VI International Conference on Computational Bioengineering
ICCB 2015
M. Cerrolaza and S.Oller (Eds)

STOCHASTIC ANALYSIS OF THE SIZE OF GENE FAMILIES IN THE GENOME AT ANY TIME OF THE EVOLUTIONARY PROCESS

EVA MARIA ORTEGA[*], JOSE ALONSO[†] AND XIAOHU LI[$]

[*] Residencia Universitaria Hnas. Oblatas Murcia, and UMH
Plaza Universidad, 30001 Murcia, Spain
e-mail: emariaop@gmail.com,

[†] Hospital Clinico Universitario Virgen de la Arrixaca
Crtra Madrid Cartagena, 30120 El Palmar Murcia, Spain

[$] School of Mathematics, Xiamen University
361005 Xiamen, China.

Key words: Size Protein families, hazard rate, reversed hazard rate, IGFR, DRPFR, bounds.

Abstract. Huynen and van Nimwegen 1998 proposed the stochastic multiplicative process, to model the size of gene families at time period t, $S_t = \alpha_1 \ldots \alpha_{t-1}\ \alpha_t$ where α_i are random multiplication environmental factors that are statistically independent at each time period, and identically distributed, assuming that at t = 0 a gene family is founded by a single ancestor, and that the duplications and the deletions are coherent with respect to the genes within one gene family. Based on this scheme, we analyze the probability distribution of the size of gene families at any time period using some results in Ortega and Li 2010,2015: we study the contraction pattern of the size of gene families using an ageing notion called Decreasing Proportional Failure Rate Property defined in Belzunce, Candel and Ruiz 1998, discuss the influence of the magnitude of the distribution of the environmental factors on the hazard rate and the reversed hazard rate of the size of gene families, and apply a stochastic bound as a quantitative measure of the expansion or the contraction of gene families in the genome. The assumptions on the distribution of the environmental factors are compatible with the lognormal model, that has been proved to fit well. The results are extended to positively correlated environmental factors by applying results in Ortega and Alonso 2014. We conclude that our results agree with the scientific evidence that if a certain gene is likely to duplicate then all the genes of its family are likely to duplicate in the genome, and that the increase in gene number with increasing biological complexity involves the expansion of families of closely related genes.

1 INTRODUCTION

The expansion or the contraction of gene families along a specific phylogenetic tree is determined by molecular processes that can be due to chance and result of natural selection. Diverse mathematical models have been proposed in the literature to describe the size of gene families in the complete genome and its probability distribution has been studied under

different biological hypothesis using statistical models and algorithms. In this paper, the Huynen and van Nimwegen (1998) model for the size of gene families at any time period t is analyzed using the results and the methods in Ortega and Li (2010,2015) and Ortega and Alonso (2014). Huynen and van Nimwegen (1998) proposed the stochastic multiplicative process, with random multiplication environmental factors that are statistically independent at each time period, and identically distributed to model the size of gene families at any time t. Let us start by the description of the size of gene families at any time period t. Assuming that at t = 0 a gene family is founded by a single ancestor, and that the duplications and the deletions are coherent with respect to the genes within one gene family, that is, if a certain gene is likely to duplicate then all the genes of its family are likely to duplicate in the genome; and analogously, if one gene is likely to be deleted then all genes of its family are as likely to be deleted from the genome. Through duplications and deletions the size of gene family will fluctuate with the possibility that the family eventually is going to extinct from the genome (in fact, no gene family lives forever in any particular genome unless other mechanisms prevent them from going extinct), and if the duplications and the deletions of genes are observed at any unit of time (observations times are unit times), then the size of gene families at time period t is $S_t = \alpha_1 \ldots \alpha_{t-1}\, \alpha_t$, with this formula that will be called Equation (1.1), where α_i $i=1,\ldots,t$ are random multiplication environmental factors drawn at each time step from a distribution function F of a random variable α with the assumptions that α_i , i = 1, ..., t are statistically independent at each period time of observation of the genome, and that Prob(α = a) is peaked around a = 1. Notice that the distribution of the environmental factor is compatible with the lognormal model. Our main objective is to provide a probabilistic approach with discussion for the random variable in Equation (1.1) under an arbitrary distribution of the environmental factor. The mathematical methods used in the paper are stochastic comparisons, bounds and ageing models for non-negative random variables, that can be seen in Shaked and Shanthikumar (2007). In Section 3, we study the contraction pattern of the size of gene families using an ageing notion called Decreasing Proportional Failure Rate Property which is defined in Belzunce et al. (1998). We also study the influence of the magnitude of the distribution of the environmental factors α_i , i = 1, ..., t on the hazard rate and the reversed hazard rate of the size of gene families by using some results in Ortega and Li (2015). In addition, we apply a numerical bound in Ortega and Li (2010,2015) for the risk function of the size of gene families at any time period, which provides a quantitative measure of the expansion or the contraction of gene families at any time of the evolutionary process. The results are extended to the case of positively correlated environmental factors by applying results in Ortega and Alonso (2014).

2 MATHEMATICAL METHODS AND PRELIMINARIES.

In this section we recall some concepts that constitute the mathematical instruments to analyze the random variable that represents the size of gene families at any time period. Consider an absolutely continuous lifetime (non-negative random variable) X, with probability density function f, cumulative distribution function F, risk function 1-F. The hazard rate r of X is defined for any x such that 1- F(x) > 0, by r(x) = f(x)/(1-F(x)). Observe that r(x) can be thought as the intensity of failure of a unit, with a random lifetime X, at time

x. The hazard rate, also called as failure rate or mortality rate, is a very known concept with many applications in probability and statistics, reliability, survival analysis, insurance and finance, and other research areas. As Shaked and Shanthikumar (2007) notices, the higher the hazard rate is, the smaller X should be stochastically.
The reversed hazard rate of X for any x such that $F(x) > 0$ is defined by $a(x) = f(x)/F(x)$ (see Keilson and Sumita (1982) and Shaked and Shanthikumar (2007)). From the reversed hazard rate, the random variable $[x - X|X \_ x]$ can be used to predict the exact times of occurrence of events because $a(x)dx$ represents the probability of failing in the interval $(x - dx, x)$, when a unit is found failed at time x. This concept is useful in casualty insurance, reliability, demography, epidemiology and medicine (forensic science) to predict times of occurrences of events. For example, the incubation times of diseases, i.e. durations from the infection until the disease occurrence, are difficult to measure because the infection time is unobserved in general. Some examples of prediction using this function are given by Keiding (1991).
The generalized failure rate of X is given by $gr(x) = xr(x)$ (see Belzunce et al. (1995,1998)). It measures at any income, the odds against advancing further to higher incomes in a proportional sense (it also represents the slope of the risk function of incomes in the Pareto diagram) (see Kleiber and Kotz (2003)). Belzunce et al. (1998) introduced the increasing proportional failure rate property, as a generalization of the increasing failure rate ageing notion, by requiring that $gr(x) = xr(x)$ be increasing. This notion was studied by Lariviere and Porteus (2001), who named it Increasing Generalized Failure Rate, denoted by IGFR.
The reversed proportional failure rate of X is defined for any $x > 0$ by $e(x) = x\, f(x)/F(x)$, also known as the elasticity function (see Dagum (1977)), where $f(x)/F(x)$ represents the reversed hazard rate at x. The decreasing reversed proportional failure rate property, denoted by DRPFR, is a related ageing notion that is characterized by an elasticity function $e(x) = x\, f(x)$ F(x) being decreasing, and was introduced in Belzunce et al. (1998). For recent probabilistic properties of the IGFR and the DRPFR notions, we refer to Ortega and Li (2010, 2015) and references therein.
Finally, we give the definitions of some stochastic comparisons that will be useful (see Shaked and Shanthikumar (2007)). Let X and Y be two absolutely continuous lifetimes, with cumulative distribution functions F_X and F_Y , hazard rates r_X and r_Y , and reversed hazard rates a_X and a_Y , respectively. X is said to be smaller than Y in:
i) the hazard rate order, denoted $X <_{hr} Y$, if $r_X(x) \geq r_Y(x)$ for all $x > 0$.
ii) the reversed hazard rate order, denoted $X <_{rhr} Y$, if for all $x > 0$, $a_X(x) \leq a_Y(x)$.

3. A DISTRIBUTIONAL PROPERTY OF THE SIZE OF GENE FAMILIES.

The following property is based on Theorem 3.3 in Ortega and Li (2015), that states the closure by products for the DRPFR notion that was pointed out by Badia (2010).
Property 3.1. Let α_i , i = 1, ..., t be independent DRPFR absolutely continuous non-negative random variables, then, $S_t = \alpha_1 \ldots \alpha_{t-1}\, \alpha_t$, is DRPFR.

The DRPFR reveals a pattern of contraction of the gene families at each time period to assess how the experimental researchers can act by influx of new genes and/or by reflective boundary conditions strongly against the deletion of the last gene within a gene family. We recall that a gene family of size one acts as a reflecting boundary for certain gene families

thus prevents them from going extinct, and alternatively, an occasional introduction of a gene from a new family into the genome, for instance by horizontal gene transfer, that is influx of new genes, may avoid that the family becomes extinct in the genome (see Huynen and van Nimwegen (1998)). Experimental studies proved that no gene family lives forever in any particular genome unless other mechanisms prevent them from going extinct. Let L denote the maximum value of the random variable of the size of gene families. The reversed hazard rate of the size of gene families is an infinitesimal as a function with domain given by the interval [0,L]. Under the assumptions of the Property 3.1, the proportional reversed failure rate of the size families converges, since it is decreasing and bounded. Using functional analysis, the product of an infinitesimal with a bounded function is also an infinitesimal, hence the proportional reversed failure rate of the size of gene families is an infinitesimal too.

4. STOCHASTIC COMPARISON OF THE SIZE OF GENE FAMILIES

From now on, we will denote $S_t = \alpha_1 \ldots \alpha_{t-1}\, \alpha_t$, and $S_t^* = \alpha^*_1 \ldots \alpha^*_{t-1}\, \alpha^*_t$, .
Next property follows from Theorem 4.2 in Ortega and Li (2015).

Property 3.2. Let $\{(\alpha_i, \alpha_i^*)|i = 1, ..., t\}$ be independent pairs of absolutely continuous IGFR non-negative random variables, such that $\alpha_i <_{hr} \alpha_i^*$, for i = 1, ..., t. Then, $S_t <_{hr} S_t^*$.

This property means that the probability of one gene deletion leading to the contraction of the family, given that the size of the family is larger than x, at any time period t, increases or decreases as at the time t = 1, i.e., at the beginning of the evolutionary process. This result agrees with the coherence assumption by Huynen and van Nimwegen (1998), which means that if a certain gene is likely to duplicate, then all the genes of its family are likely to duplicate in the genome; and if one gene is likely to be deleted, then all genes of its family are as likely to be deleted from the genome. Experimental studies show that in general, an increase in the number of large gene families is expected versus the number of small gene families as the number of genes in a genome becomes larger. This trend is supported by the Property 3.2 because for large gene families the deletion probabilities are decreasing at any future time periods. For Huynen and van Nimwegen (1998) this fact leads to competition between the gene families for space in the genome is effectively bigger in a smaller genome which leads to a smaller value of μ_a, that is, the larger competition in smaller genomes yields to shorter average lifetimes of gene families in these small genomes.
Another related property follows from Theorem 4.6 in Ortega and Li (2015).

Property 3.3. Let $\{(\alpha_i, \alpha_i^*)|i = 1, ..., t\}$ be independent pairs of absolutely continuous DRPFR non-negative random variables, such that $\alpha_i <_{rhr} \alpha_i^*$, for i = 1, ..., t. Then, $S_t <_{rhr} S_t^*$.

This property means that the probability of one gene deletion leading to the contraction of the family, given that the size of the family is smaller than x, at any time period t, increases or decreases as at the time t = 1, i.e., at the beginning of the evolutionary process.

5. STOCHASTIC BOUND OF THE SIZE OF GENE FAMILIES

For practical instances, we provide a lower bound for the risk function of the size of gene families at any time based on its distributions moments and Theorem 5.2 in Ortega and Li (2015). Since the majority of the proteins in the life come from a limited number of families, this bound is of interest in experimental studies.

Property 3.4. Let α be an IGFR absolutely continuous non-negative random variable with finite moments of the first three order, and α_i, i = 1, ..., t be independent and identically distributed as α, with $S_t = \alpha_1 \ldots \alpha_{t-1}\, \alpha_t$. Then, for any t, $x \geq 1$, with $\mu_k = E[S^k_t]$, for k = 1, 2, 3, $\Pr(S_t > x) \geq x^{[(-6(\mu_1)^3)/(2\mu_3 - 3\mu_1\mu_2 + (1+2\log(\mu_1))3\mu^3_1)]}$ for $x \leq \mu_1 \exp((2\mu_3 - 3\mu_1\mu_2 + 3\mu^3_1)/(6\mu^3_1))$, and $\Pr(S_t > x) \geq 0$, otherwise.

The power-law distribution is the limit distribution of a multiplicative stochastic process with a boundary constraint (see Sornette and Cont (1997)). If we assume that all the genes within a family are affected in the same (or at least a similar) way by the environment; and on the other hand, that in consecutive time periods, each gene family tends to expand or to shrink as a whole with a random factor α_t, then the distribution of the size of gene families in the complete genome is the result of many processes like Equation (1.1) occurring in parallel for large times t, together with the occasional introduction into the genome of a new gene family of size one. Equation (1.1) is equivalent to $\log(S_t) = \log(\alpha_1) + \ldots + \log(\alpha_t)$, and the limit for large times t gives that $\log(S_t)$ becomes normal distributed with average $\mu_t = \mu_\alpha t$ and $\sigma^2_\tau = \sigma^2_\alpha t$ by the central limit theorem. The assumptions $\mu_\alpha \leq 0$, and $\mu_\alpha / \sigma^2_\alpha < -1/2$, from Huynen and van Nimwegen respectively, allow to avoid the condition that gene families tend to become infinitely large in the limit of time going to infinity leaving the process in Equation (1.1) without a stable limit, and allow that although the size of a gene family may fluctuate for a very large time, eventually each gene family tends to become extinct in the genome) , thus these assumptions lead to conclude that the distribution of the size of gene families in the complete genome is power-law distributed. The bound given in Property 3.4 can be combined with those by the earlier asymptotic behaviour.

6. THE CASE OF POSITIVELY CORRELATED ENVIRONMENTAL FACTORS

For mathematical developments, Huynen and van Nimwegen (1998) have assumed that α_i are statistically independent random multiplication environmental factors. However it is known that an increase in the number of genes lead to an increase in the frequency of clusters of all sizes, and of the number of large clusters over the number of small cluster, and since the genes within one family have related functions, they are clustered in the genome, and they are affected by the environment in the same way at different time periods.

To finish, we deal with the dependence among environmental factors of the model in Equation (1.1) and we introduce a model for the size of gene families at any time period with positively correlated environmental factors that depend on random parameters.

Let $S_t = \alpha_1(\theta_1) \ldots \alpha_{t-1}(\theta_{t-1})\alpha_t(\theta_t)$, where $= \alpha_i(\theta_i)$, i = 1, ..., t are random multiplication environmental factors that are correlated by the positive quadrant dependence (see Shaked and Shanthikumar (2007)) of the random vector $(\theta_1, \ldots, \theta_t)$ denoted $(\theta_1, \ldots, \theta_t)$ is PQD and are

drawn from a distribution function $F_i(\theta_i)$, i=1,…,t. Given the random vector $(\theta_1^{'},...,\theta_t^{'})$, with independent components and the same marginal distributions than $(\theta_1, ...,\theta_t)$ being PQD, under the assumptions of the main result in Ortega and Alonso (2014), then $E[S_t(\theta_1, ...,\theta_t)]\geq E[S_t(\theta_1^{'},...,\theta_t^{'})]$. Using the earlier property and from Property 3.4, we get a similar bound for $E[S_t(\theta_1, ...,\theta_t)]$ in the case that the environmental factors are correlated. Notice that there are some non-parametrical statistical tests to check the PQD property.
The final conclusions can be summarized as follows. The behaviour of the random variable of the size of gene families at any time agrees with the scientific evidence that if a certain gene is likely to duplicate then all the genes of its family are likely to duplicate in the genome, and that the increase in gene number with increasing biological complexity involves the expansion of families of closely related genes.

Acknowledgements: Thanks are due to professors Felix Belzunce and German Badia for useful references in the main article where this contribution is based on.

REFERENCES

[1] Belzunce, F., Candel, J. and Ruiz, J.M. 1995. Ordering of truncated distributions through concentration curves. *Sankhya, Indian J Statist*, 57, 375-383.
[2] Belzunce, F., Candel, J. and Ruiz, J.M. 1998. Ordering and asymptotic properties of residual income distributions. *Sankhya, Indian J Statist*, 60, 331-348.
[3] Dagum, C. 1977. A new model of personal income distributions: specification and estimation. *Economie Appliquee*, 30, 413-437.
[4] Huynen, M.A. and van Nimwegen, E. 1998. The frequency distribution of Gene family sizes in complete genomes. *Molecular Biology Evolution*, 15, 583-589.
[5] Keilson, J. and Sumita, U. 1982. Uniform stochastic ordering and related inequalities. *Canadian J. Stat.*, vol. 10, 181198.
[6] Keiding, N. 1991. Age-specific incidence and prevalence: A statistical perspective (with discussion). *Journal of Royal Statistical Society A*, vol. 154, 371412.
[7] Kleiber, C., and Kotz, S. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences.* John Wiley and Sons, New Jersey.
[8] Ortega, EM. and Alonso, J. 2014. Recent issues on stochastic directional convexity, and new results on the analysis of systems for communication, information, time scales and maintenance. *Appl Stoch Mod Business Industry*, forthcoming, DOI: 10.1002/asmb.1989
[9] Ortega, EM, Li, X. 2010. New results on IGFR and DRPFR distributions and related issues. *Proceedings of XXXII National Conference on Statistics and Operations Research, SEIO2010*, A Coruna, Spain, pp.1-12.
[10] Ortega, E.M. and Li, X. 2015. Analytical results for IGFR and DRPFR distributions, with actuarial applications. *Comm. Stat. Theory Meth.*, approved for publication.
[11] Shaked, M. and Shanthikumar, J.G. 2007. *Stochastic Orders*. Springer, New York.
[12] Sornette, D. and Cont, R. 1997. Convergent Multiplicative Processes Repelled from Zero: Power Laws and Truncated Power Laws. *Journal of Physics I France*, 7, 431-444.