATICS

Elizabeth Rach

May 2004

BACHELOR OF ARTS, CORNELL UNIVERSITY

THESIS ADVISOR(S)

Richard Durrett
Department of Mathematics

Cornell University

Abstract

POWER LAW DISTRIBUTIONS OF GENE FAMILY SIZES

By Elizabeth Rach

May 12, 2004

Faculty adviser and supervisor:

Professor Rick Durrett, Department of Mathematics

Power law distributions appear in various biological and physical contexts. In the genomic world, researchers have shown that power law distributions accurately describe the sizes of gene families and protein folds. In this paper, the mathematical construction and genetic application of three models are compared: the Preferential Attachment Model, the Branching Process with Immigration Model, and the Birth Death and Innovation Model. No individual model offers a complete explanation, however, together, they serve to emphasize important components in network systems: preferential attachment, natural selection, gene birth (duplication), gene death, and gene transfer(innovation).

Table of Contents

Power Law Distributions of Gene Family Sizes

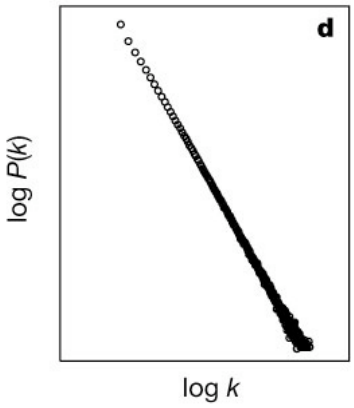
Elizabeth Rach

Introduction

A, C, T, and G are four simple letters that comprise the delicate code of all life on Earth. The sequence of any one species provides scientists with a snapshot of evolutionary time, however, little is known about the origin of such genetic networks or the interactions that govern their existence. Whole genomes can be divided into families based on phylogenetically common ancestors, and similar function regulations. Gene families can range in size from containing 3 genes to as large as 700 or 800 genes, as with zinc fingers. Research has shown that gene family sizes do not occur with equal probability in the genome. Rather, small gene families occur more frequently than large gene families, however, as overall genome size increases across species, the frequency of large gene families increases at a rate faster than the frequency of small gene families. Mathematically, the probability of observing a gene family of size k within one genome can be described by a power law distribution:

$P(k) = k^{-\gamma}$ where k = size of gene family
γ = growth rate of gene family.

Graphically, the power law distribution maintains the same shape, regardless of scale, and can be represented as a straight line on double logarithmic axes.



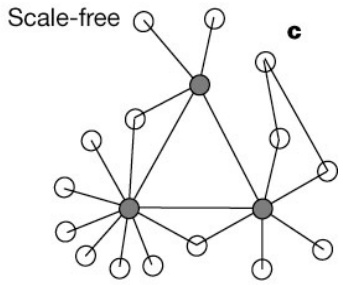
(Jeong and Tombor 2000, p. 652 Figure 1 d)

Although many studies have found power laws, a good explanation for its occurrence still remains to be discovered. In this paper, I will present three possible models that offer potential insights into the mechanisms underlying power law distributions for gene family sizes: the Preferential Attachment Model, the Branching Process With Immigration Model, and the Birth Death and Innovation Model.

Preferential Attachment

1.1 Mathematical Construction

The Preferential Attachment Model was formulated by Barabàsi and Albert(1999). In *Emergence of Scaling in Random Networks*, Barabàsi et al. describes the human genome as an oriented graph, in which genes are vertices and interactions between genes are edges. Unlike previous theory, Barabàsi defined the probability that a family gains a new member as proportional to the degree of connectedness within the family. This means that gene families with larger numbers of interactions have a higher probability of increasing in size.



(Jeong and Tombor 2000, p. 652 Figure 1 c)

Through simulations, Baràbasi et. al. showed that the model consistently produces a power law distribution.

For

m = the number of interactions between genes in a family,

x = constant,

$$\int_{m}^{\infty} (2m^2) / x^3 \, dx = 1.$$

So, Baràbasi concluded that,

for k = size of the gene family,

P(k) = probability of observing a gene family of size k,

$P(k) = \frac{2m^2}{k^3}$ for $k \geq m$, (Barabàsi and Albert 1999, p. 511).

In 2001, Andrey Rzhetsky and Shawn Gomez extrapolated Baràbasi's idea of preferential attachment for a scale-free network. They created a homogeneous continuous-time Markov Process that incorporates gene duplication (λ), and birth (μ). By letting N = the maximum size of the genome, $D(t)$ = the total number of interactions between genes, and t = time, Rzhetsky and Gomez assumed that

$$D(t + \Delta t) = D(t) + \mu D(t)\, \Delta t.$$

Next, for the purposes of graph orientation, if

$d_{i,u}(t)$ = the number of family sizes at time t with i outgoing interaction edges, and $d_{j,d}(t)$ = the number of family sizes at time t with j incoming interaction edges,

then,

$$d_{1,u}(t + \Delta t) = d_{1,u}(t) + \mu D(t)\, \Delta t - \lambda_u\, d_{1,u}(t)\, \Delta t + o(\Delta t)$$

$$d_{i,u}(t + \Delta t) = d_{i,u}(t) + \lambda_u d_{i-1,u}(t)(i-1)\, \Delta t - \lambda_u d_{i,u}(t) i \Delta t + o(\Delta t) \text{ for } 2 \leq i \leq N - 1$$

$$d_{N,u}(t + \Delta t) = d_{N,u}(t) + \lambda_u d_{N-1,u}(t)\, (N-1)\Delta t + o(\Delta t).$$

This means that if $p_{i,u}(t)$ = the proportion of classes with i interaction connections, then

$$p_{i,u}(t) = \frac{d_{i,u}(t)}{D(t)}$$

Now, by substitution,

$$p_{1,u} = \mu(1-p_{1,u}) - \lambda_u p_{1,u},$$

$$p_{i,u} = \lambda_u(i-1)\, p_{i-1,u} - \lambda_u i\, p_{i,u} - \mu\, p_{i,u} \qquad 2 \leq i \leq N-1$$

$$p_{N,u} = \lambda_u(N-1)\, p_{N-1,u} - \mu\, p_{N,u}.$$

So, the Markov matrix associated with this system of differential equations is:

$$\mathbf{Q}_u = \begin{pmatrix} -\lambda_u & \lambda_u & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \mu & -\mu - 2\lambda_u & 2\lambda_u & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \mu & 0 & -\mu - 3\lambda_u & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mu & 0 & 0 & \cdots & & -\mu - i\lambda_u & i\lambda_u & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\mu - (N-1)\lambda_u & (N-1)\lambda_u \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -\mu \end{pmatrix}.$$

(Rzhetsky and Gomez 2001, p. 992).

The solution for this system of u outgoing edges is:

$$\Pi(t) = \Pi(0)\, P_u(t) = \Pi(0)\, e^{Qut},$$

where $\Pi(t) = [p_{1,u}(t), p_{2,u}(t), \ldots, p_{N,u}(t)]$ and,

$$\Pi(0) = [1,0, \ldots, 0].$$

By solving this equation, Rzhetsky and Gomez (993) discovered that the frequency E of vertices with k incoming or outgoing edges is:

$$E_{k,in}(t) = p_{k,u}(t)$$
$$E_{k,out}(t) = p_{k,d}(t)\ .$$

1.2 Simulations

The group tested the fit of the Scale-Free Network model in MatLab by estimating the rates of outgoing interactions $\lambda_d t$, incoming interactions $\lambda_u t$ and duplication events μt, in Yeast and *E. Coli*. They minimized the data's deviations from the power law distribution by choosing parameters such that

$$\sum_{k=1}^{N} (O_{k,in} - E_{k,in})^2 + \sum_{k=1}^{N} (O_{k,out} - E_{k,out})^2 \rightarrow 0.$$

The following estimations resulted:

	$\lambda_u t$	$\lambda_d t$	μt
Yeast	17.44	33.05	33.92
E. Coli	4.63	4.63	7.29

Upon further comparison, Rzhetsky and Gomez found these values to be almost identical to listed rates in the Database of Interacting Proteins(993).

1.3 Discussion

The Scale-Free Preferential Attachment Model has shown to accurately describe a variety of biological systems such as metabolic pathways, social connections, and the World Wide Web. Its good fit to the data, however, does not imply that it is the right model for genomic applications. In *The Structure of the Protein Universe and Genome Evolution*, Koonin argues that gene families of the same size often have disparate growth rates(220). Thus, while preferential attachment may help to explain the difference in rates between small and large families, it does not answer the question of natural selection amongst families of the same size.

The Preferential Attachment Model is still valuable, however, as it exemplifies robustness against numerical fluctuations that can result from mutations, or environmental changes(Jeong and Tombor 2000, p. 651). This supports the evolutionary theory of the specialization of organisms in ecological niches. It also secures a higher probability of survival for a species. Moreover, preferential attachment may be an inherent mechanism of nature's defense.

Branching Process With Immigration

2.1 Mathematical Construction

In order to visualize this mathematical model, it may be helpful to initially build a graph. Let the x-axis represent time and the y-axis be the size of the human genome measured by the number of genes. As this model will be based on exponential growth, it will take a finite amount of time t for the maximum size of the genome N, to be fully reached. If we define a family to consist of all genes that can be traced back to a common ancestor and biologically perform the same functions, then N can be divided into separate classes of gene families of similar size. In this way, we can think of our graph as a giant square with time ranging from 0 to t on the x-axis and the size of the genome ranging from 0 to N on the y-axis.

The model is based on the idea that genes immigrate, through duplication and specialization, into a family at a constant rate r', and out of a family at rate r''. Thus, there is a net flow of immigration at rate r = r' – r''. If

1) the number of individuals immigrating at one time interval does not affect the number immigrating at different time intervals,
2) the average rate of immigration remains constant, and
3) individuals immigrate one at a time,

then, gene families grow according to a Poisson Process with parameter _. A Poisson Process is a Continuous time Markov chain (stochastic process), in which waiting times are exponentially distributed and the number of events are Poisson distributed.

As time increases, not all duplicated and specialized copies in a gene family will persist due to evolutionary constraints like natural selection and environmental adaptation. If we consider only one of the lineages in a family that are "born" and are able to persist to time t, then the descendants of a single immigrant becomes a special kind of a Markov Process called a Yule Process, in which each particle splits into two at a constant rate _ (Durrett and Schweinsberg 2003, p. 6).

For $f(t)$ = the size of a gene family at time t,

There exists $w \in R$ such that

$$\lim_{t \to \infty} \frac{f(t)}{e^{\lambda t}} \to w.$$

So, for the i^{th} gene family, where i = 1,2,3,…

$f_i(t) \sim w_i e^{\lambda t'}$ where λ = the average rate at which new genes are born.

We have changed the scale of time so that a single immigration or duplication occurs at rate $\lambda = 1$. As immigrants constantly enter at rate r, this forces the collection of Yule Processes to "compete" against each other for larger size. As a result, each k families out of N genes only experience N/k events and the size of any one gene family relative to the size of all of the gene families begun at or before a certain time t' is Beta distributed(Joyce and Tavre 1987):

$$\frac{e^{t'} w_1}{e^{t'} w_1 + e^{t'} w_2 + \ldots + e^{t'} w_k} = \frac{w_1}{w_1 + w_2 + \ldots + w_k} \sim \text{Beta } (1, k\text{-}1) .$$

By implementing a similar construct to the stick-breaking model, Durrett and Schweinsberg have shown that at time t, the expected size of a gene family born on the k^{th} event is:

N/k x (fraction of the population in families that had started by event k).

The conditional probability that a randomly chosen individual belongs to a family born on the j^{th} event is r (1/j). Thus, the conditional probability that a randomly chosen individual does *not* belong to a family born on the j^{th} event is (1- r/j). So, the fraction of the population in families that was not born on or after the k^{th} event (born before the k^{th} event) is:

$$\prod_{j=k}^{N} (1 - r/j) \quad \text{as } r/j \to 0, \quad \approx \quad \exp\left(-r \sum_{j=k}^{N} 1/j\right) \quad \approx \quad (k/N)^r .$$

By substitution, we have that the expected size of a gene family at time t born on the k^{th} event is:

$$(N/k)\,(k/N)^{r} = (N/k)^{1-r}.$$

For $S \in \mathbb{Z}$, a gene family should at least be size S if $(N/k)^{1-r} \geq S \rightarrow k < NS^{(-1/1-r)}$. So, the *number* of families at least size S is $\approx r\,N\,S^{(-1/1-r)}$, and the *number* of families of exactly size S is:

$$number = \frac{-d}{ds}\left(r\,N\,S^{(-1/1-r)}\right) = \frac{rN}{1-r}\,S^{(-1/1-r\ -1)} = \frac{rN}{1-r}\,S^{-(2-r/1-r)}.$$

Therefore, $\log(number) \approx \log(r\,N/1-r) - (2-r/1-r)\log(\text{size } S)$.

2.2 Simulations

Assumptions

The model was defined by Jiang Qian, Nicholas Luscombe and Mark Gerstein in "Protein Family and Fold Occurrence in Genomes: Power-law Behaviour and Evolutionary Model(2001)." In the paper, Qian et. al. demonstrates that power law distributions can be produced in biological phenomenon such as protein fold occurrences, and gene family sizes under the branching process with immigration model. They fit the data to the model by making two assumptions:

1) r = fixed rate of immigration

r = rate of gene acquisition – rate of gene deletion

$$r = \frac{N_{genes} - N_o}{(C-1)\,N_{genes}}$$

where N_o = initial number of genes
N_{genes} = total number of genes in the genome, and
C = average rate of gene duplication.

2) The absolute value of the slope of the graph is always at least 2. This implies that r must be between 0 and 1.

Qian claims that for $1 > r \geq 0$, an exponential distribution is converted to a power law distribution after large values of t (many generations), while for $r < 0$, a power law is never approached, regardless of the length of time.

Testing Qian's Hypothesis

For the purpose of evaluating this model, we make computer simulations to compare the accuracy of mathematically predicted parameters to actual parameters observed in living organisms. In his study, Qian focused on modeling smaller genomes such as E. Coli and S. Cerevisiae. Therefore, in our simulations, we use a genome size of N = 10,000 genes.

After 10,000 repetitions, the following linear regressions on means are produced:

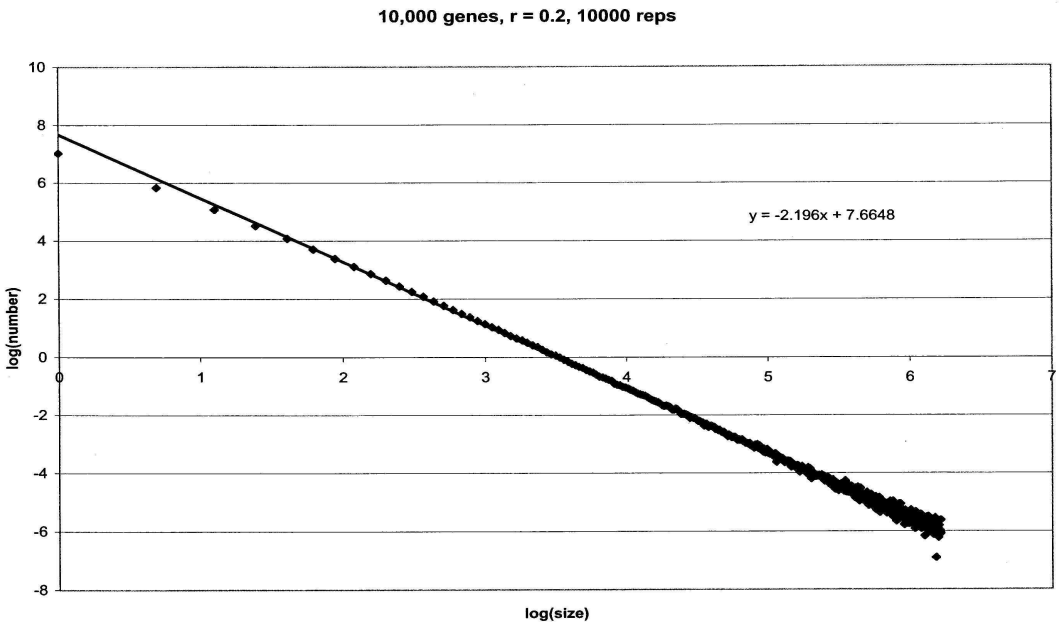
Immigration Rate	Fitted Line
r = .05	y = -2.0241x + 6.1241
r = 0.1	y = -2.0811x + 6.8936
r = 0.2	y = -2.196x + 7.6648.

From our mathematical construction,

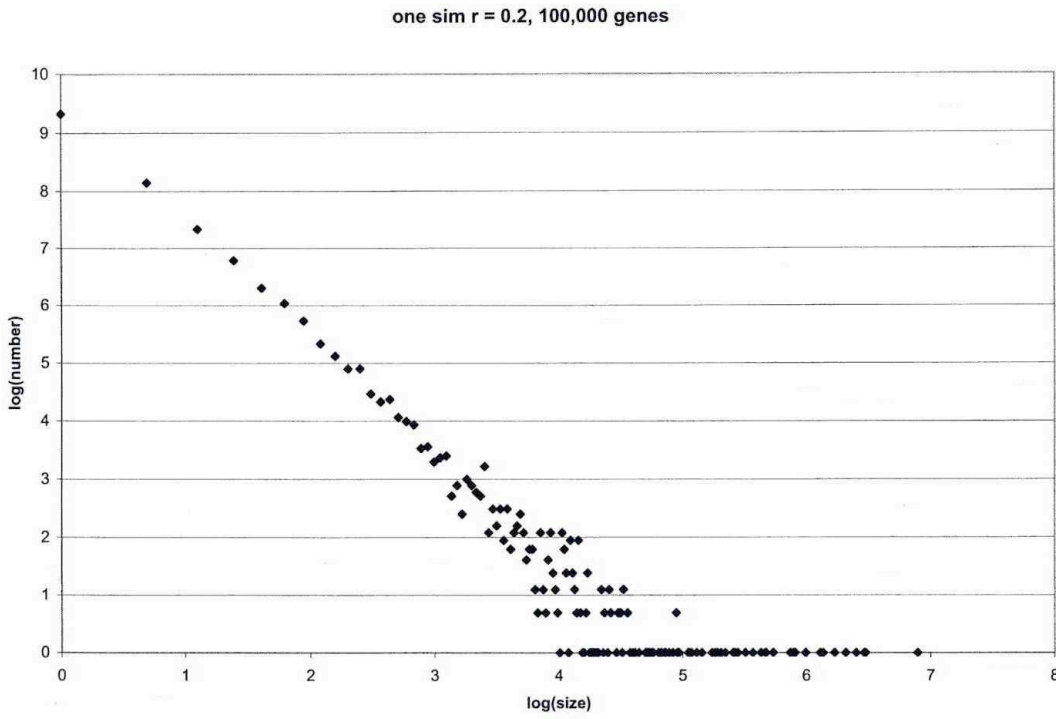
Imm. Rate	Predicted slope = -(2-r)/(1-r)	Predicted Intercept = log(rN/(1-r))
r = .05	-2.05	6.26
r = 0.1	-2.11	7.01
r = 0.2	-2.25	7.82

By calculating differences in residuals, we find that these simulations are within 5% of the actual, observed data. Thus, with these parameters, Qian's model produces the power law distribution.

Minor changes can alter the quality of the model's results. For only 1 repetition, the power law becomes less defined, regardless if the number of genes in the genome is increased to 100,000.

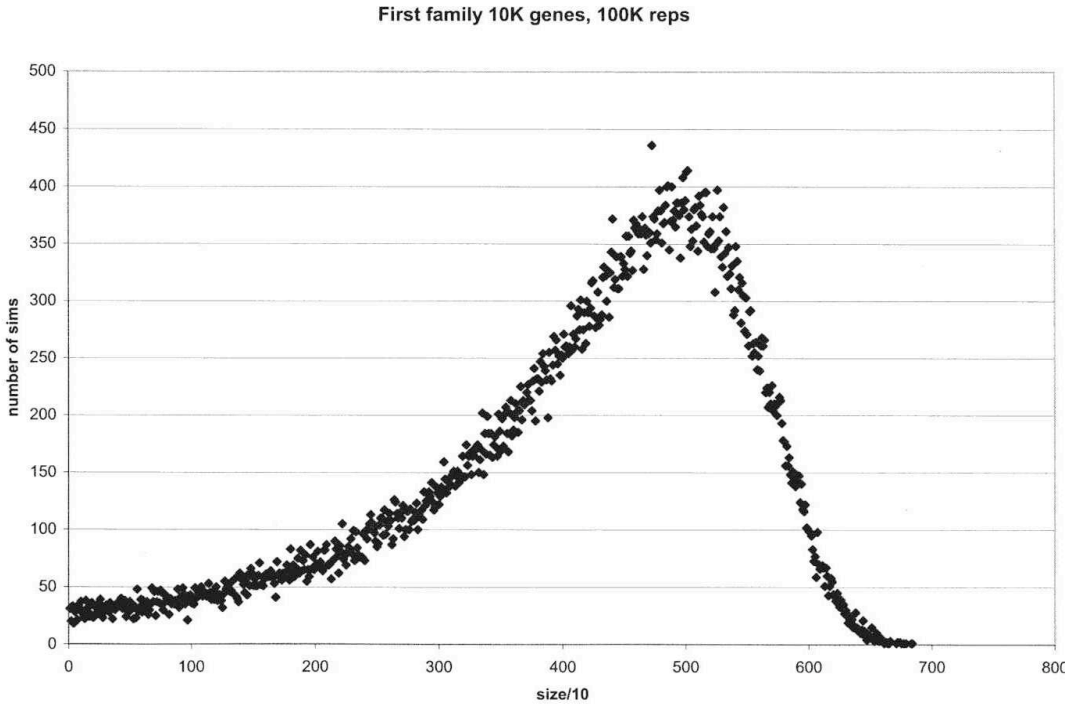


Power Law Distribution.

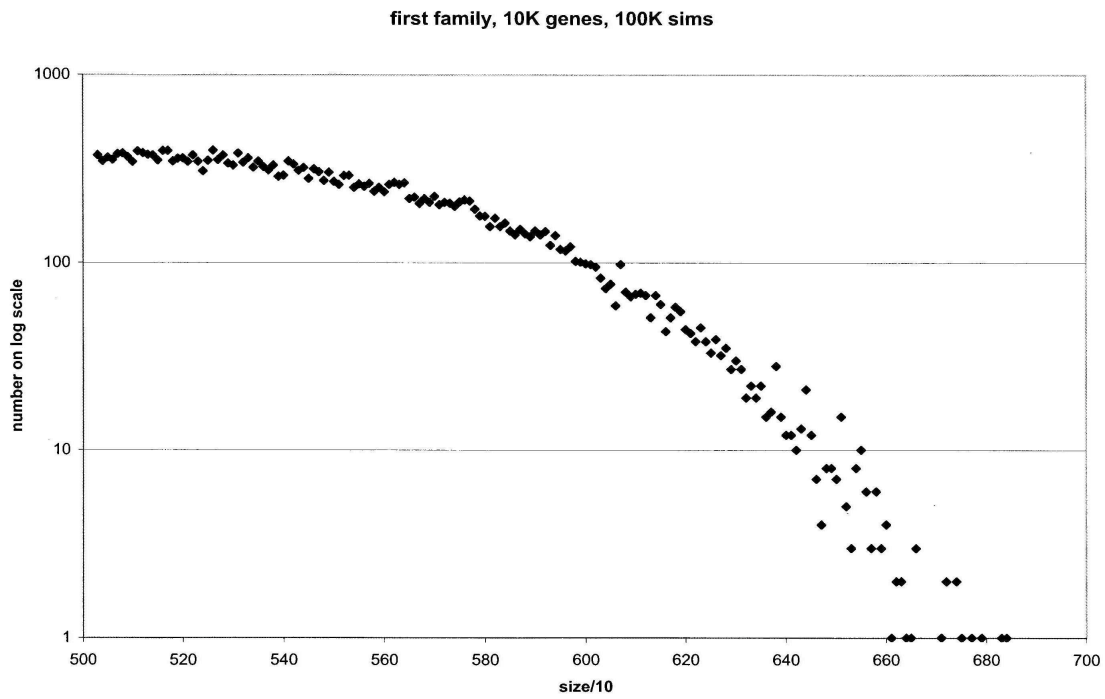


Distribution for one repetition.

The distribution is also less accurate for modeling the size of the first families. Simulations show that the regression deviates from the power law by increasing in a curvature pattern then, immediately decaying exponentially. This distribution was tested to fit a scale of 10, however, greater residuals resulted from a skewed density function and curved distribution.



Number of simulations v. distribution scaled by 10.



Gene family frequency v. distribution scaled by 10.

2.3 Discussion

The branching process with immigration model is significant because it achieves a power law distribution from a simple stochastic process. The degree of biological realism of the model, however, is questionable. In the mathematical constructions and simulations, we have shown that for $1 > r \geq 0$, $|\text{slope}| > 2$ for all organisms. At the beginning of the article, Qian et. al. state that after running regression analyses on data from over 20 organisms' actual genomes, they have found that in general, | slope | ranges from .9 - 1.2 for eukaryotes and between 1.2 – 1.8 for prokaryotes(Qian and Luscombe 2001, p. 676). This correlation directly contradicts our assumptions. It is possible to achieve smaller slopes by setting the initial population size greater than 1, however this evidence still provides support against the biological realism of the model.

A second area of concern in the model is the assignment of equal birth rates to all genes. One of the essential forces driving evolution is the difference in selective pressures acting on specific areas in the genome. It has been well

documented that certain “hotspots” continuously duplicate and specialize, while others can not see any action for generations. Thus, the designation of all genes having equal immigration rates is highly unrealistic.

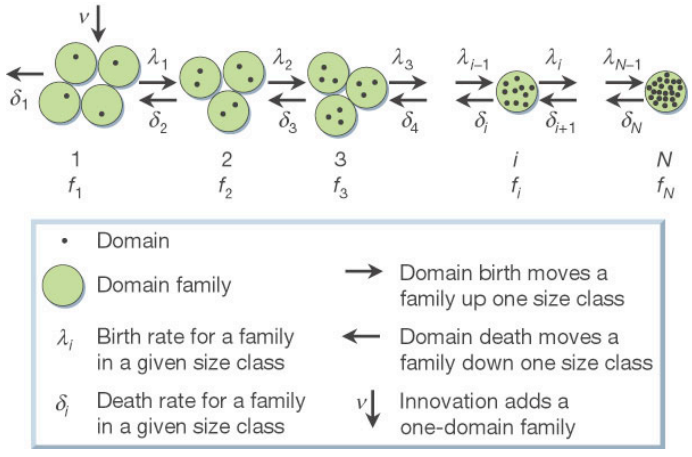
The constant inflow of genes being born is a third issue fundamental to the construction of the model. The predicted immigration rates by the model from the 20 organism’s that Qian’s research team studied are remarkably high. Mycoplasma pneumoniae and Rickettsia prowazekii are recorded in Table 1 to be .64 and .55, respectively(679). These rates are very large! That means that for each generation, about 64% of all of the Mycoplasma pneumoniae’s genome are new immigrants! While it is true that organisms with smaller genomes show higher rates of recombination and gene duplication, manipulating the rates this much to fit the model seems highly unrealistic. Often, immigration rates fall < 0, and while the article claims r can never be < 0 , if we merely compare the loss of olfactory genes in the human over evolutionary time, we can easily see that this is obviously not true. Qian even states himself that gene loss is a great factor in reductive evolution in organisms such as E. coli and Haemophilus influenzae(679). So, while the model is a great predictor for organisms in which the immigration rate of genes naturally ranges from 0 to 1, a more realistic model should accommodate all r for organisms large and small.

Regardless of these flaws, the model must still be recognized as possessing evolutionary merit. Firstly, it incorporates two major components capable of producing a power law over long evolutionary time: gene duplication and immigration. Secondly, it supports the derivation of organisms from a common ancestor by the increased conservation of the largest gene families. The largest families are typically the oldest, so, an increase in their conservation across species implies the necessity of their existence in satisfying the functional requirements of survival for a common ancestor. Thus, while the branching process with immigration model may not be the most biologically sound production of the power law distribution, it can still be accredited for supporting some basic evolutionary fundamentals.

Birth, Death, and Innovation Model (BDIM)

3.1 Mathematical Construction

The BDIM model is built on the same definitions of a gene family and the process of gene birth by duplication and divergence, as that in the Branching Process with Immigration model. Similarly, it also restricts one event to occur at a time, and defines N as the maximum number of copies of any gene, and f_i as the number of gene families of size *i*. The BDIM model further develops the constructs of the Branching Process with Immigration model, by incorporating the existence of gene death from inactivation or deletion, and the occurrence of gene family innovation from recombined coding sequence or gene transfer across species. The rate of innovation(ν) is assumed to be constant, while the birth rate (λ) and death rate (δ) vary. These dynamics of the model can be illustrated as:



(Koonin and Wold 2002, p. 220).

We can see that unlike Barbàsi's network theory, the BDIM model does not require the birth rate and death rate for a gene family of size *i* to be proportional to *i.* So, the change in gene family size over time can be characterized as:

$df_i(t)/dt$ = (families grown from size i-1) + (families diminished from size i+1) – (families of size i who grow or diminish).

Mathematically, this is:

$$df_1(t)/dt = v + \delta_2 f_2(t) - (\lambda_1 + \delta_1) f_1(t)$$

$$df_i(t)/dt = \lambda_{i-1} f_{i-1}(t) + \delta_{i+1} f_{i+1}(t) - (\lambda_i + \delta_i) f_i(t) \qquad \text{for } 1 < i < N$$

$$df_N(t)/dt = \lambda_{N-1} f_{N-1}(t) + 0 - \delta_N f_N(t).$$

For $F(t)$ = total number of gene families at time t,

$$F(t) = \sum_{i=1}^{N} f_i(t), \qquad \Rightarrow \qquad dF(t)/dt = v - \delta_1 f_1(t).$$

If enough time passes, the system can reach a globally, aymptotically, stable, unique equilibrium, where $df_i(t)/dt = 0$ for all i, $dF(t)/dt = 0$, and a balance between birth and death rates of gene families of size 1 is attained, $v = \delta_1 f_1(t)$.

Furthermore, at equilibrium, the asymptotic behavior of a solution can be defined by the relation between λ_i and δ_i. Let $\chi(i)$ be a function of power growth, such that

$$(*) \qquad \chi(i) = \lambda_{i-1} / \delta_i = i^s \theta(1 + a/i + O(1/i^2)) \qquad \text{for } a,s \in \mathbb{R}, \theta > 0.$$

Define the BDIM to be:

1.) non-balanced

if $s \neq 0$, => $\chi(i) = \lambda_{i-1} / \delta_i = i^s \theta(1+a/i + O(1/i^2))$

2.) first-order balanced

if $s = 0, \theta \neq 1$, => $\chi(i) = \lambda_{i-1} / \delta_i = \theta(1+a/i + O(1/i^2))$

3.) second-order balanced

if $s = 0, \theta = 1, a \neq 0$ => $\chi(i) = \lambda_{i-1} / \delta_i = 1+a/i+O(1/i^2)$

4.) higher-order balanced

if $s = 0, \theta = 1, a = 0$ => $\chi(i) = \lambda_{i-1} / \delta_i = 1+ O(1/i^2)$.

Let equilibrium frequencies $p_i = f_i / F_{eq}$ such that

$$p_i \sim \prod_{k=1}^{i-1} \lambda_k / \prod_{k=1}^{i} \delta_k \sim \prod_{s=2}^{i} (\lambda_{s-1} / \delta_s) .$$

If the model is balanced,

$$p_i \sim \Gamma(i)^s \theta^i \prod_{s=2}^{i} (1+ a/s) ,$$

where $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$

Through integration by parts, $\Gamma(r+1) = r \Gamma(r)$.

So, this means that

$$p_i \sim \Gamma(i)^s \, \theta^{\,i} \, \Gamma(i+a+1) / \Gamma(i+1).$$

For large i and any c,

$$\Gamma(i+c) / \Gamma(i) = (i-c-1)\,(i-c-2) \cdots (i)$$

For $i >> c$,

$$\Gamma(i+c) / \Gamma(i) \sim i^{\,c} \quad .$$

Hence, our relation becomes:

$$\Gamma(i+a+1)/\,\Gamma(i+1) \sim i^a \quad \text{, and} \qquad p_i \sim \Gamma(i)^s \, \theta^{\,i} \, i^a$$

(Karev and Wolf 2002, p. 23).

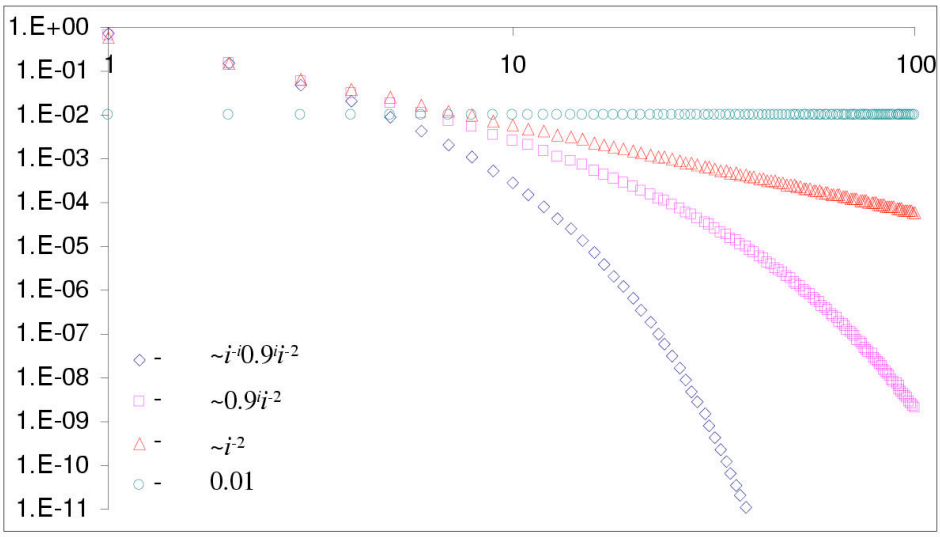
Thus, by substituting the appropriate parameters, the following BDIM asymptotics result:

1.) $p_i \sim \Gamma(i)^s \theta^{\,i} i^a$,

2.) $p_i \sim \theta^{\,i} i^a$

3.) $p_i \sim i^a$,

4.) $p_i \sim 1$.



Balanced BDIMs of different orders(Karev and Wolf 2002, p. 7).

From these results, it follows that the equilibrium frequencies p_i : increase or decrease exponentially fast with the increase of i in non-balanced BDIMs, approach Pascal, logarithmic or geometric distributions, depending on a, in first-order balanced BDIMs ($\theta < 1$), and asymptotically ascertain uniform distributions in higher-order balanced BDIMs.

Power laws are produced if and only if the BDIM is second-order balanced! This outcome is crucial in the development of the model and can be further investigated through adjustments of birth and death rates.

While all four sub-models serve to describe biological phenomenon, the *linear* BDIM most convincingly approximates genome family sizes. This will be shown in simulations.

If $\lambda \neq \delta$, a linear first-order balanced BDIM is produced with equilibrium frequencies p_i that follow a logarithmic distribution. Conversely, if $\lambda = \delta$, a linear second-order balanced BDIM is created, in which the equilibrium frequencies follow the power law $p_i \sim i^{a-b-1}$ (Karev and Wolf 2002, p. 11). By this distribution, we can see that the frequency of gene family sizes depends on the values of $a,b \in$ constants.

If $a > b$, a family's overall gene birth rate decreases faster than its gene death rate. This causes genes to compete against each other for survival, leading

to fewer larger families. On the contrary, if a < b, a family's gene birth rate decreases slower than its death rate. Selection is relaxed, and an increased number of larger families are produced.

Within a family, on average, a gene has birth rate $\lambda_i / i = \lambda + \lambda a / i$, and death rate $\delta_i / i = \delta + \delta b / i$. So, for small i, $\lambda_i / i \Rightarrow \lambda + \lambda a$, and $\delta_i / i \Rightarrow \delta + \delta b$, while for large i, $\lambda_i / i \Rightarrow \lambda$, and $\delta_i / i \Rightarrow \delta$. This means that if a and b are both positive, a gene family's overall birth and death rate decrease inversely to increasing gene family sizes. While, if a and b are both negative, the birth and death rates both increase proportional to larger gene family sizes.

The constants a and b also play a role in the ratio of birth rate to innovation rate, G(N):

$$G(N) = \sum_{i=1}^{N-1} \lambda_i f_i / \nu \quad \cong \quad \frac{\Gamma(1+b)\, N^{1+a-b}}{(1+a-b)\,\Gamma(1+a)}$$

As $N \rightarrow \infty$,

$$\text{if } 1 + a - b < 0, \qquad G(N) \rightarrow \frac{1+a}{b-a-1}$$

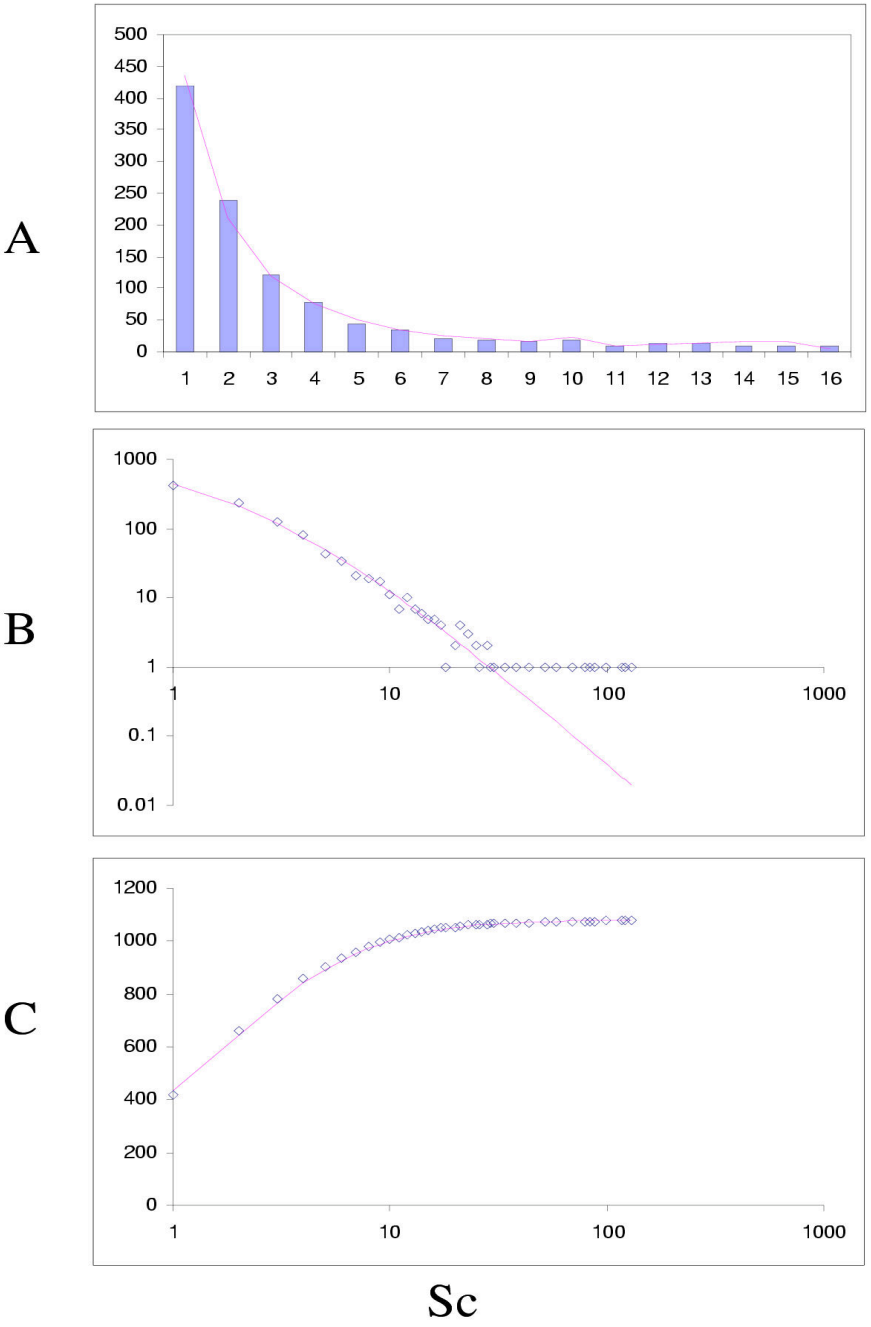
$$\text{if } 1 + a - b > 0, \qquad G(N) \rightarrow \infty.$$

Thus, the linear second-order balanced Birth Death and Innovation Model produces a power law distribution for gene family sizes that is characterized by birth and death rate parameters.

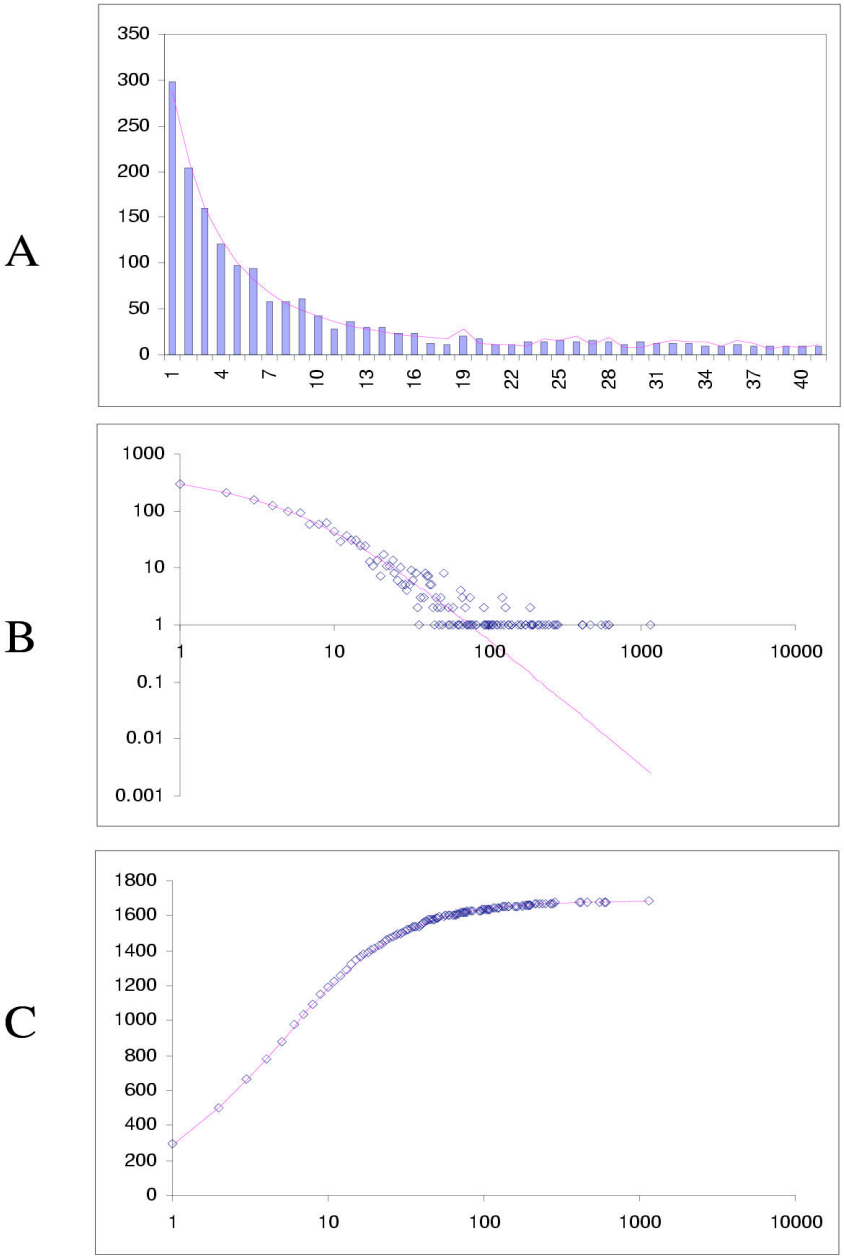
3.2 Simulations

Explicit details of the Birth, Death and Innovation Model are described by Georgy Karev and Yuri Wolf(*BMC Evol. Biology 2002)*. Included, are simulation results that test the goodness-of-fit of the BDIM to actual genomic data. Karev et. al. make these simulations by first using BLAST to categorize the genomes of 10 different organisms into bins of gene family sizes, namely Sc, *Saccharomyces cerevisiae*, Dm, *Drosophila melanogaster*, Ce, *Caenorhabditis elegans*, At, *Arabidopsis thaliana*, Hs*, Homo sapiens*, Thema, *Thermotoga maritima*, Metth*, Methanothermobacter thermoautotrophicum*, Sulso, *Sulfolobus solfataricus*, Bacsu, *Bacillus subtilis,* Eco, *Escherichia coli*. Next, he creates theoretical distributions for each genome based on appropriate parameters in the BDIM model, and statistically compares the predicted gene family sizes to the actual observed frequencies by the χ^2 test. The results by this analysis are convincing.

With all 10 genomes, the χ^2 value has a probability larger than .05 for the linear second order balanced BDIM; no significant differences between the model and observed genome data are found.



Fit of yeast *Saccharomyces cerevisiae* family size distributions to the second-order balanced linear BDIM. A. Distribution of the size of gene families. B. Gene family size distribution in double logarithmic coordinates. Magenta line: $f_i = 11521\_(i+1.55)/\_(i+4.27)$ C. Cumulative distribution of gene family size with prediction line from the second-order balanced linear BDIM(Karev and Wolf 2002, p. 16).



Fit of *Homo sapiens* gene family size distributions to the second-order balanced linear BDIM. A. Distribution of the size of gene families. B. Gene family size distribution in double logarithmic coordinates. Magenta line: f_i = 22030_(i+5.16)/_(i+7.43) C. Cumulative distribution of gene family size with prediction line from the second-order balanced linear BDIM (Karev and Wolf 2002,p.18).

Across species, the linear second-order balanced BDIM generates $a,b > 0$ such that $a < b$. Consequently, this implies that 1) selection relaxes as gene families increase, making it more likely for a gene to survive in a larger family than in a smaller family, and 2) a gene family's overall birth and death rate decrease as it gets larger in size. The first point naturally supports evolutionists believing in adaptive radiation, while the second provides mathematical evidence for observed faster rates of prokaryote evolution.

The fit of the BDIM reveals the existence of a complex relationship between the innovation rate and the birth rate. On average, the innovation rate appears to be approximately three orders of magnitude greater than the *per gene* birth rate, while simultaneously being several orders of magnitude lower than the *total family* birth rate. For smaller prokaryotic genomes, the differences are not as drastic, however, for larger eukaryotic genomes, the gradients reach several-folds in magnitude.

3.3 Discussion

The Birth, Death, and Innovation Model is important because it produces a power law distribution from realistic genetic mechanisms. It is a simple model that has impacted the field in three main ways.

Firstly, the BDIM model has shown that the distribution of gene family sizes obtains a unique equilibrium state exponentially fast. Unlike the Branching Process with Immigration model, the BDIM can accommodate all gene family sizes, large or small, by producing different power law distributions, based on specified parameters. Any change in birth, death, or innovation rates rapidly relaxes into a new stationary state. Eugene V. Koonin in "The Structure of the Protein Universe and Genome Evolution," states that this characteristic is essential in evaluating the model's realistic applications because it coincides with the well-recognized theory of punctuated equilibrium, in which paleontological records reveal brief bursts of genomic activity followed by long periods of stasis(220). Thus, the BDIM emphasizes the important role that gene birth(duplication), death and innovation(horizontal gene transfer) play in catalyzing evolution.

Secondly, the model has illustrated how a linear second-order balanced BDIM can describe a power law distribution by incorporating growth parameters that are dependent on the size of the gene family. Previously, models have been accepted that assign uniform growth rates across the entire genome. Observed distributions have shown that this assumption is not very realistic. In this regard, the BDIM has surpassed its rivals by accounting for diverse family sizes through

parameters that can be adjusted. Koonin confirms this alteration in parameters by stating that "the death rate approaches the birth rate for large families, but is considerably greater than the birth rate for small families (221)." This means that small families are more evolutionary dynamic than large families because they are forced to intensely compete against each other at birth and either quickly proliferate or perish. As with the Branching Process with Immigration model, this implies that large gene families are older than smaller gene families because they have successfully survived nature's pressures. Consequently, organisms with functionally similar large gene families can be said to have evolved from a common ancestor.

Thirdly, the BDIM has suggested that innovation be a critical component driving evolution. In the linear second-order balanced model, the rate of innovation is much higher than the *per gene* birth rate. This means that the likelihood of a new gene being created out of non-coding DNA is much lower than the probability of one being derived from the horizontal transfer, or recombination, of nucleotides on an existing chromosome. Because the mathematical equilibrium is precise, this correlation emphasizes the crucial importance that genetic innovation must have in "maintaining the balance (Karev and Wolf 2002, p. 22)."

Natural selection is another key factor in maintaining the steady state of the linear second-order balanced BDIM. While it is not incorporated directly into the mathematical construction of the model, it is implicitly recognized as the selective force that chooses the survival of certain genes over others, and regulates the magnitude of birth, death and innovation rates. There is a precise balance that must be met for the BDIM model to operate appropriately. However, in nature, complexities are not unusual.

Conclusion

Mathematical models can provide useful information about the interactions within gene families and the functions that genes control. In this paper, three possible models have been analyzed that offer potential explanations for the power law distribution of gene family sizes, namely the Preferential Attachment Model, the Branching Process with Immigration Model, and the Birth, Death and Innovation Model. Individually, all three models fail to provide complete explanations for the cause of the distribution, however, together, they serve to identify key components necessary for genomic evolution: preferential attachment, natural selection, gene birth (duplication), gene death, and gene transfer (innovation). Through the development of these essential ideas and the rapid advancements made in genetic mapping, it may soon be possible to discover the code of interactions that regulate the genomic universe.

References

Barabàsi, Albert-Làszlò and Rèka Albert. The Emergence of Scaling in Random Networks. *Nature*. 1999. v. 286, 509-12.

Durrett, Rick and Jason Schweinsberg. Approximating Selective Sweeps. http://www.math.cornell.edu/~durrett, 2003. *Theoretical Population Biology,* to appear.

Jeong, H., B. Tombor, R. Albert, Z.N. Oltvai and Albert-Làszlò Barabàsi. The Large-Scale Organization of Metabolic Networks. *Nature*. 2000. v. 407, 651-4.

Joyce, P. and S. Tavarè. Cycles, Permutations, and the Structure of the Yule Process With Immigrations. *Stochasitc Process Applications*. 1987, v. 25, 309-14.

Karev, Georgy, Yuri Wolf, Andrey Rzhetsky, Faina Berezovskaya and Eugene Koonin. Birth and Death of Protein Domains: A Simple Model of Evolution Explains Power Law Behaviour. *BMC Evolutionary Biology*. 2002, v. 2, 18-44.

Koonin, Eugene V, Yuri Wold & Georgy Karev. The Structure of the Protein Universe and Genome Evolution. *Nature*. 2002. v. 420, 218-23.

Qian, Jiang, Nicholas Luscombe and Mark Gerstein. Protein Family and Fold Occurrence in Genomes: Power-Law Behaviour and Evolutionary Model. *Journal of Molecular Biology*. 2001, v. 313, 673-81.

Rzhetsky, Andrey and Shawn M. Gomez. Birth of Scale-Free Molecular Networks and the Number of Distinct DNA and Protein Domains Per Genome. *Bioinformatics*. 2001. v. 17, 988-96.