Generalized Hypergeometric Distributions Generated by Birth-Death Process in Bioinformatics

February 2, 2022

Vladimir A. Kuznetsov[1,2], Andre Grageda[1], Davood Farbod[3*]

[1] Department of Urology, Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University, Syracuse, New York, USA 13210. E-mail: kuznetsv@upstate.edu
[2] Bioinformatics Institute, A-STAR, Singapore, Singapore 138671.
[3] Department of Mathematics, Quchan University of Technology, Quchan 9477177870, Iran. E-mail: d.farbod@qiet.ac.ir

[*] Corresponding author: Davood Farbod

Running title: *Kuznetsov V.A., Grageda A., Farbod D., / Multi-parameter skewed distributions in bioinformatics*

Abstract. Modern high-throughput biological systems detection methods generate empirical frequency distributions (EFD) which exhibit complex forms and have long right-side tails. Such EFD are often observed in normal and pathological processes, of which the probabilistic properties are essential, but the underlying probability mechanisms are poorly understood. To better understand the probability mechanisms driving biological complexity and the pathological role of extreme values, we propose that the observed skewed discrete distributions are generated by non-linear transition rates of birth and death processes (BDPs). We introduce a (3d+1)-parameter Generalized Gaussian Hypergeometric Probability ((3d+1)-GHP) model with the probabilities defined by a stationary solution of generalized BDP (g-BDP) and represented by generalized hypergeometric series with regularly varying function properties. We study the Regularly Varying 3d-Parameter Generalized Gaussian Hypergeometric Probability (3d-RGHP) function's regular variation properties, asymptotically constant slow varying component, unimodality and upward/downward convexity which allows us to specify a family of 3d-RGHP models and study their analytical and numerical characteristics. The frequency distribution of unique mutations occurring in the human genome of patients with melanoma have been analyzed as an example application of our theory in bioinformatics. The results show that the

parameterized model not only fits the 'heavy tail' well, but also the entire EFD taken on the complete experimental outcome space. Our model provides a rigorous and flexible mathematical framework for analysis and application of skewed distributions generated by BDPs which often occur in bioinformatics and big data science.

KEYWORDS: Birth-Death Process, generalized hypergeometric distributions, stationary process, limiting process, regular variation, bioinformatics data

AMS SUBJECT CLASSIFICATION: 60E05, 60J28, 60J27, 62P10, 60G52, 92D99

1 Introduction

Today, high-throughput biological systems detection methods and bioinformatics generate enormous amounts of information ('big data') about characteristics of large-scale molecular and cellular random processes which drive the evolution, interaction, complexity and integrity of such systems. Such complex data sets are subject to statistical bioinformatics and mathematical modeling that analyze and extract data to generate knowledge. These information science disciplines change biology, other fields, and introduce new technological challenges. Development of big data in bioinformatics assist in solving fundamental challenges in biomedical disciplines and presents unimaginable prospects in data technology. Accordingly, novel complex and practically useful mathematical models need to be developed and applied.

The theories of probability and stochastic processes play important roles in the modeling and analysis of big data in bioinformatics. It has been observed that at the system level, molecular and cellular characteristics often range from zero to $10^4 - 10^7$ and their empirical frequency distribution (EFD) often exhibits skewed long-right side tail [24, 25, 26, 29] which is also over-dispersed [25, 26, 29]. The diverse EFD forms are observed in the normal and pathological processes, that probabilistic properties are essential, but underlie probability mechanisms are poorly understood.

The Birth-Death Process (BDP) with different functional forms of birth and death rates as the continuous-time and discrete states random process is one of the biologically reasonable models frequently used in bioinformatics and large-scale biomolecular data analysis. Importantly, the stationary distributions that are defined by the BDP could exhibit a skewed form having a power law-like right-side tail [25, 26, 27]. We have used linear BDP to model the gene expression level statistics that identified the skewed stationary probability distribution called Kolmogorov-Waring Distribution (KWD) and applied this model to data analysis of the Cancer Genome Anatomy Project (CGAP, NIH) which generated hundreds of sequencing and microarray data sets. The KWD model has been successfully quantified and applied to the analysis of Serial Analysis of Gene Expression (SAGE) and microarray data sets collected by CGAP and led us to the discovery of the power law-like frequency distribution at the gene expression level in the transcriptome of different bulk tissue samples (including cancer and normal tissues), stem cells, cell lines of many eukaryotes. The KWD not only proper fits well to such skewed EFDs with power law-like fat tail, but also has explained observed systematic changes of the frequency distribution form as a function of sample size and biomedical conditions.

Currently, there are many successful applications of BDP generated skewed frequency distribution models in diverse fields of bioinformatics and systems biology. For example, in an analysis of protein domain occurrences in proteins of a proteome of

organisms [25, 26, 28, 34], the number of expressed genes in the transcribed genome of cells [25, 26, 28], the DNA-protein binding avidity of the transcription factors (TFs) - proteins bound the specific DNA sites (BS) and regulate gene transcription [29, 32, 40], the number of R-loops formation sequences per gene region in the genome of an organism [41], the number of mutations in a gene of cancer cells [35].

Using skewed distributions according to the variety and diversity of bioinformatics data, two- and more parameter distribution models with linear birth and death rates as a function of the state of the process have been developed and used in big data analysis. Two-parameter Waring distribution [25, 26], three-parameter KWD [28, 30, 31], regularly varying hypergeometric frequency distributions with two, three and more parameters [2, 3, 4, 5, 6, 11, 13, 15] have been developed and applied to the analysis of the EFD generated by bioinformatics data sets.

In bioinformatics data analysis the multi-parameter probability models based on the BDP have been also developed [3, 6, 12, 21, 26, 28, 29, 30, 31, 34]. Kuznetsov has developed the generalized KWD (g-KWD) [26, 28, 29, 30, 31], applications of BDP with several parameters to protein domains evolution in proteins could be found in [21, 34], a spatial subclass of "multi-parameter regularly varying generalized hypergeometric distribution of the second type" has been introduced by Danielian et al.[12] whose considered the skewed distributions generated by BDP via analysis of bioinformatics data.

We specifically refer to the pioneer studies by Kemp [22] and Maki [33] who have used stationary Probability Distribution Function (s-PDF) generated by BDP to develop a basic theory of the multi-parameter distributions. Both scientists have considered the rational functions to model transition rates of integer argument $x \in \{0, 1, 2, ...\}$ [22, 33]; see [19, 29] for more references.

Referring to the previous studies [3, 6, 26, 28, 30, 31] Danielian et al.[12] have applied a regular variation theory to the skewed distributions generated by the BDP occurring via analysis of bioinformatics data, introduced multi-parameter regularly varying generalized hypergeometric distribution of the second type. This theory specified some mathematical properties of the EFD, called *statistical (empirical) facts*, which have been observed in bioinformatics data sets and are systematically reproducible in biomolecular systems. The facts follow skewness to the right, regularly varying at infinity, unimodality, stability by estimated parameter values, and convexity. These properties are also commonly observed in many other evolving complex systems. It was proposed that if a PDF (or probability law) satisfies these essential properties, then the PDF could properly be a mathematical framework for bioinformatics applications and statistical data analysis [3, 6]. However, more systematic, rigorous theory and data-driven applications need to be developed.

To better understand probabilistic mechanisms driving complexity and pathological role of extreme values in biosystems, we hypothesized that the observed skewed distribution of random event(s) varying at a large scale in these systems is a result of non-linear and mutually independent transition rate functions defined by several probabilistic mechanisms. We propose that the asymptotic theory of the generalized skewed distributions with extreme value tail generated by BDP provide a rigorous mathematical framework in a search for plausible probabilistic mechanisms essential for the biosystem's evolution and pathological regime. Our motivation is to select and develop mathematically adequate family of non-linear probability distributions by verifying the *statistical facts*, offer a family of the models in particular the stationary probability distributions with more than two parameters and non-linear transition rate

functions of state of the BDP. In this study, we propose a new *(3d+1)-Parameter Generalized Gaussian Hypergeometric Probability ((3d+1)-GHP)* distribution based on the hypergeometric series function and then provide some biological applications for it.

Note 1. Danielian et al.[12] have studied a practically meaningful subfamily of Generalized Hypergeometric Probability Distribution [19, 22] called by the authors the Generalized Hypergeometric Distribution of the Second Type. In this paper we consider a subfamily of (3d+1)-GHP as *Regularly Varying 3d-Parameter Generalized Gaussian Hypergeometric Probability Distribution (3d-RGHP).*

The paper is formed as follows. In section 2 introduces the (3d+1)-GHP model as a stationary distribution of generalized BDP (g-BDP). In section 3 we prove the *statistical facts* for the subfamily 3d-RGHP as a model proper for bioinformatics data sets and biomedical systems. Section 4 presents numerical examples of PDF of the 3d-RGHP, and an analysis of the PDF in the log-log plot. Section 5 provides results of the (3d+1)-GHP and 3d-RGHP fitting to the EFD of unique mutations that occurred in melanoma genes genome-wide. Section 6 provides a discussion and conclusion.

2 Generalized BDP with (3d+1)-GHP

Suppose that a random process is a family $\mathbf{X} = \{X(t), t \in T\}$ of random variable values $X = i,\ i = \{0, 1, 2, ..\}$ such that the process is a Markov process with continuous time $t \in T \geq 0$ and countable number of states $X = i$ where $i \in S = \{0, 1, 2, ...\}$.

Continuous time BDP is a practically applicable subset of the Markov chains models [16, 23]. The BDP $\{X(t), t \geq 0\}$ is described as a random time movement of the object between neighbor states with non-negative and finite transition rates. BDP model allows change state only for one step for short time interval $(t, t + \delta t]$: $E_i \to E_{i+1}$ as well as $E_i \to E_{i-1}$ if $i \geq 1$. If $i = 0$, the theory allows $E_0 \to E_1$. Assume that if the process at time t is in E_i, then during $(t, t + \delta t]$ the transitions $E_i \to E_{i+1}$ have probability $\lambda_i(t)\delta t + o(\delta t)$, $E_i \to E_{i-1}$ have probability $\mu_i(t) + o(\delta t)$, probability no change E_i in state i be $1 - (\lambda_i(t) + \mu_i(t))\delta t + o(\delta t)$ and a probability of more than one state change occurs equals $o(\delta t)$. $o(h)$ means that $\lim_{h \longrightarrow 0} \frac{o(h)}{h} = 0$, called negligible with respect to h.

Let $\{p_i(t) = P(X(t) = i)\}$, where $i \in S = \{i = 0, 1, 2...\}$, denote the probability mass function (PMF) at time $t \in [0, \infty)$ generated by the stochastic process $\{X(t), t \geq 0\}$. According to the theory, at each infinitesimal time interval $(t, t + \delta t]$ where $\delta t \to 0$ the transition functions are differentiable functions and we can calculate birth and death transition rates

$$\lambda_i(t) = \lim_{\delta t \longrightarrow 0+} \frac{P(X(t+\delta t = i+1)|X(t)=i)}{\delta t} \tag{1}$$

$$\mu_i(t) = \lim_{\delta t \longrightarrow 0+} \frac{P(X(t-\delta t = i-1)|X(t)=i)}{\delta t} \tag{2}$$

where $\infty > \lambda_i(t) > 0$ for $i \geq 0$, $\infty > \mu_i(t) > 0$, $i \geq 1$, $\infty > \mu_0(t) \geq 0$.

In this section, we generalize the BDP model, considering the transition rates as non-linear multi-parameter functions of states. According to basic assumptions of the BDP model, when at time t a state of a population is i, the probability of a single birth occurring during the short time interval $(t, \delta t)$ is defined by the conditional probabilities (1-2) which does not depend on time interval δt and for also at $\delta t \longrightarrow 0$

$$\begin{cases} P_{i,i+1}(\delta t) = \lambda_i(t)\delta t + o(t), \quad P_{i,i-1}(\delta t) = \mu_i(t)\delta t + o(t), \ i = 0, 1, 2, ..., \\ \\ P_{i,j}(\delta t) = o(t), \ |i-j| > 1, \end{cases} \tag{3}$$

and we conclude that

$$P_{ii}(\delta t) = 1 - (\lambda_i(t) + \mu_i(t))\delta t + o(t) \tag{4}$$

as $\delta t \longrightarrow 0$ and

$$\infty > \lambda_i(t) > 0, \ \infty > \mu_i(t) > 0, i = 1, 2, ...; \ \mu_0(t) \geq 0. \tag{5}$$

Using (1-2) and (3-4) and assuming a differentiation property of the probability functions $p_i(t), i = 0, 1, 2, ...$, the total changes of each probability in the BDPs, we have Kolmogorov's BDP differential difference equations at $\delta t \longrightarrow 0$ [16], [23]:

$$\frac{dp_0(t)}{dt} = -\lambda_0(t)p_0(t) + \mu_1(t)p_1(t) \tag{6}$$

$$\frac{dp_i(t)}{dt} = -(\lambda_i(t) + \mu_i(t))p_i(t) + \lambda_{i-1}(t)p_{i-1}(t) + \mu_{i+1}(t)p_{i+1}(t) \tag{7}$$

where $i = 1, 2, ...$ We also assume that

$$0 < \lambda_i(t) < \infty, \ i = 0, 1, 2, ...; \ 0 < \mu_i(t) < \infty, \ i = 1, 2, ... \tag{8}$$

The BDP defined by (6-7) at any time is the following conditions

$$\sum_{i=0}^{\infty} p_i(t) = 1 \tag{9}$$

and other two Kolmogorov's axioms [23, 37]. Also, according to Markov chain property the probability measure of the BDP is a right-continuous step-function, suggesting its 'regularity' property.

Additional specification of the properties of the BDP allows us to define the distribution laws with potential applications.

Let us consider long-time regimes of the BDP related to stationary and limiting characteristics and the corresponding probability functions [9, 16].

Let $n(t_1, t_2)$ be a random number of events of the process within time interval $[t_1, t_2]$ at time $s \in [0, t]$. The BDP is called stationary (or time-homogeneous) if the probability distributions of r.v.'s $n(t_1, t_2)$ and $n(t_1 + s, t_2 + s)$ are coincident $P(n(s, s + t) = i) = P(n(0, t) = i)$ where $i = 1, 2, ...$ for $s \in [0, t]$. We will consider the stationary process BDP with non-zero limiting distribution. This proposes that all transition rate functions of the process are positive and there are no adsorption states, in particular we put $\mu_0(t) = 0$.

Let $\hat{p}_i,\ (i = 0, 1, 2, ...)$ be a probability of stationary process in state i and $\bar{\lambda}_i, \bar{\mu}_i$ be transition rates of the process in this state.

Definition 1. A PDF of random variable $X = i, i \in S = \{0, 1, 2, ...\}$ is a s-PDF of the process $\{X(t), t \in [0, \infty)\}$ given by (6-7) if the probabilities $\hat{p}_i,\ (i = 0, 1, 2, ...)$ are defined by the non-trivial solution the system of algebraic equations

$$\bar{\lambda}_0 \hat{p}_0 = \bar{\mu}_1 \hat{p}_1 \tag{10}$$

$$(\bar{\lambda}_i + \bar{\mu}_i)\hat{p}_i = \bar{\lambda}_{i-1}\hat{p}_{i-1} + \bar{\mu}_{i+1}\hat{p}_{i+1} \tag{11}$$

where $i = 1, 2, ...$ These equations provide a singular point solution of the system of differential-deference BDP equations at

$$\frac{dp_i(t)}{dt} = 0. \tag{12}$$

Note that probabilities of X of PDF is a discrete distribution function which also called a PMF.

Thus, we have

$$\hat{p}_i = \hat{p}_0 \cdot \prod_{k=1}^{i} \frac{\bar{\lambda}_{k-1}}{\bar{\mu}_k}, \tag{13}$$

and

$$\hat{p}_0 = \Big(1 + \sum_{i=1}^{\infty} \prod_{k=1}^{i} \frac{\bar{\lambda}_{k-1}}{\bar{\mu}_k}\Big)^{-1}, \tag{14}$$

where $0 < \hat{p}_0 < 1,\ 0 < \hat{p}_i < 1,\ i = 1, 2, ...$

Note that

$$Q = \sum_{x=1}^{\infty} \prod_{i=1}^{x} \tau_i < \infty, \tag{15}$$

where $x = 1, 2, ...,\ \tau_i = \frac{\bar{\lambda}_{i-1}}{\bar{\mu}_i} \leq \nu < 1$ provides well-known necessary and sufficient conditions for the existence of the non-trivial stationary solution Eqs.(13-14).

To characterize asymptotic properties of the probability distribution in time, the time of the random process after its initiation can be discriminated on two intervals: $[0, t_n)$ called time of the transition phase (non-stationary process phase) and next time interval $[t_n, \infty)$, which refers to the long-time behavior, is called limiting or steady state phase.

We consider long-time asymptotic solution of BDP (6-7) allows determining this solution using

$$\frac{dp_i(t)}{dt} \asymp 0, \tag{16}$$

at $t \longrightarrow \infty$ in states $i = 0, 1, 2, ...$ The solution is the PMF that is also named steady-state.

The probabilities of PMF in the steady-states we denote by $p_i^*,\ i = 0, 1, 2, , ...$ In this study we assume that in the $\lim_{t\to\infty}\ p_0(t) > 0$.

Let us define the non-zero limiting probability function of (6-7) which takes r.v. $X(t)$ at $t \longrightarrow \infty$:

$$\lim_{t\longrightarrow\infty} P(X(t) = i) = \lim_{t\longrightarrow\infty} p_i(t) = p_i^*$$

Let i be an initial state and j be a limiting in time state of a trajectory given by BDP (6-7) where $\{i, j\} \in S = \{0, 1, 2, ..., \}, i \neq j$, stationary PMF exists and is defined by the probabilities $\hat{p}_i,\ i \in S = \{0, 1, 2, ..., \}$.

Definition 2. The PMF is the limiting PMF (defined by probabilities $p_i^*,\ i = 0, 1, 2, ...,$) of the BDP if, for any transition from initial state i to a state j, the transition rates are defined by transition rates of the stationary distribution.

Based on the axioms of a probability and a definition of the Markov chain, we assume that in the $\sum_{i=0}^{\infty} p_i^* = \sum_{i=0}^{\infty} p_i(t) = 1, t \in T = [0, \infty), i \in S = \{0, 1, 2..., \}$.

Note that the expression (15) is applicable for calculation of the limiting distribution after certain period of 'transition phase' of the process when starting from some $i = i_j$ at time $t > t_c > 0$ the condition $\tau \leq \nu_i < 1$ happen for all $i > i_j$ (namely, in the right tail of the probability function).

The BDP $\{X(t), t \geq 0\}$ is an ergodic Markov chain if it is possible to go from every state to every state. This proposes that for any initial distribution of state probabilities $p_i(t = 0), i \in S = \{0, 1, 2, ..., \}$ the transition rates are positive and the limiting probability distribution representing by p_i^* of random variable i exists due to stationary distribution exists. The conditions $\frac{dp_i(t)}{dt} \asymp 0, (i = 0, 1, 2, ...)$ assume the asymptotic converge of $p_i(t)$ to $\hat{p}_i$ when $t \longrightarrow \infty$.

Thus, we have

$$\lim_{t\longrightarrow\infty} p_i(t) = p_i^* \ \longrightarrow \hat{p}_i = \hat{p}_0 \prod_{s=1}^{i} \tau_s > 0 \tag{17}$$

This suggests a long-time asymptotic converge of the BDP equations solutions to an isolated (stationary) solution.

Thus, the long-time limiting distribution is independent of the initial state while the stationary distribution is determined by the unique initial state distribution (it is a non-trivial singular solution of the BDP equations). The long-time limiting distribution is an asymptotic distribution, while the stationary distribution a special initial state distribution.

Now, having the above definitions and assumptions, we consider that the birth and death intensity rates between states may be approximated by polynomial or rational functions with integer argument $x \in \{0, 1, 2, ...\}$ [20, 22, 26, 28].

Let as consider a stationary BDP which birth and death rates given by quadratic functions of state $x = \{0, 1, 2, ...\}$ given with three positive parameters sets $\{a, b, \theta_\lambda\}$ for birth rate and two parameters for death rate functions $\{c, \theta_\mu\}$ respectively. We consider these functions as the following

$$\begin{cases} \bar{\lambda}_x = \theta_\lambda (x + a)(x + b), \\ \bar{\mu}_x = \theta_\mu (x(x + c)). \end{cases} \tag{18}$$

Below we will show that (18) allow to define the solution of stationary BDP in Gaussian hypergeometric series (GHS) terms [19, 17].

The GHS is a series $\sum_{n=0}^{\infty} f_n(z)$ such that $\frac{f_{n+1}}{f_n}$ is a rational function of n [19, 17]. On factorization of the two polynomials in degree n, we obtain

$$\frac{f_{n+1}}{f_n} = \frac{(n + a_1)(n + a_2)...(n + a_p)}{(n + b_1)(n + b_2)...(n + b_q)(n + 1)} z \tag{19}$$

where z is the argument of the function, which may be equal to one (monic), or may be not. The factor $(n + 1)$ may be result in factorization or may be not. Decomposition of (19) into rational form is depending on $p + q$ constants and argument z.

For $p = 2$ and $q = 1$, $z = \theta$ where θ is a positive real value, we have the GHS

$$\sum_{n=0}^{\infty} f_n(\theta) = f_0 \sum_{n=0}^{\infty} \frac{\alpha^{[n]} \beta^{[n]}}{\gamma^{[n]}\ n!} \theta^n =: f_0\ {}_2F_1(\alpha, \beta; \gamma; \theta) \tag{20}$$

where $n \in \{0, 1, 2, ..., \}$. The terms of series are characterized by parameters α, β, γ of Pochhammer's symbols and series argument θ. Pochhammer's symbol is standard ascending factorials notation $a^{[n]}$ that denotes $a^{[n]} = a(a + 1)...(a + n - 1) = \frac{\Gamma(a+n)}{\Gamma(a)}$, $1^{[n]} = n!$. In the literature, the (20) are associated with the families of Generalized Hypergeometric Probability (GHP) [19, 22, 36].

It is straightforward to show that using the expression (13-14) and the birth and death rates given by (18), the probabilities $\hat{p_x}$ are converted to the terms power series function of an argument θ with exponent parameter $x \in \{0, 1, 2, ...\}$

and the coefficients defined as the rational functions given by $a^{[x]}, b^{[x]}$ and $(c+1)^{[x]}$

$$\hat{p}_x = \hat{p}_0 \prod_{i=1}^{x} \frac{\bar{\lambda}_{i-1}}{\bar{\mu}_i} = \hat{p}_0 \cdot \prod_{i=1}^{x} \frac{(i+a-1)(i+b-1)}{i(i+c)}\theta = \hat{p}_0 \cdot \frac{a^{[x]}b^{[x]}}{x!(c+1)^{[x]}}\theta^x \quad (21)$$

where $\theta = \frac{\theta_\lambda}{\theta_\mu}$ and $0 < \hat{p}_0 < 1, 0 < \hat{p}_x < 1,\ x = 1, 2, ...$

Finally, putting $a = \alpha$, $b = \beta$, $(c+1) = \gamma$ in the (21), we obtain the probabilities $p_i, i \in S = \{0, 1, 2, ...\}$ in the forms of GHS terms and $\hat{p}_0 = \frac{1}{f_0}$

$$\hat{p}_0 = \Big(1 + \sum_{x=1}^{\infty} \frac{a^{[x]}b^{[x]}}{(c+1)^{[x]}\ x!}\theta^x\Big)^{-1} = \frac{1}{{}_2F_1(a, b; c+1; \theta)} > 0. \quad (22)$$

Note, at $b = 1$ or $a = 1$ the probabilities (21-22) are the probabilities of KWD with linear birth and death rate functions of $X = x$ [25, 26], and if $b \neq 1$ or $a \neq 1$ the probabilities (21-22) could be considered as a specific case of g-KWD where the birth and death rate functions are polynomials of x with degree 2.

Corollary 1. The probabilities $\hat{p}_x, x = 0, 1, 2, ...$ of s-BDP with transition rates (18) are defined by GHS terms and calculated by the following

$$\frac{\hat{p}_{x+1}}{\hat{p}_x} = \theta \frac{(x+a)(x+b)}{(x+1+c)(x+1)} \quad (23)$$

at $\hat{p}_0$ defined by (22).

At $0 < \hat{p}_x < 1, x = 1, 2, ...,$ where the parameters (a, b, c) are defined by the properties of the birth and death rate functions (18), $(0 < a < \infty,\ 0 < b < \infty,\ 0 < c < \infty)$. Referring to (23) and due to $0 < \hat{p}_x < 1, x = 0, 1, 2, ...,$

$$\lim_{x \longrightarrow \infty} \frac{\hat{p}_{x+1}}{\hat{p}_x} = \theta \quad (24)$$

where the values of the parameter θ should be positive and limited by interval $(0, 1]$.

In this study we consider a generalization of BDP with birth and death rate functions defined as a finite product of the functions given by quadratic birth and death rate functions of $X = x$ in form (18) which parameters may have different non-negative values. We introduce the generalized form of the transition rate functions as follows

$$\begin{cases} \bar{\lambda}_x = \theta_\lambda^d \Big((x+a_1)(x+a_2)...(x+a_d)\Big)\Big((x+b_1)(x+b_2)...(x+b_d)\Big), \\ \bar{\mu}_x = \theta_\mu^d \Big(x^d\Big)\Big((x+c_1)(x+c_2)...(x+c_d)\Big), \end{cases} \quad (25)$$

where $\bar{\lambda}_x$ and $\bar{\mu}_x$ are the birth and death functions of random variable $X = x$ respectively, where $x \in S = \{1, 2, ..., x, ...\}$ and X the is a countable random

variable. a_j, b_j, c_j are the parameters $0 < a_j < \infty,\ 0 < b_j < \infty,\ 0 < c_j < \infty$ where $j = 1, 2, ..., d$ are the elements of the 3d-parameter set $V = \{a_j, b_j, c_j\}$. $\theta_\lambda = const > 0$, $\theta_\mu = const > 0$; θ_λ^d and θ_μ^d are the highest powers of x in the polynomials with degree d. Let denote $\theta = \frac{\theta_\lambda}{\theta_\mu}$.

Using (13-14) and (25) we can present the probabilities $\{\hat{p}_x\}$ with (3d + 1) unknown parameters of the birth and death rate functions as follows

$$\hat{p}_x = p_x(\alpha) = P(X = x),\ \ x = 0, 1, 2, ...,\ \ \alpha = \{a_j, b_j, c_j, \theta\},\ j = 1, 2, ..., d \quad (26)$$

takes the following form

$$\begin{cases} \hat{p}_x(\alpha) = \hat{p}_0(\alpha) \prod_{k=1}^{x} \frac{\bar{\lambda}_{k-1}}{\bar{\mu}_k} = p_0(\alpha) \cdot \theta^{xd} \prod_{k=1}^{x} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k+a_j-1)(k+b_j-1)}{(k+c_j)} \\ \hat{p}_0(\alpha) = \left[1 + \sum_{x=1}^{\infty} \theta^{xd} \prod_{k=1}^{x} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k+a_j-1)(k+b_j-1)}{(k+c_j)}\right]^{-1} \end{cases} \quad (27)$$

where $x \in S = \{1, 2, ..., x, ...\}$. $\alpha = \{a_j, b_j, c_j, \theta\}$ is a set the unknown parameters such that $0 < a_j < \infty,\ 0 < b_j < \infty,\ 0 < c_j < \infty$, where $j = 1, 2, ..., d$, $0 < \theta \leq 1$. In general case, d could be an unknown parameter of the model.

Note, $\hat{p}_0(\alpha)$ is the probability normalization factor providing $\sum_{x=0}^{\infty} \hat{p}_x(\alpha) = 1$. $\hat{p}_0(\alpha) > 0$ provides a condition of an existence of the s-PDF. At $\hat{p}_0 > 0$ the stationary distribution existence is well-defined by convergence of Gaussian hypergeometric series. $\hat{p}_0$ is a function of all parameters and could be interpreted as a probability of non-observed events in experiment outcome space, that is defined by parameters of the distribution.

Having (27), we can obtain the stationary probability measure $F_x(\alpha) = P(X \leq x, \alpha)$ (also called as the distribution of random variable X)

$$F_x(\alpha) = \hat{p}_0(\alpha) \sum_{s=0}^{x} \theta^{sd} \prod_{k=1}^{s} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k + a_j - 1)(k + b_j - 1)}{(k + c_j)} \quad (28)$$

Using Pochhammer's symbol we obtain (27) in the form

$$\begin{cases} \hat{p}_x(\alpha) = \hat{p}_0(\alpha) \theta^{xd} \prod_{k=1}^{x} \frac{1}{k^d} \cdot \prod_{j=1}^{d} \frac{a_j^{[x]} b_j^{[x]}}{(c_j+1)^{[x]}} \\ \hat{p}_0(\alpha) = \left[1 + \sum_{x=1}^{\infty} \theta^{xd} \prod_{k=1}^{x} \frac{1}{k^d} \cdot \prod_{j=1}^{d} \frac{a_j^{[x]} b_j^{[x]}}{(c_j+1)^{[x]}}\right]^{-1} \end{cases} \quad (29)$$

Note, $\hat{p}_x, x = 0, 1, 2, ...$ in (27) (as well as (29)) could be presented by generalized hypergeometric series [17, 19].

This follows from recursive form of state probabilities (27) as

$$\frac{\hat{p}_{x+1}(\alpha)}{\hat{p}_x(\alpha)} = \theta \frac{(a_1+x)(a_2+x)...(a_d+x)(b_1+x)(b_2+x)...(b_d+x)}{(1+x)^d(c_1+x+1)(c_2+x+1)...(c_d+x+1)}. \tag{30}$$

Corollary 2. The g-BDP defined by birth and death rate functions given by (25) generates the stationary probabilities (27) and a probability measure (28) called here the (3d+1)-GHP.

The $\{\hat{p}_x(\alpha)\}$ at 0, $d = 1, 2, ...$, provides a broad family of s-PDF (27) with the functionally diverse properties of the birth and death transition rate functions (25).

Note that (30) could be used to study asymptotic properties of the (3d+1)-GHP, specifically in the case when both birth and death rate functions are growing when an argument increases $x \to \infty$.

In this case

$$\lim_{x \longrightarrow \infty} \frac{\hat{p}_{x+1}}{\hat{p}_x} = \theta. \tag{31}$$

Corollary 3. Let us consider (30). If $\theta > 1$ then a stationary BDP process not exists; if $0 < \theta < 1$ then the PMF tail is unimodal, strictly decreases and the tail becomes longer if $\theta \to 1-$; if $\theta = 1$ the existence and asymptotic properties of the stationary distribution is defined by the relationships of the parameters $\{a_j, b_j, c_j\}, j = 1, 2, ..., d$ (see the next section).

Let $x \in S = \{1, 2, ...\}$ and consider an asymptotic property of birth and death rate functions defined by (25) at $x \to \infty$. We have

$$\lim_{x \longrightarrow \infty} \bar{\lambda}_x = \infty, \lim_{x \longrightarrow \infty} \bar{\mu}_x = \infty, \lim_{x \longrightarrow \infty} \frac{\bar{\lambda}_{x+1}}{\bar{\lambda}_x} = 1, \lim_{x \longrightarrow \infty} \frac{\bar{\mu}_{x+1}}{\bar{\mu}_x} = 1, \lim_{x \longrightarrow \infty} \frac{\bar{\lambda}_x}{\bar{\mu}_x} = \theta^d. \tag{32}$$

The last expression is most interesting because it shows that when the rates of the both BDPs increase at $x \to \infty$ their ratio is positive constant and if the $0 < \theta < 1$ than the rate of death will be larger than the rate of birth. Thus, the stationary BDP exists; furthermore, the PMF tail becomes shorter if d be larger.

Let propose that PMF given by (27) exists and PMF unique at $\theta = 1$. In this case $\lim_{x \longrightarrow \infty} \frac{\bar{\lambda}_x}{\bar{\mu}_x} = 1$ and at the fixed other parameters the PMF tail be the longest. Thus, our model gives us a characterization of 'moderate growth' for birth and death transition functions at asymptotic increasing the of $X = x$. Our previous studies have shown that a parameterization of the PFDs using skewed EFDs often provides $\theta = 1 - \epsilon,\ 0 \leq \epsilon << 1$, referring to the case of most long tail of PDFs.

Now, we can consider the (27) to identify the relationships between parameters of the transition rate functions satisfying unimodality, convexity, regularity, skewness - the important mathematical characteristics commonly observed in

large scale data sets of bioinformatics. We will to consider several basic asymptotic properties of our model in the case $\theta = 1$.

3 Extracting a regularly varying distribution

Let us have the following definitions.

Definition 3 ([12]). The distribution $p_x(\alpha)$ varies regularly at infinity with exponent $(-\rho)$ if it can be stated as

$$p_x(\alpha) = x^{-\rho} L(x)(1 + o(1)), \quad x \longrightarrow \infty, \tag{33}$$

where $L(x) > 0$ for $x = 1, 2, ...$, and for $s = 2, 3, ...$, $\lim_{x \longrightarrow \infty} \frac{L(sx)}{L(x)} = s^{-\rho}$.

Definition 4 ([12]). Assume that, for $s = 2, 3, ...$, the limit exists

$$\lim_{x \longrightarrow \infty} \frac{L(sx)}{L(x)} = 1 \tag{34}$$

then we say $p_x(\alpha)$ shows the asymptotically constant slowly varying component if in the formula (34)

$$\lim_{x \longrightarrow \infty} L(x) = L \in \mathbb{R}^+.$$

Now, we propose the following theorem.

Theorem 1. *The model (3d+1)-GHP* (27) varies regularly at infinity if and only if $\theta = 1$ and the exponent of the regular variation of $\hat{p}_x(\alpha)$ is equal to

$$-\rho = -\sum_{j=1}^{d} (c_j + 2 - a_j - b_j) < -1 \tag{35}$$

Meanwhile, $\hat{p}_x(\alpha)$ shows the asymptotically constant slowly varying component as

$$L = \hat{p}_0(\alpha) \cdot \frac{\Gamma(c_1 + 1)...\Gamma(c_d + 1)}{\big(\Gamma(a_1)...\Gamma(a_d)\big)\big(\Gamma(b_1)...\Gamma(b_d)\big)} \in \mathbb{R}^+. \tag{36}$$

Proof. By using formula

$$\prod_{k=1}^{x} \frac{k}{q+k} \approx \frac{\Gamma(q+1)}{x^q}, \quad x \longrightarrow \infty, \quad q \in \mathbb{R}^+,$$

we obtain an approximation as follows (for $j = 1, 2, ..., d$):

$$\hat{p}_x(\alpha) \approx \hat{p}_0(\alpha) \cdot \theta^{xd}$$

$$\times \left(\left(\frac{\Gamma(c_1+1)}{x^{c_1}} \cdots \frac{\Gamma(c_d+1)}{x^{c_d}} \right) \times \left(\frac{x^{a_1-1}}{\Gamma(a_1)} \cdots \frac{x^{a_d-1}}{\Gamma(a_d)} \right) \times \left(\frac{x^{b_1-1}}{\Gamma(b_1)} \cdots \frac{x^{b_d-1}}{\Gamma(b_d)} \right) \times 1^d \right) \tag{37}$$

$$= \hat{p}_0(\alpha) \cdot \theta^{xd} \frac{\Gamma(c_1+1)\ldots\Gamma(c_d+1)}{\big(\Gamma(a_1)\ldots\Gamma(a_d)\big)\big(\Gamma(b_1)\ldots\Gamma(b_d)\big)} \cdot x^{-\sum_{j=1}^{d}(c_j+2-a_j-b_j)}$$

when $x \longrightarrow \infty$ and based on (33) we have $\rho = \sum_{j=1}^{d}(c_j + 2 - a_j - b_j)$ such that $(-\rho)$ in (35) is the exponent of regular variation of $\hat{p}_x(\alpha)$. From (37) we turns out $\hat{p}_x(\alpha)$ in (27) varies regularly at infinity if and only if $\theta = 1$.

Additionally, by a limit $L = \lim_{x \longrightarrow \infty} L(x)$, we consider asymptotically constant slowly varying component (if exists). From (37) we see that L presented in (36) is an asymptotically constant slowly varying component for $\hat{p}_x(\alpha)$. The Theorem 1 is proved. □

Remark 1. Let, in the (3d+1)-GHP model (27), $\theta = 1$. Then, the exponent $(-\rho) = -\sum_{j=1}^{d}(c_j + 2 - a_j - b_j)$ of regular variation holds the condition $\rho > 1$.

Hence, now based on the Theorem 1 and from (3d+1)-GHP distribution family (27) we can extract 3d-RGHP model with the following PMF:

$$\begin{cases} \hat{p}_x(\alpha) = \hat{p}_0(\alpha) \cdot \prod_{k=1}^{x} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k+a_j-1)(k+b_j-1)}{(k+c_j)}, \\ \\ \hat{p}_0(\alpha) = \left[1 + \sum_{x=1}^{\infty} \prod_{k=1}^{x} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k+a_j-1)(k+b_j-1)}{(k+c_j)} \right]^{-1} \end{cases} \tag{38}$$

where $\hat{p}_0(\alpha)$ is the normalization factor, $x = 0, 1, 2, ...,$ $\alpha = (a_j, b_j, c_j)$, $j = 1, 2, ..., d$, is the unknown parameter such that $0 < a_j < \infty$, $0 < b_j < \infty$, $0 < c_j < \infty$, $j = 1, ..., d$. Also $a_1, ..., a_d$ and $b_1, ..., b_d$ are called numerator parameters and $c_1, ..., c_d$ denominator parameters. We call the model (38) as 3d-RGHP.

Having (28), the stationary probability measure $F_x(\alpha) = P(X \leq x, \alpha)$ at $\theta = 1$ is the following

$$F_x(\alpha) = \hat{p}_0(\alpha) \sum_{s=0}^{x} \prod_{k=1}^{s} \frac{1}{k^d} \prod_{j=1}^{d} \frac{(k+a_j-1)(k+b_j-1)}{(k+c_j)}. \tag{39}$$

We call (39) as 3d-RGHP. Its PMF is written in the form (38)

Remark 2. Assume that, in the PMF (38) $d = 1$. Then it receives the three-parameter regularly varying generalized hypergeometric frequency distribution as a special case (see [11, 14]).

3.1 The log-log plot of 3d-RGHP

In this subsection $\ln \hat{p}_x(\alpha)$ versus $\ln x$ is considered. Let us establish the following theorem.

Theorem 2. *The 3d-RGHP model (38) is upward/downward convexity.*

Proof. From $\hat{p}_x(\alpha)$ in (38) we have

$$t_x = \frac{\ln \hat{p}_x(\alpha)}{\ln x} = \frac{1}{\ln x}\Big(m_1 + \sum_{k=1}^{x} \ln\Big(\frac{1}{k^d}\prod_{j=1}^{d}\frac{(k+a_j-1)(k+b_j-1)}{(c_j+k)}\Big)\Big), \quad (40)$$

where $x = 2, 3, ...,$ and $m_1 = \ln \hat{p}_0(\alpha)$.

Now, it can be written

$$\begin{aligned}(\ln x)t_x = m_1 + \textstyle\sum_{k=1}^{x}\Big(&g_{a_1}(k) + ... + g_{a_d}(k) + g_{b_1}(k) + ... + g_{b_d}(k)\\ &- g_{c_1}(k) - ... - g_{c_d}(k)\Big), \quad x = 2, 3, 4, ...,\end{aligned} \quad (41)$$

where for $\epsilon \in \mathbb{R}^+$ and $t \in [1, \infty)$ and $g_\epsilon(t) = \ln(1 + \frac{\epsilon}{t})$. The function $g_\epsilon(t)$ is bounded and increasing. Compared to Danielian and Astola [11], it is readily seen that for $x = 1, 2, ...,$

$$\sum_{k=1}^{x} g_\epsilon(k) = \int_1^x g_\epsilon(t)dt + s(\epsilon) + \delta_x(\epsilon), \quad (42)$$

where $s(\epsilon)$ is constant and $\lim_{x\longrightarrow\infty} \delta_x(\epsilon) = 0$.

If we set

$$m_2 = m_1 + s(a_1) + ... + s(a_d) + s(b_1) + ... + s(b_d) - s(c_1) - ... - s(c_d)$$

then by substituting (42) into (41) we get the following when $x \longrightarrow \infty$

$$\begin{aligned}(\ln x)t_x = m_2(1 + o(1)) + \textstyle\int_1^x \Big(&g_{a_1}(x) + ... + g_{a_d}(x) + g_{b_1}(x) + ... + g_{b_d}(x)\\ &- g_{c_1}(x) - ...g_{c_d}(x)\Big)dt.\end{aligned} \quad (43)$$

Moreover, compared to Danielian and Astola [11], for $\epsilon \in (0, \infty)$ and $x = 1, 2, ...,$

$$\textstyle\int_{0^-}^x g_\epsilon(t)dt = (x + \epsilon)\ln(x + \epsilon) - x\ln x - \epsilon\ln\epsilon + x. \quad (44)$$

By substituting (44) into (43) it can be obtained when $x \longrightarrow \infty$

$$\frac{\ln \hat{p}_x(\alpha)}{\ln x} = a_1+a_2+...+a_d-d+b_1+b_2+...+b_d-d-c_1-c_2-...-c_d+\frac{m_3}{\ln x}(1+o(1)), \tag{45}$$

(m_3 is some real constant). By (45) it follows that, at least for large values of x, the deviations of $\ln \hat{p}_x(\alpha)$ versus $\ln x$ from the straight line

$$\sum_{j=1}^{d}(a_j + b_j - c_j - 2) = constant$$

are small which shows upward/downward convexity. The proof of Theorem 2 is finished. □

3.2 Convexity and unimodality

Unimodality of the 3d-RGHP and convexity on the right side PDF tail are essential features for frequency distributions in bioinformatics data sets. For details about this see [3, 6, 10, 21, 24, 25, 26, 28, 29, 30, 32, 34]. So, in this subsection, we shall attempt to prove such features for the 3d-RGHP model. We give a theorem as follows.

Theorem 3. *The 3d-RGHP model (38) is unimodal and asymptotically downward convex.*

Proof. From $\hat{p}_x(\alpha)$ in (38) we obtain recursive formula for $x = 0, 1, 2, ...,$ as follows

$$\frac{\hat{p}_{x+1}(\alpha)}{\hat{p}_x(\alpha)} = \frac{(a_1 + x)(a_2 + x)...(a_d + x)(b_1 + x)(b_2 + x)...(b_d + x)}{(1 + x)^d(c_1 + x + 1)(c_2 + x + 1)...(c_d + x + 1)}. \tag{46}$$

Let us rewrite the recursive (46) as

$$\frac{\hat{p}_{x+1}(\alpha)}{\hat{p}_x(\alpha)} = \frac{(a_1 + x)(b_1 + x)}{(1 + x)(c_1 + x + 1)} \times \frac{(a_2 + x)(b_2 + x)}{(1 + x)(c_2 + x + 1)} \times ... \times \frac{(a_d + x)(b_d + x)}{(1 + x)(c_d + x + 1)}. \tag{47}$$

We investigate unimodality of 3d-RGHP $\hat{p}_x(\alpha)$ for some cases.

Case 1. Assume that, in (47) (in its first expression of right side), $min(a_1, b_1) < 1$. For example if $a_1 < 1$ then $\frac{(a_1+x)}{(1+x)} < 1$ and increases when x increases. In view of $b_1 < c_1 + 1$, the expression $\frac{(b_1+x)}{(c_1+x+1)} < 1$ and increases when x increases. So, their product $\frac{(a_1+x)(b_1+x)}{(1+x)(c_1+x+1)} < 1$ and increases when x increases. Similarly, this situation can be verified for $\frac{(a_2+x)(b_2+x)}{(1+x)(c_2+x+1)}$ and $\frac{(a_d+x)(b_d+x)}{(1+x)(c_d+x+1)}$. Therefore, based on (47)

$$\frac{(a_1+x)(b_1+x)}{(1+x)(c_1+x+1)} \times \frac{(a_2+x)(b_2+x)}{(1+x)(c_2+x+1)} \times ... \times \frac{(a_d+x)(b_d+x)}{(1+x)(c_d+x+1)} < 1,$$

and the model $\hat{p}_x(\alpha)$ decreases and is downward convex. This concludes that 3d-RGHP model $\hat{p}_x(\alpha)$ is unimodal.

Case 2. Suppose that $w_j = (c_j + 2 - a_j - b_j)$, $j = 1, 2, ..., d$. From (47) let us write

$$\begin{cases} \frac{(a_1+x)(b_1+x)}{(1+x)(c_1+x+1)} = 1 - \frac{w_1}{(c_1+x+1)} + \frac{(a_1-1)(b_1-1)}{(1+x)(c_1+x+1)}, \\ \frac{(a_2+x)(b_2+x)}{(1+x)(c_2+x+1)} = 1 - \frac{w_2}{(c_2+x+1)} + \frac{(a_2-1)(b_2-1)}{(1+x)(c_2+x+1)}, \\ . \\ . \\ . \\ \frac{(a_d+x)(b_d+x)}{(1+x)(c_d+x+1)} = 1 - \frac{w_d}{(c_d+x+1)} + \frac{(a_d-1)(b_d-1)}{(1+x)(c_d+x+1)}. \end{cases} \tag{48}$$

From (48), it can be concluded that if $(a_j - 1)(b_j - 1) \leq w_j$ then the Case 1 is received. Moreover, if $(a_j - 1)(b_j - 1) > w_j$ then for enough large x the $\hat{p}_x(\alpha)$ is again downward convex and in our cases is unimodal.

The proof of Theorem 3 is completed. □

The following corollary is given.

Corollary 4. Based on the Theorems 1, 2 and 3 we conclude that PMF and corresponding 3d-RGHP defined by probabilities (38) can be considered as a *new* non-linear and multi-parameter probabilistic model for the needs of data analysis and statistical inferences in bioinformatics and relevant fields.

4 Numerical analysis of the 3d-RGHP parameter: skewness variation and limitation of power law property

In Theorem 2, section 3, we proved that the deviations of $\ln \hat{p}_x(\alpha)$ versus $\ln x$ from the straight line could be small, at least for some large values of x. However, the theory does not estimate how long the power law-like region on x should be, how soon it may occur, and how stable this extreme region may be when the parameter set is changed. To answer these questions, we fixed the value $d = 2$ and $\theta = 1$ and use 4 distinct sets of parameters.

We considered 3d-RGHP in several special cases. We considered $\theta = 1$ as the case corresponding to regularly varying function property. Furthermore, we set $d = 2$ and thus modeled the 6-RGHP for $\alpha = (a_1, a_2, b_1, b_2, c_1, c_2)$. We gradually

increased all or some of the parameters. The values of ρ corresponding to the selected six-parameter sets were calculated. All values are given as follows

$$\alpha = (0.5, 0.9, 1.8, 1.9, 2.6, 2.7), \qquad -\rho = -4.2$$
$$\alpha = (0.7, 1.8, 2.7, 3.9, 4.6, 5.9), \qquad -\rho = -5.4$$
$$\alpha = (0.8, 1.9, 3, 5, 6.4, 9.1), \qquad -\rho = -8.8$$
$$\alpha = (0.95, 2.3, 3.6, 4.7, 5.9, 55), \qquad -\rho = -53.35$$

We plotted the PMF of 3d-RGHP (38) for the parameter sets at $d = 2$ (i.e. 6-RGHP). As proved in Theorem 3, Figures 1(A-D) show that for the different parameter sets the PMFs are unimodal (mode is observed at $x = 1$). However, if the gradual increasing all or some of the parameters, a curvature of the right-side tail of PMF was negatively corbeled fatness of the tail end. We also observed significant shift and variation of the power law-like right-side tail of the PMFs. Indeed, Figures 1(E-H) show the log-log plots of the double-truncated 3d-RGHP for the case 6-RGHP with different values of the parameters presented in Figures 1(A-D). We specified the cutoff values of random variable x interval that provide approximately similar slops of the function tails (asymptotic power law-like exponents). The left-side cutoffs and the right end of the intervals follows: In Figure 1-E, $x \in [5, 15]$; Figure 1-F, $x \in [10, 20]$; Figure 1-G, $x \in [15, 25]$; Figure 1-H, $x \in [20, 30]$. Figures 1(E-H) show that the power law tail interval is highly sensitive to parameter values, and a significant fraction of the probabilities within random variable intervals are significantly deviated from the straight line. Our numerical analysis observations are agreed with the variation of the value of regular variation ρ, at which is (negative) exponent of the PMF (at $\theta = 1$).

Our numerical analysis observations are agreed with the variation of the value of regular variation ρ, at which is (negative) exponent of the PMF (at $\theta = 1$).

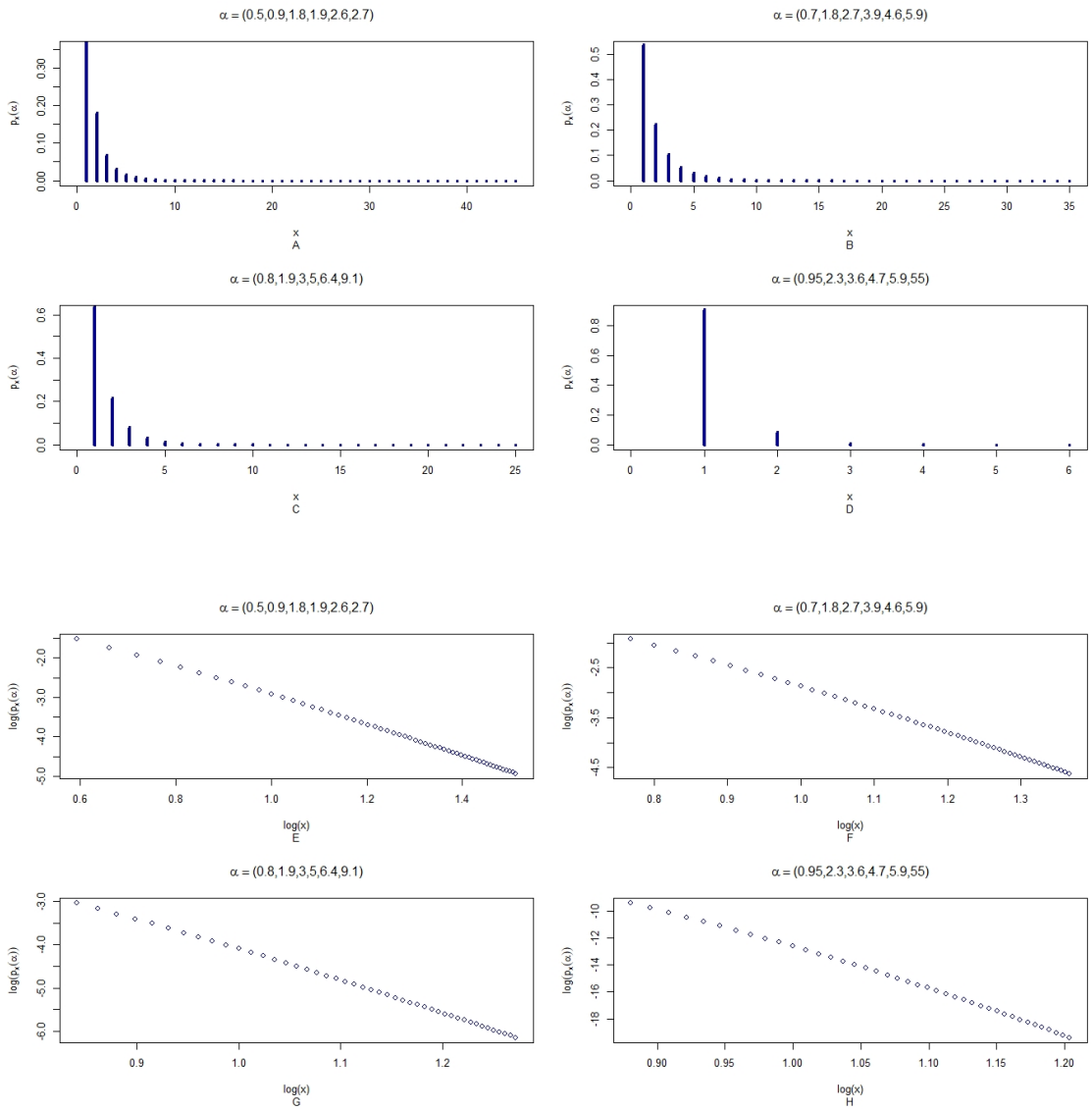

Figure 1: Illustrations of the PMF and log-log plots of 3d-RGHP model (38) when $d = 2$ (6-RGHP) for possible values of six parameters $\alpha = (a_1, a_2, b_1, b_2, c_1, c_2)$.

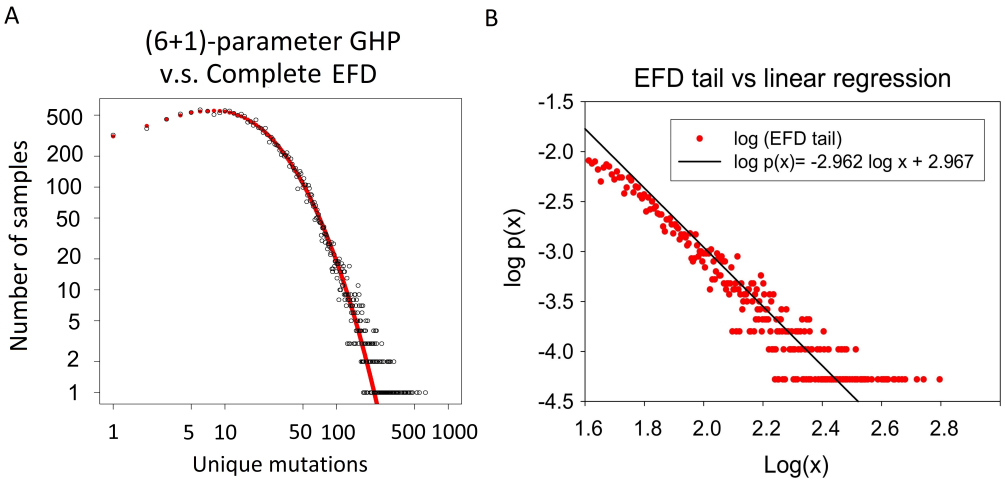

Figure 2: Frequency distribution of superficial melanoma mutations. 7-parameter Generalized PDF. A: Example of stationary BDP-derived PDFs model's parameterization using complex form of the EFD. The frequency distribution of the number of unique mutated samples of gene loci detected in the large collection of human melanomas have modeled by (6+1)-parameter GHP distribution. Estimated parameters and goodness of fit criteria values are $\hat{a}_1 = 0.743$; $\hat{a}_2 = 2.341$; $\hat{b}_1 = 17.323$; $\hat{b}_2 = 24.584$; $\hat{c}_1 = 28.375$; $\hat{c}_2 = 15.578$; $\hat{\theta} = 0.99$; $\hat{\rho} = 2.962$. $\Psi = 5.641$. B: Regression model with negative slope $-\hat{\rho} = -2.962$ approximates asymptotic decreasing the EFD tail. Correlation coefficient between experimental and predicted data is 0.874; $p - value < 0.0001$ (t-test, F-test). The EFD r.v. x interval [65-625] is shown on axis X on $\log_{10}$ scale.

5 Application of the 3d-RGHP model. Distribution of the number of unique mutated samples in genes of melanoma patient's genomes

To demonstrate the ability of our (3d+1)-GHP (27) and 3d-RGHP (38) models to analyze complex forms of the EFD with heavy tail on the right, we have downloaded somatic mutation data for the genes in several human cancers from the COSMIC database (https://cancer.sanger.ac.uk/cosmic). We studied the frequency distribution of the number of unique gene loci mutated samples detected in the large collection of human melanomas. Our data set used in this study, includes 18,996 unique mutations, collected from 587,357 human superficial melanoma's gene samples.

Figure 2A shows that the EFD is an unimodal, skewed and has a heavy tail on the right that is in the log-log plot. The mean value of the number of unique mutated gene samples consists of 30.9, median 20, mode 5. One mutation per gene was observed in 320 samples, a maximum number of unique mutated samples per gene equals 626.

We parameterized the (3d+1)-GHP (27) with 7 parameters. We utilized the bounded quasi-Newton limited-memory BFGS (L-BFGS-B) algorithm to estimate the parameters, using our goodness of fit Ψ criteria [26, 28].

We were able to parameterize our (6+1)-parameter PDF using (27)at a high level of the goodness of fit criteria value (Figure 2A). Furthermore, our results show that the parameterized PDF fits well not only to the long tail of the EFD but also to the entire random variable domain of the EFD (Figure 2A). Seven estimated parameter values, calculated $\hat{p}_0$ and the Ψ criteria value are showed on the legend of Figure 2.

We observed that the estimated parameter $\hat{\theta} = 0.99$ and $\hat{p}_0 = 0.01$. According to these estimates, we suggest (see Discussion) that we could use the parameters to analyze the Theorem 1 and Remark 1 predictions referring to parameter ρ - the power law exponent at asymptotic values of x (observed number of unique melanoma somatic mutations in gene samples). According to the estimated values of our model parameters and Theorem 1 $\hat{\rho} = 2.962$. Figure 2B shows that the linear regression function with a negative slope equals -2.962 fits to the experimentally observed tail end of the EFD for the r.v. values larger then 115 unique gene loci mutated samples.

6 Discussion and Conclusion

In this study, we introduced and characterized a multi-parameter family of generalized GHP models, generated by the stationary g-BDP and defined by PMF (27) and associated probability measure called (3d+1)-GHP (28). We specified a generalized probability subfamily called the 3d-RGHP as a regularly varying model associated with PMF (38).

Our 3d-RGHP model represents a broad family of regularly varying PDFs with long right tails. To offer 3d-RGHP as a *new* EFD in bioinformatics, the *statistical facts*, such as regular variation at infinity, the exhibition of asymptotically constant slowly varying component, upward/downward convexity, unimodality, and convexity, were established. The non-linear dependence of the transition rates from a state of the stationary BDP allows developing many applications. In particular, the model presents a rigorous and flexible analytical framework for analysis of the skewed distribution functions generated by BDP observed in big data bioinformatics.

The number of unique mutated samples in a gene of melanoma genome shows a skewed EFD distribution with a long right tail. Skewness is strongly characterized by values and order of mode=5, median=20 and mean=31. This proposes that the mutagenesis in melanoma cells is a highly-intensive and driven multiple probabilistic mechanisms. Our model successfully describes the complexity of

the EFD and the results of our estimates (Figure 2A) that demonstrated the ability of model (27) to quantify statistical characteristics of the EFD in the contexts of the skewed long tail and its potential power law properties.

Meanwhile, according to our results, a probability of non-observed somatic mutation consists of only 1.1% suggesting that the experimental outcome space is almost fully detected. Prediction of the p_0 estimation provides a statistical basis for unbiased PDF identification.

Importantly, Figure 2A showed that our (6+1)-GHP PDF accurately fits the EFD on the entire random variable domain $X = x;\ x \in [1, 625]$. Assuming that birth and death distributions are stationary processes, we noted that mutagenesis in the melanoma genome has an extreme (near 'critical') regime defined by the value of $\hat{\theta} = 0.99$, that according to our theory is corresponding to regularly varying transition rates $\theta = 1$. Furthermore, by our estimation, the experimentally observed power law tail end exponent equals $-\hat{\rho} = -2.961$. Figure 2B shows that the linear regression with negative slope equals -2.962 fits well a segment of r.v. x at the tail end of our EFD supporting our Theorem 1 and Remark 1. These results suggest that our multi-parameter probabilistic models may be practically used for real bioinformatics and other big data analysis.

Interestingly, these results suggest that the probability values of the power law exponent asymptotic are reduced significantly faster at $x \to \infty$ in comparison to the standard Zipf law, having a slower decay exponent parameter ($-\rho \approx -2$). This value is commonly observed in many EFDs with a power law-like tail. Thus, even the EFD has a long skewed tail region on the right, the end of the tail of the region with asymptotic power law property could be relatively short. Figures 2(A-B) showed that the right side tail of our parameterized PDF monotonically deviates from the direct line when x-value increases. The EFD shows the region with power-law properly is overdispersed. A possible explanation of the reduction of a power-law property region is the non-linear dependence of the ratio of birth and death rate functions from state value.

Figures 1(A-D) showed the PMFs at used parameter sets are unimodal with mode value 1. Numerical analysis showed that the length and shape of the right-side tails are significantly varied with parameter value changes. However, the PMF tail's length and shape are sensitive to the parameter values. Figures 1(E-H) demonstrated the log-log plot of the double-truncated distribution representing the power law tails of the PDF presented in Figures 1(A-D) respectively. For each set of the parameters, the cut-off value interval was defined according to the linearity of the power law exponents of the tails in the log-log plot. The log-log plots of Figures 1(E-H) showed that on relatively long intervals of random variable on the right-side of the 3d-RGHP, the power law-like tail can be significantly deviated from the straight line.

Thus, our numerical analysis of the 3d-RGHP and an example of the real data parameterization of the model suggest that in dependent of the parameter value sets, the right-side tail of the PMF could be increased or decreased however, power law-like exponent asymptotic may be approached at very large r.v. values at the tail end; such interval could be relatively short, unstable and

sensitive to sample size.

In these contexts, we point out that the concept of scale-free distribution [8] model and its multiple variants, that focus on the power law tail properties of the distribution, ignore the experiment outcome space completeness and the non-linear and multiple probabilistic mechanisms, that makes such theories quite restricted, non-rigorous and practically non-applicable [10, 21, 25, 26, 27, 28, 29, 30, 31, 32, 38, 39].

Our approach to modeling of the stochastic process considering complexity, non-linear interactions/connectivity, integrity, completeness, birth and death mechanisms, stationary and transitions allows deep and flexible theory development and provides a natural bridge between modeling, algorithms and statistical methods development and the real-life applications needs. Importantly, our theory could be used for statistical hypotheses testing in big data analysis, and our model estimates and predictions are testable.

Big data sets exist and are rapidly collected not only in bioinformatics and biomedical sciences but also in all kinds of data-intensive areas corresponding to modern physics, astronomy, material sciences, geophysics, internet, technology, epidemiology, economics, finance, social sciences, community networks, visitors monitoring, etc. Skewed distributions of complex form and extreme characteristics regularly occur in data sets from these disciplines. We suggest that our g-BDP models may be useful in many such cases.

In summary:

We suggest that the many observed skewed discrete distributions are generated by non-linear and multiple independent stochastic process and their combinations;

Our 3d-RGHP model presents a flexible analytical framework for analysis of the skewed distribution functions generated by BDP observed in bioinformatics and big data science;

Our results suggest that the 3d-RGHP model can be used for statistical analysis and identification of probabilistic mechanisms and significant characteristics of evolving systems whose EFD exhibits complex form and a heavy tail at critical regime;

Our model could help better understand stochastic processes driving evolution of biological complexity and the pathological role of extreme values.

References

[1] J. ASTOLA AND E. DANIELIAN (2004) *Regularly varying distributions generated by birth-death processes*. Tampere International Center for Signal Processing (TICSP), Series no. 27, Tampere, Finland.

[2] J. ASTOLA AND E. DANIELIAN (2006) On regularly varying distributions generated by birth-death processes. *Facta Universitatis (NIS): Series Electronics Energetics*, **19**(1):109-131.

[3] J. ASTOLA AND E. DANIELIAN (2007) *Frequency distributions in biomolecular systems and growing networks*. TICSP Series no. 31, Tampere, Finland.

[4] J. ASTOLA, E. DANIELIAN AND K. GASPARIAN (2007) The maximum likelihood estimators for distributions with moderate growth. *Proceedings 6-th International Conference on Computer Sciences and Information Technologies (CSIT)*, September 24-28, Yerevan, Armenia, pp. 91-94.

[5] J. ASTOLA, K. GASPARIAN AND E. DANIELIAN (2007) Moments estimators for hypergeometric distributions. *Proceedings of the TISCP Workshop on Spectral Methods and Multirate Signal Processing (SMMSP)*, September 1-2, Moscow, Russia, pp. 233-234.

[6] J. ASTOLA, E. DANIELIAN AND S. ARZUMANYAN (2010) Frequency distributions in bioinformatics. a review, *Proceedings of the Yerevan State University: Physical and Mathematical Sciences*, **223**(3):3-22.

[7] N.H. BINGHAM, C.M. GOLDIE AND J.L. TEUGELS (1987) *Regular Variation*, Cambridge University Press, Cambridge.

[8] A.L. BARABASI, R. ALBERT AND H. JEONG (1999) Mean-field theory for scale-free random networks, *Physica A: Statistical Mechanics and its Applications*, **272**(1-2):173-187.

[9] S. BISWAS (1995) *Applied stochastic processes*. A biostatistical and population oriented approach, John Welly Sons, New York.

[10] A.D. BROIDO AND A. CLAUSET (2019) Scale-free networks are rare. *Nature communications*, **10**(1):1017.

[11] E. DANIELIAN AND J. ASTOLA (2006) On regularly varying hypergeometric distributions. *Proceedings International TICSP Workshop on Spectral Methods and Multirate Signal Processing*, Florence, Italy, September 2-3; In Astola et al. (eds.): TICSP Series 34, Tampere, Finland, 127-132.

[12] E. DANIELIAN, R. CHITCHYAN AND D. FARBOD (2016) On a new regularly varying generalized hypergeometric distribution of the second type. *Mathematical Reports*, **18(68)**, 2:217-232.

[13] D. FARBOD (2015) Asymptotic properties of maximum likelihood estimators for a generalized Pareto-type distribution. *Journal of Contemporary Mathematical Analysis*, **50**(1):44–51.

[14] D. FARBOD (2020) Estimation of parameters for a generalized hypergeometric distribution generated by birth-death process. *Journal of Mathematical Researches*, **6**(3):417-428, (in Persian).

[15] D. FARBOD AND K. GASPARIAN (2013) On the maximum likelihood estimators for some generalized Pareto-like frequency distribution. *Journal of the Iranian Statistical Society (JIRSS)*, **12**(2):211-233.

[16] B. GNEDENKO (1998) *Theory of Probability.* 6-th Ed., London: Routledge.

[17] I.S. GRADSHTEYN AND I.M. RYZNIK (2007) *Table of Integrals, Series and Products.* Academic Press, New York and London.

[18] A. GRAGEDA, D. FARBOD AND V.A. KUZNETSOV (2022) Modeling and identification of skewed distributions in genomics, proteomics and bioinformatics determined by birth-death processes, *Under Submission.*

[19] N.L. JOHNSON, S. KOTZ AND A.W. KEMP (1993) *Univariate discrete distributions*, John Wiley, New York.

[20] J.N. KAPUR (1978) Application of generalized hypergeometric functions to generalized birth and death processes. *Indian Journal of Pure and Applied Mathematics*, 9:1059-1069.

[21] G.P. KAREV, Y.I. WOLF, F.S. BEREZOVSKAYA AND E.V. KOONIN (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology*, 4:32.

[22] A.W. KEMP (1968) A wide class of discrete distributions and the associated differential equations. *Indian Journal of Statistics, Series A*, **30**:401-410.

[23] A.N. KOLMOGOROV (1956) *Foundations of the Theory of Probability.* Chelsea, New York.

[24] V.A. KUZNETSOV (2001) Distributions associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP Journal on Applied Signal Processing*, **4**, 258-296.

[25] V.A. KUZNETSOV (2002) Statistics of the number of transcripts and protein sequence encoded in the genome. In: Zhang W, Shmulevich I (eds), Computational and statistical approaches to genomics, 1st ed., 125-171, Springer, Boston.

[26] V.A. KUZNETSOV (2003) Family of skewed distributions associated with the gene expression and proteome evolution. *Signal Processing*, **33**(4):889-910.

[27] V.A. KUZNETSOV (2003) Stochastic model of evolution of conserved protein coding sequences. *AIP Conference Proceedings, American Institute of Physics*, **665**, 369-380.

[28] V.A. KUZNETSOV (2003) Hypergeometric stochastic model of evolution of conserved protein coding sequence in the archaeal, bacterial and eukaryotic proteomes. *Fluctuation and Noise Letters*, **3**:295-324.

[29] V.A. KUZNETSOV (2017) *Mathematical modeling of avidity distribution and estimating general binding properties of transcription factors from genome-wide binding profiles*. In: Tatarinova,T. V. and Nikolsky, Y. (eds), Biological Networks and Pathway Analysis, Springer, New York, 193-276.

[30] V.A. KUZNETSOV, G.D. KNOTT AND R.F. BONNER (2002) General Statistics of Stochastic Process of Gene Expression in Eukaryotic Cells. *Genetics*, Genetics Society of America, **161**, 1321-1332.

[31] V.A. KUZNETSOV, V.A. PICKALOV, O.V. SENKO AND G.D. KNOTT (2002) Analysis of the evolving proteomes: Predictions of the number for protein domains in nature and the number of genes in eukaryotic organisms. *Journal of Biological Systems*, **10**(4):381-407.

[32] V.A. KUZNETSOV, O. SINGH, P. JENJAROENPUN (2010) Statistics of protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome, *BMC genomics*, 11, Suppl 1 (Suppl 1):S12.

[33] D.P. MAKI (1976) A note on rates of convergence in estimating absolutely continuous distribution functions, *Journal of Approximation Theory*, **16**(3):276-280.

[34] A.S. NOVOZHILOV, G.P. KAREV AND E.V. KOONIN (2006) Biological applications of the theory of birth-and-death processes. a review, *Briefing in Bioinformatics*, **7**(1):70-85.

[35] G.S. OW, A.V. IVSHINA, G. FUENTES AND V.A. KUZNETSOV (2014) Identification of two poorly prognosed ovarian carcinoma subtypes associated with CHEK2 germ-line mutation and non-CHEK2 somatic mutation gene signatures, *Cell Cycle*, **13**(14):2262-2280.

[36] J. RODRIGUES-ALI, A. CONDE-SANCHEZ, A.J. SAEZ-CASTILLO AND M.J. OLMO-JIMENEZ (2007) A new generalization of the Waring distribution, *Computational Statistics and Data Analysis*, **51**, 6138-6150.

[37] A.N. SHIRYAEV AND A. BULINSKY (2017) *Theory of random processes*, Moscow Sate University, Moscow, (In Russian).

[38] M.P.H. STUMPF AND M.A. PORTER (2012) Critical Truths About Power Laws. *Science*, **335**(6069):665-666.

[39] M.P.H. STUMPF, C. WIUF AND R.M. MAY (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks, *Proceedings of the National Academy of Sciences USA*, **102**(12):4221-4224.

[40] C.L. WEI, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome, *Cell*, **124**(1):207-219.

[41] T. WONGSURAWAT, P. JENJAROENPUN, C.K. KWOH AND V.A. KUZNETSOV (2012) Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity, *Nucleic Acids Research*, **40**(2):e16.