PROCEEDINGS OF THE YEREVAN STATE UNIVERSITY

Physical and Mathematical Sciences 2010, № 3, p. 3–22

REVIEWS

Mathematics

FREQUENCY DISTRIBUTIONS IN BIOINFORMATICS: THE DEVELOPMENT

J. ASTOLA[1*], E. A. DANIELIAN[2], S. K. ARZUMANYAN[2]

[1] *Academician of the Academy of Science of Finland*
[2] *Chair of Probability Theory and Mathematical Statistics, YSU*

The mathematical investigation of large-scale biomolecular sequences is being carried out by analyzing the properties of events arising in such sequences. This survey is devoted to discussion of results in this field. Based on general empirical facts being fulfilled for all frequency distributions, we discuss the axiomatics suggested by J. Astola and E. Danielyan.

The axiomatics postulates the regular variation of frequency distribution with asymptotically constant slowly varying component, the form of its shape, and the stability by parameters. The verification of the axiomatics fulfillment for well-known frequency distributions is done.

The paper describes also methods of construction of new parametric families of frequency distributions. These methods are: usage of stationary distributions of birth-death process, special functions, stable densities, etc.

The problem of stability by parameters is formulated the results on stability by parameters in terms of various classical metrics are given. The conditions of regular variation for different families of frequency distributions are formulated.

Keywords**:** frequency distributbiomion, olecular sequence, regular variation, convexity, stability by parameters, asymptotic expansion.

1. Introduction. Discovering the evolution by investigating the variety of large-scale biomolecular systems there is no other way but to characterize the frequency distributions (FDs), say $\{P_n\}$, of events being important for systems' functioning. The variety and diversity of such systems do not allow to figure out and suggest a *universal* approximation for FD, i.e. suggest a *universal* model, which might be suitable in all possible situations. Based on huge datasets of biomolecular systems it has been possible to extract *only* some *common* information (statistical facts) being applicable almost to all situations for empirical FDs. Those are:

1. $\{P_n\}$ has a *skew* to the right, $P_n>0$ for all n, $\sum P_n = 1$.

The conception of *skewness* for biologist is based on *intuition* and on the shapes of graphs of empirical FDs. The *quantitative* aspects of the *skweness*

[*] E-mail: jaakko.astola@fut.fi

conception were not even exploited. Only for *Power Law* and *Pareto Law* defined below the parameter ρ was declared as a measure of skewness.

2. $\{P_n\}$ exhibits *Power Law* behavior as $n \to +\infty$ (see [1–6]).

Random variable (RV) $\xi \geq 0$ has *Power Law*, if (P denotes a probability)

$$P_n = P(\xi = n) = c(\rho)n^{-\rho}, \ 1 < \rho < +\infty, \ n \geq 1, \ c(\rho) = \left(\sum_{n\geq 1} n^{-\rho}\right)^{-1}. \quad (1.1)$$

The *Power Law* is used for estimation of the connectivity number in metabolic networks [2], of the rates of protein synthesis in protein sets of prokaryotic organisms [7], of the number of expressed genes in eukaryotic cells [8, 9], of DNA sequencing structures [9], etc.

The *Power Law* is of interest in *self-organized* growing biomolecular networks, because of its *scale-invariant* property: $P_m = \left(c(\rho)\right)^{-1} P_n P_s$ for integers $n \geq 1$, $s \geq 1$, $m = n \cdot s$. *Self-organization* means, that if we know local FDs on successive two *fractals*, then we may extraporate the FD for all system [10].

3. The *log-log plot* ($\log P_n$ versus $\log n$) of most empirical FDs $\{P_n\}$ *systematically* deviated from the straight line and show the upward/downward convexity [1–6].

That is why many new statistical FDs have been proposed. Those are: *Pareto Law* (generalization of *Power Law*)

$$P_n = c(\rho,b)(n+b)^{-\rho}, \ \text{-}1 < b < +\infty, \ 1 < \rho < +\infty, \ n \geq 1,$$

$$c(\rho,b) = \left(\sum_{n\geq 1}(n+b)^{-\rho}\right)^{-1} \ \text{(see [9]);} \quad (1.2)$$

Warring Distribution

$$P_n = \left(1 - \frac{p}{q}\right)\prod_{k=1}^{n}\frac{p+k-1}{q+k,}, \ 0 < p < q < +\infty, \ n \geq 1, \ P_0 = 1 - \frac{p}{q}, \text{ etc.} \quad (1.3)$$

Constructing new FDs the advantage is given to parametric ones, because by changing the parameters one hopes to find out the *best* approximation for unknown FD.

4. The small changes in *environment* do not have a dramatic influence on the structure of biomolecular system.

We may call this fact the *adaptivity* or the *robustness*.

We may trust or not assumptions of statistical models that lead to different empirical FDs for the events occurrence number in biomolecular systems. But, due to the probability theory, the replacement of observations' *independence* in models by various type of *weak dependence* cannot have essential impact on the behavior of P_n for *large n*.

Anyway, statistical models even being important *do not describe* the functional *mechanism* of biomolecular systems.

2. On the Mechanism. The *dynamic* of the biomolecular large-scale systems many authors try to explain with the help of *birth-death* models with various types of intensities. Their *stationary* solutions generate *skewed* to the right distributions as it requires the empirical *fact 1.*

In general, the development of any evolutionary large-scale complex biomolecular system is a result of two *fundamental* phenomena: Darwin *natural selection,* random *mutation*. The functioning is explained with the help of *standard birth-death* process [11, 12], say

$$\{\xi(t) : t \geq 0\}, \tag{2.1}$$

which is a homogeneous Markov Process with continuous time and countable number of states 0,1,2,... Moreover, conditional probability $P_{ij}(t) = P(\xi(s+t) = = j / \xi(s) = i)$ doesn't depend on s, and for $t \to 0$

$$P_{ii+1}(t) = \lambda_i t + o(t)\,,\ P_{i+1i}(t) = \lambda_{i+1} t + o(t)\,,\ i = 0,1,2,...,\ P_{ij}(t) = o(t)\,, |i-j| > 1,$$
$$\lambda_i > 0, \quad \mu_{i+1} > 0.$$

Among the numerous publications devoted to birth-death models we like to point out a pioneer paper of Yule [13], paper of Simon [14] and some recent ones (see, for instance, Granzel and Schubert [15], Bornholdt and Ebel [16], Oluic-Vicovic [17], Kuznetsov [18], Astola and Danielian [12, 19]).

The Yule's birth model is designed to describe the evolution of new species within a genus. Its *stationary* solution has the *Power Law* as a limit case.

The Kuznetsov's birth-death model describes the expression process of the genes in the eukaryotic cells, which exhibits a *strong stochastic component* (SSC), a *chaotic movement* (CM) of mutation, and a *skewed* FD of the number of events. In the model *coefficients* of the process are *linear*.

The latest one, birth-death model by Astola and Danielian gives a wide generalization of previous models. Here the *coefficients* are *non-linear* and the stationary solutions present FDs of moderate growth, i.e. $\lim\limits_{n\to\infty}(P_{n+1}/P_n) = 1$.

The stationary distribution of the process (2.1) exists iff

$$\sum_{n\geq 1}\prod_{k=1}^{n} \varepsilon_k < +\infty \tag{2.2}$$

with $\varepsilon_n = (\lambda_{n-1}/\mu_n)$, $n \geq 1$, and takes the form

$$P_n = P_0 \prod_{k=1}^{n} \varepsilon_k\,, \qquad P_0 = \left(1 + \sum_{n\geq 1}\prod_{k=1}^{n} \varepsilon_k\right)^{-1}. \tag{2.3}$$

$\{\varepsilon_n\}_1^\infty$ presents a sequence of ratios of "birth" and "death" coefficients.

Let us interpret the process (2.1) as expressed genes process in the eukaryotic cells, which exhibits a SSC, a CM. It is a discrete process with many protein coding genes in an "off" state. The production of the mRNA occurs in sporadic pulses with specific mRNA transcripts starts from initiation of the transcription of the mRNA molecule of the specific gene at moment zero. Then the mRNA molecule exports from nucleus to cytoplasm of the cell where the transcript is degrades. It leads to a new mRNA copies and degradation of transcripts. We indicate the gene expression level by integers $n = 0,1,2,\ldots$, assuming that it is a random process, and denote its distribution at a moment t by $\{P_n(t)\} = \{P(\xi(t) = n)\}$. The process is described as a standard birth-death process (2.1) and $\xi(t)$ denotes the random number of mRNA transcripts per a cell in transcripton at a moment t. This mechanism realize the way how molecules are chosen to be included into organism over time. It is a mixture of molecular sequences being before in organism and new ones, so called mutant new sequences from other organisms. In "stable" evolution process the intensities $\{\lambda_{n-1}\}$ of "birth" and $\{\mu_n\}$ of "death" do not depend on t and lead to stationary solution (2.2), (2.3). The summarized intensities take the form

$$\lambda_{n-1} = a + \lambda^*_{n-1}\,,\ \ \mu_n = b + \mu^*_n,\ \ n \geq 1\,, \tag{2.4}$$

where a >0, b >0 and λ^*_n, μ^*_n present the intensities of CM and SSC (2.4) at state n. Here λ^*_n>0, μ^*_n>0, $\lim\limits_{n\to\infty}\lambda^*_n = \lim\limits_{n\to\infty}\mu^*_n = +\infty$.

3. General Representation. The empirical *fact 1* is enough for the following conclusion: any *FD* $\{P_n\}$ may be presented in the form (2.2), (2.3) (see [20]). Indeed, due to *fact 1*, for $n \geq 1$ we have

$$P_n = P_0 \prod_{k=0}^{n-1} \frac{P_{k+1}}{P_k}\,.$$

Denoting $\varepsilon_n = \left(P_n / P_{n-1}\right)$, $n \geq 1$, we come to the first equality in (2.3). The last equality in *fact 1* leads to the last equality in (2.3). Now, P_0>0 is equivalent to (2.2).

The reverse statement is also true: any distribution of the type (2.2), (2.3) satisfies empirical *fact 1.*

According to variety and diversity of biomolecular sequences *new* parametric FDs were needed. Kuznetsov suggested three-parametric *Kolmogorov-Warring Distribution* [18]. Astola and Danielian built three-parametric *Regular Hypergeometric Distribution* [21], which takes the form

$$P_n = P_0 \prod_{k=0}^{n-1} \frac{\left(\hat{p}_1 + k\right)\left(\hat{p}_2 + k\right)}{\left(1+k\right)\left(\hat{q} + k\right)},\ \ n \geq 1,\ \ P_0 = \frac{\Gamma(\hat{q}-\hat{p}_1)\Gamma(\hat{q}-\hat{p}_2)}{\Gamma(\hat{q}-\hat{p}_1-\hat{p}_2)\Gamma(\hat{q})}\,, \tag{3.1}$$
$$0 < \hat{p}_1 + \hat{p}_2 < \hat{q} < +\infty\,,$$

where $\Gamma(\cdot)$ denotes the *Euler's Gamma-Function.*

Several variations of three-parametric *Regular Pareto type Distribution* have been proposed in [22–24], which finally acquires the following form ($\prod\limits_{m=1}^{0} \equiv 1$):

$$\begin{cases} P_n = C\left(\rho,b,c\right)\dfrac{1}{\left(n+b\right)^\rho}\displaystyle\prod_{m=1}^{n-1}\left(1+\dfrac{c-1}{\left(m+b\right)^\rho}\right), n \geq 1, \\ C\left(\rho,b,c\right) = \left(\displaystyle\sum_{n\geq 1}\dfrac{1}{\left(n+b\right)^\rho}\prod_{m=1}^{n-1}\left(1+\dfrac{c-1}{\left(m+b\right)^\rho}\right)\right)^{-1}, 0<c<+\infty, -1<b<+\infty, 1<\rho<+\infty. \end{cases} \tag{3.2}$$

Easily seen that from (3.1) for $\hat{p}_1 = 1, \hat{p}_2 = p, \hat{q} = q+1,$ we get the *Warring Distribution* (see (1.3)), and from (3.2) for c=1 we obtain the *Pareto Law* (see (1.2)).

Although the FDs (3.1), (3.2) were constructed using other principles than the stationary solutions of standard birth-death process, but it is clear that they can be presented in the form (2.2), (2.3). Note that for FDs (1.1), (1.2) and (3.2) we have to put P_0 equal to normalization factor, and find P_0 from the equality $\sum P_n = 1$. After this manipulation it is obvious that, for instance,

$$\varepsilon_{n+1} = \frac{(\hat{p}_1 + n)(\hat{p}_2 + n)}{(1+n)(\hat{q}+n)},\ n \geq 0\text{, in case of (3.1),}$$

$$\varepsilon_{n+1} = \left(1 - \frac{1}{n+b+1}\right)^{\rho}\left(1 + \frac{c-1}{(n+b)^{\rho}}\right),\ n \geq 0,\ \text{in case of (3.2),}$$

where $\{\varepsilon_n\}$ is a sequence of coefficients in general representations (2.2), (2.3) of these FDs.

For $m \geq 1$ consider m-parametric $FD\{P_n(\vec{c}_m)\}$, where $\vec{c}_m = (c_1, c_2, ..., c_m)$ and c_i, $i = \overline{1,m}$, are *parameters*. It is convenient to choose parameters to be *independent*, and ranges of their changes be also *independent*. It means that there are no relationships among them of equality and of inequality types respectively.

All FDs presented above have independent parameters. But for *Warring Distribution* and *Regular Hypergeometric Distribution* the independence of ranges of parameters' changes does not take place. The situation is improved by making the linear transformations of parameters:

$$\rho = q + 1 - p,\ p = p \quad \text{and} \quad \rho = \hat{q} + 1 - \hat{p}_1 - \hat{p}_2\ \hat{p}_1 = \hat{p}_1\ \hat{p}_2 = \hat{p}_2 \qquad (3.3)$$

in the first and in the second cases correspondingly.

Definition 1. We say that $\{P_n(\vec{c}_m)\}$ is well-defined, if the coefficients $\varepsilon_1, ..., \varepsilon_m$ of its general representation uniquely define parameters $c_1, ..., c_m$.

Theorem 1 (see [25]). All above presented FDs are well-defined.

4. Regular Variation. Due to characteristic property $P_n = Ln^{-\rho}$, $L = c(\rho) \in R^+ = (0, +\infty), n \geq 1$ (see (1.1)), the empirical *fact 2* has been interpreted in mathematical sense in [19, 22] as a *regular variation* of FD.

Definition 2 (see [26, 27]). The sequence $\{X_n\}$ of positive numbers varies regularly as $n \to +\infty$ with exponent $\alpha \in R^1 = (-\infty, +\infty)$, if for any integer $s \geq 2$

$$\lim_{n \to \infty} (X_{s \cdot n} / X_n) = s^{\alpha}. \qquad (4.1)$$

The case $\alpha = 0$ presents the slowly varying sequence, which is usually denoted by $\{L(n)\}$.

The Definition 2 is equivalent to the representation $X_n = n^{\alpha} \cdot L(n), n \geq 1$, with some arbitrary chosen $L(0) > 0$.

In general, the sequence $\{L(n)\}$ may show quite different behavior as $n \to +\infty$.

Lemma 1 (see [28], p. 6–8). Let $0 \le \underline{L} \le \overline{L} \le +\infty$. Then, there is a slowly varying sequence $\{L(n)\}$ such that $\underline{L} = \varliminf_{n\to\infty} L(n), \overline{L} = \varlimsup_{n\to\infty} L(n)$.

But, according to the properties of all before known FDs, in [23] the following general property of FDs was suggested.

Property 1. FD $\{P_n\}$ varies regularly as $n \to +\infty$ with exponent $(-\rho)$, $1 < \rho < +\infty$, and exhibits an asymptotically constant slowly varying component (ACSVC) L, i.e.

$$P_n = L(n)\cdot n^{-\rho}, n \ge 1, \text{ and } \lim_{n\to\infty} L(n) = L \in R^+ . \tag{4.2}$$

For example (see [22, 23]), the following statement holds.

Theorem 2. The FDs (3.1) and (3.2) satisfy Property 1 with

$$\rho = \hat{q} + 1 - \hat{p}_1 - \hat{p}_2, L = \frac{\Gamma(\hat{q}-\hat{p}_1)\cdot\Gamma(\hat{q}-\hat{p}_2)}{\Gamma(\hat{p}_1)\cdot\Gamma(\hat{p}_2)}\cdot\frac{1}{\Gamma(\hat{q}-\hat{p}_1-\hat{p}_2)},$$

ρ is the parameter in (3.2), and

$$L = C(\rho,b,c)\prod_{n\ge1}\left(1+\frac{c-1}{(n+b)^\rho}\right).$$

In [29] the following $2m$-parametric FD was considered:

$$\begin{cases} P_n = P_0\prod_{k=1}^{n}\prod_{i=1}^{m}\dfrac{k+\hat{p}_i}{k+\hat{q}_i}, n \ge 1, \quad \hat{p}_i > 0, \hat{q}_i > 0, i = \overline{1,m}, \\ P_0 = \left(1+\sum_{n\ge1}\prod_{k=1}^{n}\prod_{i=1}^{m}\dfrac{k+\hat{p}_i}{k+\hat{q}_i}\right)^{-1}, \quad \rho = \sum_{i=1}^{m}(\hat{q}_i - \hat{p}_i) > 1. \end{cases} \tag{4.3}$$

Theorem 3. The FD (4.3) exhibits the asymptotic expansion

$$P_n = \frac{L}{n^\rho} + \frac{M}{n^{\rho+1}} + o\left(\frac{1}{n^{\rho+1}}\right), n \to +\infty, \tag{4.4}$$

where

$$L = P_0 \cdot \prod_{i=1}^{m}\frac{\Gamma(\hat{q}_i+1)}{\Gamma(\hat{p}_i+1)} \in R^+ , \tag{4.5}$$

$$M = -L\frac{\rho + \sum_{i=1}^{m}\left(\hat{q}_i^2 - \hat{p}_i^2\right)}{2} \in (-\infty, 0).$$

The FD (4.3) is a generalization of (3.1). Theorem 3, in particular, says that $\{P_n\}$ of type (4.3) varies regularly as $n \to +\infty$ with exponent $(-\rho)$ and exhibits ACSVC L (see (4.4), (4.5)), i.e. satisfies Property 1. This is the content of the first term at the right-hand-side of expansion (4.4). The second one gives additional information on "smoothness" of $\{P_n\}$, which agrees with the empirical *fact 4*. Indeed, the "smoothness" of *continuous* functions comes to light, if they can be presented in the form of *Taylor's Series*. The expansion (4.4) is the *analog* of such "smoothness" for a *discrete* case.

The asymptotic expansion (4.4) is natural for all known FDs. That is why we may even postulate it as an *Extended Property 1* for FDs.

Denote by m_α the moment of order $\alpha \in R^+$ *of FD* $\{P_n\}$. If $\{P_n\}$ varies regularly with exponent $(-\rho)$, $1<\rho<+\infty$, then $m_\alpha<+\infty$ for $\alpha<\rho-1$ and $m_\alpha=+\infty$ for $\alpha>\rho-1$(see [26]). If $\{P_n\}$ exhibits ACSVC, then $m_{\rho-1}=+\infty$. *The Problem* of asymptotic behavior of *truncated* moment $\mu_{\rho-1}(x)=\sum\limits_{n<x} n^{\rho-1}P_n$ as $x\to+\infty$ arises.

Theorem 4 (see [22, 30]). Let FD$\{P_n\}$ satisfies Property 1. Then,

$$\mu_{\rho-1}(x)=(L\cdot\ln x)(1+o(1)),\ x\to+\infty. \tag{4.6}$$

For concrete FDs even more terms of asymptotic (4.6) can be obtained.

For instance, for FD (1.3) (see [30] and [22], p. 140–145): if $q-p=1$ and p is an integer, then $\mu_1(x)=p\left\{\ln x+(C-A_1(p))+2x^{-1}+O(x^{-2})\right\}$, $x\to+\infty$, where $A_1(p)=1+\sum\limits_{n=1}^{p}n^{-1}$; if $q-p=2$ and p is an integer, then

$$\mu_2(x)=2p(p+1)\left\{\ln x+(C-A_2(p))+\frac{2p-1}{2x}+O(x^{-2})\right\},\ x\to+\infty,$$

where $A_2(p)=\dfrac{3p+4}{2(p+1)}+\sum\limits_{n=1}^{p}n^{-1}$. Everywhere C denotes the *Euler's constant.*

5. Convexity and Monotonicity. For a sequence $\{X_n\}$ of positive numbers we consider the following two types of *convexity*: for n=0,1,2,…

1) $X_{n+2}-2X_{n+1}+X_n<(>0)$ upward (downward) convexity,

2) $(X_n/X_{n+1})<(>)(X_{n+1}/X_{n+2})$ log-upward (log-downward) convexity.

The *Problem* of comparison of these convexities arises.

Lemma 2 (see [24, 31]). The upward (log-downward) convexity of $\{X_n\}$ implies its log-upward (downward) convexity.

For $\{X_n\}$ the existence of a pair "upward and log-downward convexity" contradicts Lemma 2. But the pair "downward and log-upward convexity" may exist (see [12], p. 60–61).

Let $\{X_n\}$ and $\{Y_n\}$ be positive sequences. Obviously, if $\{X_n\}$and $\{Y_n\}$ are log-upward (log-downward) convex, then $\{X_n\cdot Y_n\}$ is of the same type.

What can we say about $\{X_n\pm Y_n\}$? It turns out that, for instance, the following statement holds.

Lemma 3 (see [31]). Let $\{t_n\}=\{Y_n/X_n\}, n\ge1$, be downward convex. If $\{X_n+Y_n\}$decreases and is log-downward convex, then $\{X_n+Y_n\}$is of the same type.

Lemma 2 is of interest in the way of constructing FDs with given properties, because of the following statement. Let $\rho\in(1,+\infty), L\in R^+, s\ge1$ be an integer, $M_i\in R^1\setminus\{0\}, i=\overline{1,s}, 0<\alpha_1<\alpha_2<...<\alpha_s$, are given.

Corollary 1. The sequence $\left\{ \frac{L}{n^{\rho}} + \sum_{i=1}^{s} \frac{M_i}{n^{\rho+\alpha_i}} \right\}$ decreases and is log-downward convex starting from some index $n_0 \geq 1$.

With the help of this statement the following general result is proved.

Theorem 5 (see [31]). There is a decreasing and log-downward convex *FD* $\{P_n\}$ satisfying asymptotic expansion with above a priori given constants

$$P_n = \frac{L}{n^{\rho}} + \sum_{i=1}^{s} \frac{M_i}{n^{\rho+\alpha_i}} + o\left(\frac{1}{n^{\rho+\alpha_s}} \right), n \to +\infty \tag{5.1}$$

(compare to (4.4)).

Using the method developed in [31], it is possible to build FD of type (5.1) with any finite number of log-upward/log-downward convex pieces in its graph, the last of which decreases and is log-downward convex.

Very often it is easier to prove such kinds of statement using *continuous analogs* (CA) of the sequence $\{X_n\}$ of *positive* numbers. We say that the function f(t) defined on $[0,+\infty)$ is a CA of $\{X_n\}$, if *f* is continuous on $[0,+\infty)$, and $f(n) = X_n, n \geq 0$. The "smoothness" of *f* allows to apply methods of Mathematical Analysis.

The *linear* CA (LCA) of $\{X_n\}$ sometimes is *more preferable* for other purposes among all other continuous ones. The shape of its graph is formed by possibly *minimal* number of convex pieces.

Definition 3 [22]. We say that $f(t)$ defined on $[0,+\infty)$ is the LCA of $\{X_n\}$, if: $(a)\, f(n) = X_n, n \geq 0$; $(b)\, f(t)$ is continuous on $[0,+\infty)$; $(c)\, f(t)$ is linear on each $[n, n+1], n \geq 0$.

It can be easily proved that $\{X_n\}$ and its LCA are unimodal (or not) with the same mode simultaneously, and have the same intervals of *monotonicity* and *convexity*.

Now let us discuss the properties of *monotonicity* and *convexity* of known FDs. The *famous* ones (see (1.2), (1.3)) are *decreasing* and *log-downward convex*.

For FDs (3.1), (3.2) constructed at the second stage of development we combine the results from [23] and [24] in the following *statement*.

Theorem 6. FDs (3.1), (3.2) are unimodal. Their graphs are formed by no more than two monotone, and no more than three log-convex (convex) pieces.

6. Around Empirical Fact 3. Due to substantiated Property 1 (see (4.2)),

$$\frac{\log P_n}{\log n} = (-\rho) + \frac{\log L(n)}{\log n} \text{ for n=1,2,...,} \tag{6.1}$$

where in case of *Power Law* $L(n) = c(\rho)$ doesn't depend on *n* (as a rule, the biologists deal with the log-log plot of $\{P_n\}$). According to empirical *fact 3* one may conclude that the upward (downward) convexity of $(\log P_n / \log n)$ is only the result of piecewise convexity of $\{\log L(n)\}$, where, obviously, the last piece is *downward* convex. Note that $\lim_{n\to\infty} (\log L(n) / \log n) = 0$, which follows from the following property of $\{L(n)\}$: for any $\varepsilon \in (0,1)$ and n large enough the inequality $n^{-\varepsilon} < L(n) < n^{\varepsilon}$ holds. Now, writing (6.1) in the form $\log P_n = (-\rho)\log n + {} + \log L(n)$ we conclude that $\{\log P_n\}$ is piecewise convex, and may affect on the type of convexity of $\{P_n\}$ only in *initial* finite interval. But in finite interval the number of log-convex (convex) pieces for $\{P_n\}$ cannot be more than *finite*. So, this number on $[0,+\infty)$ is finite too. Next argument: any biomolecular system comes to a structure with minimal "energy expenditure", which affects on FDs. The situation, when $\{P_n\}$ has more than one log-convex (convex) piece under "slow mutation" leads to *unnecessary* "energy expenditure". (The famous FDs (1.1), (1.2), (1.3) have exactly *one log-downward* (*downward*) convex piece). But building the mathematical theory of FDs in bioinformatics with arbitrary speed of "mutation" one has to allow for FDs to have more than one (at least two or three) log-convex (convex) pieces. Such a situation may be explained with the help of evolution process' functioning. Indeed, the value $\lambda^*_{n-1} - \mu^*_n$ (see (2.4)) as $n \to +\infty$ creates nonlinear deterministic "shift" over time n in new transcripts. If $\left(\lambda^*_{n-1} / \mu^*_n\right) < 1$ over time n, then the SSC manages the situation. The deterministic "shift" of SSC leads to the *log-downward (downward) convexity o*f $\{P_n\}$. The CM doesn't take part in stabilization process. If $\lambda^*_{n-1} / \mu^*_n \approx 1$ for "large" massif of n, then the condition $(a/b) < 1$ (see (2.4)) stabilizes the process. The CM gets possibility to make observed affect on the process. It's maximal influence can be easily seen on the initial massif on indices, around the mode (now intensities of SSC and CM are comparable). Just around the mode *log-upward (upward)* small *convex* piece of $\{P_n\}$ appears.

Above said and Theorems 5, 6 we propose for the following

Property 2. FD$\{P_n\}$ is unimodal and it's graph is formed by no more than three log-convex (convex) pieces, the last of which is log-downward (downward) convex.

The Property 2 with convex (log-convex) pieces has been suggested in [23] (in [24]). For the FD (4.3) the following statement takes place [24].

Theorem 7. Let for vectors $(\hat{p}_1,\ldots,\hat{p}_m)$ and $(\hat{q}_1,\ldots,\hat{q}_m)$ in (4.3) the numbers $\hat{p}_{(i)}$ and $\hat{q}_{(i)}, i=\overline{1,m}$, present the i-th order statistics respectively. Then the conditions.

$$\hat{p}_{(1)} < \hat{q}_{(1)}, \ldots, \hat{p}_{(m)} < \hat{q}_{(m)} \tag{6.2}$$

are sufficient for $\{P_n\}$ of type (4.3) to be decreasing and log-downward convexity.

The following *Problem* stays unsolved: find necessary and sufficient conditions for the fulfillment of Property 2 with one, two, three log-convex pieces for FD (4.3).

Let $\{\varphi_n\}$ be increasing sequence of positive numbers with $\lim\limits_{n\to\infty} \varphi_n = +\infty$, $\lim\limits_{n\to\infty} (\varphi_{n+1} / \varphi_n) = 1$. Then the sequence $\{\psi_n\}$, where $\psi_n = 1 + (\mu\varphi_n / b)$, $\mu > 0$, $b > 0$, $n \ge 1$, possesses the same properties.

A deep investigation on the stationary distribution (2.2)–(2.3) with $\lambda^*_{n-1}=\lambda\varphi_{n-1},\ \mu^*_n=\mu\varphi_n,\ n\ge 1$, and with $\lambda^*_{n-1}=\lambda\varphi_{n,}\ \mu^*_n=\mu\varphi_n,\ n\ge 1$, in (2.4) has been done in [12, 19, 32]. The following FD was extracted $\left(\prod_{m=1}^{0}\equiv 1\right)$:

$$P_n=P_0\frac{c}{\psi_n}\prod_{m=1}^{n-1}\left(1+\frac{c-1}{\psi_m}\right),\ n\ge 1,\quad P_0=\left(1+c\sum_{n\ge 1}\frac{1}{\psi_n}\prod_{m=1}^{n-1}\left(1+\frac{c-1}{\psi_m}\right)\right)^{-1} \tag{6.3}$$

with *either* $0<c<1, \sum\psi_n^{-1}=+\infty$, or $0<c<+\infty, \sum\psi_n^{-1}<+\infty$.

Theorem 8. Let $\psi_0=1, \{\psi_n\}$ increases, $\lim\limits_{n\to+\infty}(n/\psi_n)=0$. Then:

1. $\{P_n\}$ varies regularly with exponent $(-\rho)$, iff $\{\psi_n\}$ varies regularly with exponent ρ.

2. If $1<\rho<+\infty$, then $\sum\psi_n^{-1}<+\infty$.

3. $\{P_n\}$ and $\{\psi_n\}$ exhibits ACSVCs L and L_1 simultaneously. Moreover, $L=(P_0+c-1)/L_1$.

The statements 1 and 2 in Theorem 8 are established in [12], the statement 3 is proved below.

Put $a_{n+1}=\psi_{n+1}P_{n+1}, n\ge 0\ \ (a_0=\psi_0P_0=P_0)$. Then, due to (6.3), $a_{n+1}=a_n+(c-1)P_n=\ldots=P_0c+(c-1)\sum\limits_{k=1}^{n-1}P_k$. So, we obtain the reverse equalities

$$\psi_n=\left(P_0c+(c-1)\sum_{k=1}^{n-1}P_k\right)/P_n,\ n\ge 1. \tag{6.4}$$

Since, $\psi_n=n^{\rho}L_1(n), P_n=n^{-\rho}L(n)$, therefore, from (6.4) we obtain $L(n)=\left(P_0c+(c-1)\sum\limits_{k=1}^{n-1}P_k\right)/L_1(n),\ n\ge 1$. Letting $n\to+\infty$ we prove the statement 3.

Theorem 9 (see [12], p. 54–55,67–68). Let the conditions of Theorem 8 hold, and $\{\psi_n\}$ varies regularly with exponent $\rho\in(1,+\infty)$. If $\{\psi_n\}$ is downward and log-upward convex, then $\{P_n\}$ decreases and is log-downward convex.

According to Theorems 8, 9 under the conditions of Theorem 9 the FD (6.3) satisfies Properties 1 and 2.

7. Back to General Representation. Below let the sequence $\{\varepsilon_n\}$ of positive numbers be the sequence of *coefficients* in *general representation* (2.2), (2.3) of FD $\{P_n\}$. So, $\varepsilon_n=(P_n/P_{n-1}), n\ge 1$. It follows that: (a) for a given n we have $\varepsilon_n>(<)1$, iff $P_n>(<)P_{n-1}$; (b) $\{\varepsilon_n\}$ increases (decreases), iff $\{P_k\}$ is log-downward (log-upward) convex. Thus, one may reformulate the Property 2 in terms of $\{\varepsilon_n\}$.

Property 2*. $\{\varepsilon_n\}$ is formed by no more than three monotone pieces, where the last one increases and is located under the straight line *y*=1.

Now assume that the Property 2^* holds for the FD $\{P_k\}$. We are going to study the question of task on convergence of series (see (2.2)), i.e.

$$\sum_{n\geq 1}\prod_{k=1}^{n}\varepsilon_k<+\infty \tag{7.1}$$

for FDs of *moderate growth*, which means that $\lim\limits_{n\to\infty}\left(P_n/P_{n-1}\right)=1$.In terms of $\{\varepsilon_n\}$ it is equivalent to the existence of limit

$$\lim_{n\to\infty}\varepsilon_n=1. \tag{7.2}$$

Any regularly varying FD$\{P_n\}$, due to Property 1, is a FD of *moderate growth*.

To obtain a *sufficient* condition for the validity of (7.1) one can use the *Kummer's Test* (see 3.37, p. 116–117 [33]). Namely, let D_n be a sequence of positive numbers such that the *positive* limit (finite or infinite) exists $\lim\limits_{n\to\infty}\left(D_{n-1}\varepsilon_n^{-1}-D_n\right)$. Then (7.1) holds.

Due to (7.2), one may replace the last limit by the following one

$$\lim_{n\to\infty}\left(D_{n-1}-\varepsilon_n D_n\right)>0. \tag{7.3}$$

It can be easily seen that (7.3) holds, if

$$\varepsilon_n=1-\frac{\rho}{n^{\alpha}}+o\left(\frac{1}{n^{\alpha}}\right),\, n\to+\infty,\, \rho\in R^{+},\, \alpha\in(0,1)\ \text{(we take } D_n=n^{\alpha}\text{)}, \tag{7.4}$$

$$\varepsilon_n=1-\frac{\rho}{n}+o\left(\frac{1}{n}\right),\, n\to+\infty, 1<\rho<+\infty\ \text{ (we take } D_n=n\text{)}. \tag{7.5}$$

This test was used in [34] in order to prove (7.1) for

$$\varepsilon_n=1-\frac{1}{n}-\frac{1}{n\ln n}-....-\frac{1}{n\ln n\ldots\underbrace{\ln\ln\ldots\ln n}_{K}}-\frac{\rho}{n\ln n\ldots\underbrace{\ln\ln\ldots\ln n}_{K+1}}\left(1+o(1)\right),\quad n\to+\infty,$$

where $K\geq 0$ is a *natural* number and $\rho\in(1,+\infty)$.

All considered cases of $\{\varepsilon_n\}$'s asymptotic behavior are examples of FD $\{P_k\}$ of *moderate growth*. Now let's discuss the Property 1.

Theorem 10 (see [20]).

1) The condition (7.5) implies the regular variation of $\{P_k\}$ with exponent ($-\rho$).

2) Under the condition (7.5) the existence of ACSVC for $\{P_k\}$ is equivalent to the limit relation

$$\lim_{n\to+\infty}\sum_{k\geq n}\left(1-\varepsilon_k-\frac{\rho}{k}\right)=0. \tag{7.6}$$

Let the following asymptotic expansion with $\alpha\in\left(\frac{1}{2},1\right)$ holds for $\{P_n\}$:

$$P_n=\frac{L}{n^{\rho}}+\frac{M}{n^{\rho+\alpha}}+o\left(\frac{1}{n^{\rho+1}}\right),\quad n\to+\infty,\ L\in R^1,\quad M\in R^1\backslash\{0\}, \tag{7.7}$$

which in particular implies that $\{P_n\}$ varies regularly at infinity with exponent ($-\rho$) and exhibits ACSVC *L*, i.e. the Property 1 takes place. We easily verify that

$$\varepsilon_n=\left(1-\frac{1}{n}\right)^{\rho}\cdot\frac{1+Rn^{-\alpha}+o\left(n^{-1}\right)}{1+R\left(n-1\right)^{-\alpha}+o\left(n^{-1}\right)}=$$

$$=\left(1-\frac{\rho}{n}+o\left(\frac{1}{n}\right)\right)\left(1+\frac{R}{n^{\alpha}}+o\left(\frac{1}{n}\right)\right)\left(1-\frac{R}{n^{\alpha}}+o\left(\frac{1}{n}\right)\right)=1-\frac{\rho}{n}+o\left(\frac{1}{n}\right),\quad n\to+\infty,$$

where $R=\left(M/L\right)$. Various forms of $\{P_k\}$'s asymptotic expansions, similar to (7.7), may lead to the form (7.4) for $\{\varepsilon_n\}$, which allows us to formulate the reverse to statement (1) in Theorem 10.

In particular, let's replace the Property 1 by more strong but also natural

Improved Property 1. For FD $\{P_k\}$ the asymptotic expansion (4.4) holds.

The Improved Property 1 implies the asymptotic expansion (7.5). Note that for all presented above FDs the Improved Property 1 takes place.

Several relationships between $\{P_k\}$'s and $\{\varepsilon_n\}$'s asymptotic expansions are obtained in [31, 35].

For investigation of the properties of FDs $\{P_k\}$ it is natural to formulate them in terms of $\{\varepsilon_n\}$.

*Property 1**. $\{\varepsilon_n\}$ satisfies asymptotic expansion (7.5).

8. Continuity by Parameters. The *simplest* form of empirical *fact 4* with respect to parameters of FDs in mathematical sense is the *continuity* of $\{P_n\}$ by parameters. Sometimes FDs are given in the form of their *Generating Functions* (GFs).

Let us consider the *finite-parametric* FD $\{P_n(\vec{c}_m)\}$. The parameters are $c_1,\ldots,c_m$ and $\vec{c}_m=\left(c_1,\ldots,c_m\right)$. The GF of FD $\left\{P_n\left(\vec{c}_m\right)\right\}$ is defined as follows: for any $x\in\left[0,1\right]$

$$P\left(x,\vec{c}_m\right)=\sum_{n\geq0}P_n\left(\vec{c}_m\right)x^n\,. \tag{8.1}$$

Having GF (8.1), it is possible to establish the *continuity* of $\{P_n(\vec{c}_m)\}$ by parameters $c_1,\ldots,c_m$ with the help of *Continuity Theorem* for GF (see XI.6, p. 262 [36]).

Continuity Theorem. Let $\left\{P_n^{(k)}\right\}$ be a sequence of FDs. Then, in order $P_n^{(k)}\to P_n$ as $n\to+\infty$ for fixed *n* it is necessary and sufficient the following convergence:

$$P_k\left(x\right)=\sum_{n\geq0}P_n^{(k)}x^n\to\sum_{n\geq0}P_nx^n=P\left(x\right)\ \text{as}\ k\to+\infty\ \text{for any}\ x\in\left[0,1\right].$$

This idea has been developed in [22], p. 33–36, for famous FDs. For instance, in case of FD (1.3) with the help of hypergeometric series (see 9.100, p.1040, [37]) and its integral representation (see 9.111, p. 1040, [37]) we have for GF of (1.3):

$$P(x,p,q)=(q-p)\int_0^1(1-t)^{q-1}(1-tx)^{-p}\,dt\,. \tag{8.2}$$

Due to *Continuity Theorem*, if $p\to p'$, $q\to q'$, then the GF (8.2) with parameters p and q tends to GF (8.2) with parameters p' and q', if it is possible to pass to the limit under the sign of integral, which is the case in this situation.

In some cases the integral in (8.2) may be evaluated in a closed form. Due to 9.121.24, p. 1041 [37], for $p=1/2$, $q=1$ one may obtain

$$P\left(x,\frac{1}{2},1\right)=\frac{1}{1+\sqrt{1-x}},\quad x\in[0,1]\ \text{(see also [38]).}$$

9. Stability by Parameters. Let $\{P_n(\vec{c}_m)\}$ be *m*-parametric FD with $\vec{c}_m\in\Omega\subset R^m$. In general, the stability *property* by parameters is formulated as follows:

$$\{P_n(\vec{c}_m)\}\ \text{is stable with respect to parameters}\ c_1,\dots,c_m$$
$$\text{in terms of some classical metric, say}\ \rho. \tag{9.1}$$

Here explanations are needed. All *non-trivial* metrics in R^m are equivalent to the following one $\sum_{k=1}^{m}\left|c_k-c_k^{'}\right|=\left|\vec{c}_m-\vec{c}_m^{\,'}\right|$, where $\vec{c}_m=(c_1,\dots,c_m)\in\Omega$, $\vec{c}_m^{\,'}=\left(c_1^{'},\dots,c_m^{'}\right)\in\Omega$. In the set of sequences $\{P_n(\vec{c}_m)\}$ with different collections of parameters $\vec{c}_m$ one has to introduce some *classical* metric $\rho\left(\{P_n(\vec{c}_m)\},\{P_n(\vec{c}_m^{\,'})\}\right)$ "suitable" to $\{P_n(\vec{c}_m)\}$. For simplicity we write $\rho\left(\vec{c}_m,\vec{c}_m^{\,'}\right)$ instead of $\rho\left(\{P_n(\vec{c}_m)\},\{P_n(\vec{c}_m^{\,'})\}\right)$. Below K is a *convex compact* in Ω.

Definition 4. We say that FD $\{P_n(\vec{c}_m)\}$ is ρ-stable with respect to $c_1,\dots,c_m$, if for any $K\subset\Omega$

$$\lim_{|\vec{c}_m-\vec{c}_m^{\,'}|\to 0}\rho\left(\vec{c}_m,\vec{c}_m^{\,'}\right)=0 \tag{9.2}$$

uniformly on $\vec{c}_m$, $\vec{c}_m^{\,'}\in K$.

Then, the empirical *fact 4* for finite-parametric FD with respect to parameters takes the mathematical form (9.2).

The form of K may be chosen *simple*, if parameters and ranges of their changes are independent. Parameters (ranges of their changes) are independent, if there are no relations of *equality* type (of *inequality* type) among them.

All presented above FDs have independent parameters. But for FDs (1.3) and (3.1) the independence of parameters' changes ranges doesn't take place. The situation can be improved with the help of *linear* transformations:

$\rho=q+1-p,\ p=p$ for FD(1.3); $\rho=q+1-\hat{p}_1-\hat{p}_2,\ \hat{p}_1=\hat{p}_1,\ \hat{p}_2=\hat{p}_2$ for FD(3.1).

Now, for the FD $\{P_n(\vec{c}_m)\}$ with independent parameters and independent ranges of their changes one may choose K in the form

$$K=\prod_{i=1}^{m}\left[\underline{c}_i,\bar{c}^i\right], \tag{9.3}$$

where $\left(\underline{c}_1,\ldots,\underline{c}_m\right)\in\Omega$, $\left(c_1',\ldots,c_m'\right)\in\Omega$, $\underline{c}_i\le\bar{c}_i$ for $i=\overline{1,m}$.

The following *metrics* for $\left\{P_n\left(\vec{c}_m\right)\right\}$ in bioinformatics are usually used [12], [22]:

$$\delta\left(\vec{c}_m,\vec{c}_m^{\,'}\right)=\sup_{n\ge0}\left|\sum_{k=0}^{n}\left(P_k\left(\vec{c}_m\right)-P_k\left(\vec{c}_m^{\,'}\right)\right)\right|, \quad (\text{Uniform Metric})$$

$$\varepsilon\left(\vec{c}_m,\vec{c}_m^{\,'}\right)=\sum_{n\ge0}\left|P_n\left(\vec{c}_m\right)-P_n\left(\vec{c}_m^{\,'}\right)\right| \qquad (\text{Metric in Variation}).$$

The last one is the particular case with p=1 of l_p*-metrics,* and, obviously, $\delta\left(\vec{c}_m,\vec{c}_m^{\,'}\right)\le\varepsilon\left(\vec{c}_m,\vec{c}_m^{\,'}\right)$. Thus, if $\left\{P_n\left(\vec{c}_m\right)\right\}$ is ε-stable, then it is δ-stable too.

Generally speaking, the *reverse* statement to the last one is not true. But for the FD (6.3) under the conditions of Theorem 8 it can be proved that $\delta(c,c')=1/2\varepsilon(c,c')$ (see [12], Chapter 4).

The stability problems for introduced FDs in terms of ε- and δ-metrics are in the center of attention of several publications [39–43].

It is of interest the following *Stability Criterion* for $\left\{P_n\left(\vec{c}_m\right)\right\}$ in terms of l_p*-metrics* (see [22, 43]). We assume the following conditions hold:

1. The FD $\left\{P_n\left(\vec{c}_m\right)\right\}$ allows a representation in the form $P_n\left(\vec{c}_m\right)=\left(g_n\left(\vec{c}_m\right)/g\left(\vec{c}_m\right)\right)$, $g_n\left(\vec{c}_m\right)\ge0$, $n\ge0$.

2. There is $\vec{c}_{m+}\in K$ such that for all $n\ge0$ we have $g_n\left(\vec{c}_{m+}\right)=\max\limits_{\vec{c}_m\in K}g_n\left(\vec{c}_m\right)$.

3. There is $\vec{c}_{m-}\in K$ such that $\vec{c}_{m-}=\min\limits_{\vec{c}_m\in K}g\left(\vec{c}_m\right)$.

4. The FD $\left\{P_n\left(\vec{c}_m\right)\right\}$ satisfies Property 1.

Let $-\rho=-\rho\left(\vec{c}_m\right)$ be the *exponent* of $\left\{P_n\left(\vec{c}_m\right)\right\}$'s regular variation and $p>\left(1/\rho\left(\vec{c}_{m+}\right)\right)$.

Theorem 11. $\left\{P_n\left(\vec{c}_m\right)\right\}$ is l_p-stable on K, iff

$\lim\limits_{\left|\vec{c}_m-\vec{c}_m^{\,'}\right|\to0}\left|g_n\left(\vec{c}_m\right)-g_n\left(\vec{c}_m^{\,'}\right)\right|=0$ uniformly on $\vec{c}_m$, $\vec{c}_m^{\,'}\in K$ for every $n\ge0$.

Theorem 11 was applied to FDs (3.1) and (3.2) in order to prove for them the l_p-stability by parameters.

Another approach to stability of finite-parametric FDs has been suggested in [44]. It is based on *monotonicity* property of functions in case, when FD may be presented in the form of such functions combined by finite number of operations of finite or infinite sums, product, ratio, convolution.

10. Semi-Group Property. We already mentioned that the *Power Law* (1.1) is of interest in *self-organized* growing biomolecular networks, because of its *scale-invariant* property. Trying to figure out other FDs for application here one has to analyze the *properties* of such networks. Together with the self-organization there is the *second* peculiarity. The FD must be of the *same* type in united interval as it is in each fractal forming the interval in order to *extrapolate* the FD in united interval and in *whole* system. The fractals may be chosen with *approximately equal lengths* in the way, which allows to postulate either the *independence* or some type of "*weak*" *dependence* between the numbers of event's occurrences on each fractal. These random numbers are characterized by local FDs on fractals. Now, instead of *scale-invariance* the *semi-group* property has to take place. In contradiction to *scale-invariance* property, where the operation of *multiplication* is used, the *semi-group* property implies that the convolution of FDs of the *same* type equals to FD of *exactly* this type. Such *semi-group* property is *intrinsic* for normal, Cauchy's, Levy's distribution functions and for many other very useful ones. The *semi-group* property holds, for instance, for the *four-parametric* family of *Stable Laws* (see [45, 46]). Moreover, the conception of *regular variation* and the *semi-group* property for empirical FDs` continuous analogs are *closely connected* and *supplement each other* from the point of view of Probability Theory. It is just the time to notice that *Stable Laws* not only satisfy *semi-group* property, but also Property 1.

Below we introduce more powerful than the *semi-group* property, and, obviously, more *restrictable* property, which extracts the family of *Stable Laws.*

Definition 5. We say that distribution function S is stable, if for any $a_i \in R^1$, $b_i \in R^+$, $i = 1,2$, there are numbers $a \in R^1$, $b \in R^+$ such that

$$S\left(\frac{x-a_1}{b_1}\right) * S\left(\frac{x-a_2}{b_2}\right) = S\left(\frac{x-a}{b}\right), \quad x \in R^1,$$

where * denotes the sing of convolution.

Let us describe the *parameters* of Stable Laws. The first *essential* parameter $\alpha \in (0,2]$ is the *exponent,* which defines the *exponent* $(-\rho)$ of *Stable Law* density's regular variation $\rho = \alpha + 1$.

Excluding Normal Law $\alpha = 2$any Stable Law $\alpha \in (0,2)$ has infinite variance.

Denoting by S_α the *Stable Law* with exponent $\alpha \in (0,2)$ consider its two *tails*: $S_\alpha(-x)$ (left tail) and $1 - S_\alpha(x)$ (right tail) for $x \in R^+$. The *second essential* parameter for S_α is *asymmetry*, i.e. the value of limit

$$\beta = \lim_{x\to\infty} \frac{1 - S_\alpha(x) - S_\alpha(-x)}{1 - S_\alpha(x) + S_\alpha(-x)} \in [-1,1]$$

(the ratio of the tails difference and sum), which *always exist*. In other words, the *asymmetry* is nothing else, but the measure of skewness f*or* $S_\alpha(x)$. Due to empirical *fact 1*, we are interested in *Stable Laws* with maximal skewness to the right, i.e. in $S(x)$ with $\beta = +1$.

The remained two parameters (shifting parameter and scale factor) are *non-essential.*

The next condition, which has to be fulfilled, if we want to use *Stable Laws* in bioinformatics, consists in following. The extracted densities of *Stable Laws*, which assumed to be continuous analogs of FDs, must be concentrated in $[0,+\infty)$.

Denote by $S(x;\alpha,\beta)$ a stable density with exponent α and asymmetry β .

For our purposes we may use not only $S(x;\alpha,\beta), 0<\alpha<1$ (only in this case $S(x;\alpha,1)$ is concentrated on $[0,+\infty)$), but also $2\cdot S(x;\alpha,0), 0<\alpha<2,$ for $x\in[0,+\infty)$.

The density $S(x;\alpha,0)$ for $x\in R^1$ is *symmetric,* so, $2\cdot S(x;\alpha,0)$ for $x\in[0,+\infty)$ is concentrated on $x\in[0,+\infty)$ and has skewness to the right.

Now, the following families of *two-parametric* densities

$$\begin{cases}\left\{\hat{f}_{\alpha,\sigma}(x)=\sigma^{-1/\alpha}S\left(x\cdot\sigma^{-1/\alpha};\alpha,1\right),\ 0<\alpha<1,\ \sigma\in R^+\right\},\\ \left\{f_{\alpha,\sigma}(x)=2\sigma^{-1/\alpha}S\left(x\cdot\sigma^{-1/\alpha};\alpha,0\right),\ 0<\alpha<2,\ \sigma\in R^+\right\}\end{cases} \tag{10.1}$$

are *condidates* to be *continuous analogs* of FDs (see [22]).

Finally, note that the densities (10.1) are formed by no more than three convex pieces and are unimodal (see Property 2).

11. Disc retization of Densities. Besides the way of *new* FDs construction based on standard birth-death process with various forms of intensities, there is a couple of other known ways. The first one consists on construction based on *discretization* of densities, which are concentrated on $[0,+\infty)$ and satisfy Properties 1 and 2. We already have such an example: two-parametric *Stable Densities* (10.1).

Let $f(t),\ t\in[0,+\infty)$, be a *continuous* density satisfying Properties 1 and 2.

Definition 6 (see [22]). We say that FD $\{P_n\}$ of the type

$$P_n=\int\limits_n^{n+1} f(t)dt,\ n\geq 0, \tag{11.1}$$

is the discretization of *f*.

It is very important that the discretization conserves the properties of monotonicity, convexity, unimodality of the density $f(t)$, i.e. at least the Properties 1 and 2 hold. It remains only to verify the Property 3.

Theorem 12 (see [41]). The discretizations of type (11.1) of densities (10.1) are stable with respect to parameters α and σ in terms of Metric in Variation.

A slightly different form of *discretization* of *Stable Densities* was used in publications [48–50].

$$P_n=\frac{f(n)}{\sum\limits_{K\geq 0} f(K)},\ n\geq 0, \tag{11.2}$$

where $f(t)$ presents the corresponding *Stable Density.*

Unfortunately, the *closed* form of *Stable Densities* is possible to obtain only for normal, Cauchy and Levy Laws. For others there are only representations in the form of convergent series [45, 46]. That is why above introduced types of *discretizations* for *Stable Densities* lead to complex expressions. At the same time, the *Laplace Transform* of *Stable Density* with asymmetry $\beta=1$, due to Theorem 3.1, p. 43 [46], always exists

$$\rho_\alpha(s)=\int\limits_{-\infty}^{+\infty} e^{-sx}dS_\alpha(x)=\begin{cases}\exp\left(-s^\alpha\right) \text{ for } 0<\alpha<2, \alpha\neq 1,\ s\in R^+,\\ \exp\left(-s+s\log s\right) \text{ for } \alpha=1, s\in R^+.\end{cases} \tag{11.3}$$

Here only one representative with given parameters of shifting and scaling are taken. Also for any *Stable Density* the right-side *Laplace Transform* has a *closed* form. We may demonstrate how it is possible to build FDs with the help of *Laplace Transform.* Let

$$\rho(s)=\int\limits_0^\infty e^{-sx}f(x)dx,\quad s\geq 0, \tag{11.4}$$

be the *Laplace Transform* of continuous on $[0,+\infty)$ density $f(x)>0$, which is concentrated on $[0,+\infty)$. It is easy to prove the following statement.

Lemma 4. The function $\rho(1-z),\ 0\leq z<1$, presents the GF of some FD.

Note that Lemma 4 is true, even if the lower limit in integral (11.4) equals to some number $-\alpha$, where $\alpha\in R^+$, and $f(x)>0$ is concentrated on $[-\alpha,+\infty)$.

Due to Lemma 4, one may suggest a new type of discretization, in particular, based on *Laplace Transform* of *Stable Densities* (see (11.3) too).

From 1960s, after the appearance of a series of papers by Mandelbrot and his successors, who sketched the use of *Stable Laws* in Economics and Biology, it comes out that *Stable Laws* have to be attached to *Special Functions* of Mathematical Analysis. Several *Special Functions* are connected with *Stable Densities* [51]. For instance, let

$$E_\sigma(X)=\sum_{n\geq 0}\frac{X^n}{\Gamma(n\sigma+1)},\ \sigma\in R^+,\ \text{(see [45])}$$

be the *Mittag-Leffler* function, where $\Gamma(\cdot)$ denotes the Euler's *Gamma-Function.* Then (see, for instance, [46], p. 169)

$$\alpha E_\alpha(-s)=\int\limits_0^\infty e^{-sx}X^{1-1/\alpha}S(X^{-1/\alpha};\alpha,1)dX,\quad s\geq 0.$$

12. Method of Special Functions. The way of FDs construction based on different forms of *discretizations* of either *Stable Densities,* or their *Laplace Transforms* may be referred as a *variation* of Method of Special Functions.

There are other ideas, whose realizations can be interpreted as variations of Method of Special Functions. For instance, we search various Special Functions of Mathematical Analysis, which have representations in the form of positive convergent series and also Integral Representations. Then, forming the ratios of the n-th term and the sum we construct the probability P_n for the FD $\{P_n\}$. In this way the Warring, Hypergeometric, Pareto FDs and many other useful ones may be obtained. Let us illustrate this way on simple example of Warring Distribution (1.3).

Consider the *hypergeometric* series (see 9.100, p. 1039, [37]), which is a Special Function:

$$F(\alpha,\beta,\gamma,z)=1+\sum_{n\geq 0}\frac{(\alpha)_n(\beta)_n}{(\gamma)_n(n+1)}z^n \tag{12.1}$$

for *positive* values of arguments, where $(x)_n = x(x+1)\cdots(x+n),\ n \geq 0.$ The series (12.1) is convergent in the following cases (see 9.102, p. 1040, [37]):

$(a)\ 0 < z < 1;\quad (b)\ 1 \leq z < +\infty,\ \ \alpha + \beta - \gamma < 0;\quad (c)\ \mathrm{z} = 1;\ \ \alpha - \beta + \gamma \geq 1.$

The following integral representation holds for $\gamma > \beta > 0$ (see 9.111, p.1040, [37]):

$$F(\alpha,\beta,\gamma,z) = \frac{1}{B(\beta,\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tz)^{-\alpha}dt, \qquad (12.2)$$

where $B(x,y)$ denotes the *Beta Function.*

$P_n,\ n \geq 1$, of the form (1.3) means that, due to $\sum_{n\geq 0} P_n = 1$, we have $(P_0)^{-1} = F(p,1,q+1,1) = \dfrac{1}{B(1,q)}\displaystyle\int_0^1 (1-t)^{q-p-1}dt = (1-\frac{p}{q})^{-1}$, i.e. $P_0 = 1-(p/q)$, where (12.1) and (12.2) were used. In our case the conditions (b) and $\gamma > \beta$ hold.

One more way for FDs construction, which may be characterized as an *addition* to Method of Special Functions, consists in following. Let the FD $\{P_n\}$ satisfies the general representation (2.3), (2.2) and Properties 1 and 2. We present (2.3) in the form

$$P_n = P_0\exp\left\{\sum_{k=1}^{n}\log(1-\delta_n)\right\},\ \ \delta_n = 1-\varepsilon_n,\ n \geq 1, \qquad (12.3)$$

where, due to Properties 1, 2 and their various improvements in terms of $\{\varepsilon_n\}$, we have $\lim_{n\to\infty}\delta_n = 1$, $\{\delta_n\}$ is monotone starting from some index n_0, etc. So, $\{\delta_n\}$, in particular, is a slowly varying sequence possessing "good" properties. We use the way of replacement of sums in (12.3) by integrals

$$\int_0^t \log(1-\delta(u))du,\ t \in [1,+\infty), \qquad (12.4)$$

which doesn't change the *qualitative* behavior of distributions. In this way one has to choose the interpolation $\delta(t)$ for the sequence $\{\delta_n\}$. For instance, we may use the following statement (see Theorem 1, p. 55, [28]).

Lemma 5. There is a slowly varying function $\delta(t)$ such that:

(a) $\delta(n) = \delta_n,\quad n \geq 1;$

(b) *δ(t)* is infinite differentiable;

(c) $\lim_{t\to+\infty}(\delta(t)/\delta(n)) = 1$ for $t \in [n,n+1]$;

(d) $\delta(t)$ is monotone, if $\{\delta_n\}$ is monotone;

(e) $\delta(t)$ is log - downward convex, if $\{\delta_n\}$ is log - downward convex.

By this operation, which is called a *dediscretization*, from (12.3) we come to a "smooth" probability density, defined on $[0,+\infty)$:

$$f(t) = f(0)\cdot\exp\left\{\int_0^t \log(1-\delta(u))du\right\}, \qquad (12.5)$$

where $f(0)$ may be obtained from the following equality

$$\int_0^\infty f(t)dt = 1, \text{ or } f\left(0\right) = \left(\int_0^\infty \exp\int_0^t \log(1-\delta(u))dt\right)^{-1}.$$

At the next step we choose various forms of $\delta\left(t\right)$, for which the integral (12.4) is possible to evaluate and get closed expressions for it.

Finally, any of many variations of the reverse operation, i.e. discretization leads to *new* FDs.

The manner of dediscretization has been introduced and developed in [22, 52] on example of distributions of moderate growth. The general approach for FDs of the form (2.3), (2.2) is presented in [53].

Received 02.06.2009

REFERENCES

1. **Apic G., Gough J., Teichmann S.A.** J. Mol. Biol., 2001, v. 301, № 2, p. 311–325.
2. **Jeong H., Tombor B., Albert R., Ottval Z.N.** Nature, 2000, v. 407, № 6804, p. 651–654.
3. **Rzhetsky A., Gomes S.M.** Bioinformatics, 2001, v. 17, № 10, p. 988–996.
4. **Wagner A., Fell D.A.** Proc. Roy. Soc. London, B 268, 2001, № 1478, p. 1803–1810.
5. **Wolf Y.I., Kalev G., Koonin E.V.** BioEssays, 2002, v. 24, № 2, p. 105–109.
6. **Wuchty S.** Mol. Biol. Evol., 2001, v. 18, № 9, p. 1694–1702.
7. **Ramsden J.J., Vohradsky J.** Phys.Rev., E 56, 1998, № 6, p. 7777–7780.
8. **Kuznetsov V.A.** EURASIP, J. Apll. Signal Process., 2001, № 4, p. 285–296.
9. **Kuznetsov V.A.** Statistics of the Number of Transcripts and Protein Sequences encoded in Genome. In: W.Zhang, I.Shmulevich (Eds.) Computational and Statistical Methods to Genomics. Boston: Kluwer, 2002, p. 125–171.
10. **Kauffman S.A.** The Origins of Order: Self-Organization and Selection in Evolution. New York: Oxford Univ. Press, 1993.
11. **Saaty T.** Elements of Queuing Theory. Dower Publications, 1983.
12. **Astola J., Danielian E.** Tampere: TICSP Series, № 27, 2004, p. 1–94.
13. **Yule J.U.** Trans. Roy. Soc. London, B 213, 1924, p. 21–87.
14. **Simon H.A.** Biometrica, 1955, v. 42, p. 425–440.
15. **Glanzel W., Schubert A.** Inform Process. Manager, 1995, v. 31, № 1, p. 69–80.
16. **Bornholdt S., Ebel H.** Phys. Rev., E 64 (3–2), 2001, p. 035104(4).
17. **Oluic-Vicovic V.** J. Am. Soc. Inform. Sci., 1998, v. 49, № 10, p. 867–880.
18. **Kuznetsov V.A.** Signal Process., 2003, v. 83, № 4, p. 889–910.
19. **Danielian E., Astola J.** Facta Universitatis (NiŠ), 2004, v. 17, p. 405–419.
20. **Danielian E.A., Avagyan G.P.** Matem. v Visshey Shcole, 2008, № 4(4), p. 17–23 (in Russian).
21. **Danielian E., Astola J.** Tampere: TICSP Series № 34, p. 127–132.
22. **Astola J., Danielian E.** Tampere: TICSP Series № 31, 2006, p. 1–251.
23. **Arakelian A.G.** The Stability of Frequency Distributions in Biomolecular Models: The abstract of Ph.D. Thesis. Yerevan: YSU, 2007 (in Russian).
24. **Yakovlev S.P.** The Analysis of Analytic Properties Multi-dimensional Frequency Distributions: The abstract of Ph.D. Thesis. Yerevan: SIUA, 2008 (in Russian).
25. **Avagyan G.P.** The Analysis of Properties of Regularly Varying Distributions: The abstract of Ph.D. Thesis. Yerevan: SIUA, 2009 (in Russian).
26. **Seneta E.** Regularly Varying Functions. Lecture Notes in Mathematics. Springer–Verlag, 1976.
27. **Bingham N.H., Goldie C.M., Tiegels J.L.** Regular Variation. Cambridge Univ. Press, 1986.
28. **Danielian E.** Tampere: TICSP Series № 12, 2001, p. 1–80.
29. **Arutyunian G.S., Yakovlev S.P.** Matem. v Visshey Shcole, 2008, № 4 (2, 3), p. 60–63 (in Russian).
30. **Arakelyan A.H., Mehrdy K.A.** Matem. v Visshey Shcole, 2007, № 3 (1), p. 5–13 (in Russian).
31. **Danielian E.A., Avagyan G.P.** Doklady NAN Armenii, 2009, № 109 (1), p. 21–31.

32. **Astola J. Danielian E.** Facta Universitatis (NiS), 2006, v. 19, p. 109–131.
33. **Thomson B.S., Bruckner J.B., Bruckner A.M.** Elementary Real Analysis. Upper Saddle River–New Jersey: Pentice–Hall, 2001, 677 p.
34. **Arzumanyan S.K.** Vestnik GIUA, 2009, № 12/2, p. 34–40 (in Russian).
35. **Avagyan G.P.** Inform. Techn. And Control, 2009, № 1, p. 8–18 (in Russian).
36. **Feller W.** An Introduction to Probability Theory and its Applications. V. 1. 1st edition. John Wiley and Sons, 1957.
37. **Gransteyn I.S., Ryznik I.M.** Table of Integrals, Series and Products. New York and London: Academic Press, 1965.
38. **Danielian E., Astola J.** On Generating Functions of Pareto and Warring Distributions. M.: In Bregovic R., Gotchev A (eds.) TICSP Workshop on Spectral Methods and Multirate Signal Proc., 2007, v. 37, p. 235–237.
39. **Arakelyan A.H.** Matem. v Visshey Shcole, 2006, № 2(1), p. 5–10 (in Russian).
40. **Arakelyan A.H.** Matem. v Visshey Shcole, 2006, № 2(2), p. 70–75 (in Russian).
41. **Astola J., Danielian E.A., Arakelyan A.H.** Doklady NAN RA, 2008, 108(2), p. 99–109 (in Russian).
42. **Yakovlev S.P.** Proc. of Engineer. Ac. of Armenia, 2007, № 4(3), p. 26–33.
43. **Yakovlev S.P.** Modelirovanie, Optimizatsia, Upravlenie. Yerevan: SIUA, 2008, № 11(1), p. 139–144 (in Russian).
44. **Astola J., Danielian E.A., Yakovlev S.** Proc. of Engineer. Ac. of Armenia, 2007, № 2, p. 265–272.
45. **Feller W.** Introduction to Probability Theory and its Applications. V. 2. 1st edition. John Wiley and Sons, 1966.
46. **Zolotarev V.M.** Transl. of Math. Monographs., Amer. Math. Soc., 1980, v. 65.
47. **Astola J., Danielian E.A., Arakelyan A.H.** Doklady NAN RA, 2007, № 107(1), p. 26–36.
48. **Farbod D.** Far East Journal of Theoretical Statistic. India, 2008, v. 26(1), p. 121–128.
49. **Farbod D., Gasparian K.V.** Matem. v Visshey Shcole, 2009, v. 5(1), p. 50–54 (in Russian).
50. **Farbod D., Gasparian K.V.** Statistica Bologna, 2008, v. 68, № 3-4, p. 134–140 (in Russian).
51. **Zolotarev B.M.** TVP. M., 1995, № 39, p. 354–362 (in Russian).
52. **Astola J., Danielian E.** Facta Universitatis (NiS), 2007, v. 20, № 2, p. 119–146.
53. **Arzumanyan S.K.** Vestnik-77 GIUA (Politechnik): Sb. Nauchn. i Metod. Statey. Part 1, 2010, v. 1 (in Press).

Proc. of the Yerevan State Univ. Phys. and Mathem. Sci., 2010, № 3, p. 3–22.

Յա. Աստոլա, Է. Ա. Դանիելյան, Ս. Կ. Արզումանյան

Հաճախականային բաշխումները կենսաինֆորմատիկայում. զարգացումը

Մեծ չափերի կենսամոլեկուլային հաջորդականությունների մաթեմատիկական ուսումնասիրությունը կատարվում է այդ հաջորդականություններում առաջացող պատահույթների վերլուծության օգնությամբ: Ակնարկը նվիրված է այդ բնագավառում ստացված արդյունքների քննարկմանը: Բոլոր հաճախականային բաշխումների համար ճշմարիտ էմպիրիկ փաստերի հիման վրա քննարկվում է Յա. Աստոլայի և Է. Դանիելյանի կողմից առաջարկված աքսիոմատիկան: Վերջինս ընդունում է ասիմպտոտիկորեն հաստատուն (դանդաղ փոփոխվող) բաղադրիչով հաճախականային բաշխման կանոնավոր փոփոխումը, այդ բաշխման գրաֆիկի կառուցվածքը և կայունությունն ըստ պարամետրերի:

Ստուգվում է աքսիոմատիկայի կատարումը պրակտիկայում օգտագործվող հաճախականային բաշխումների համար: Պարամետրական հաճախականային բաշխումների նոր ընտանիքների համար նկարագրված են նրանց կառուցման մեթոդներին վերաբերող արդյունքներ. ծնման և վախճանման պրոցեսի ստացիոնար բաշխումների, հատուկ ֆունկցիաների, կայուն խտությունների և այլ երևույթների օգտագործմամբ: Ձևակերպված է ըստ պարամետրերի կայունության խնդիրը, բերված են տարբեր դասական մետրիկաների տերմիններով կայունությունը հաստատող արդյունքներ: Տրված են հաճախականային բաշխումների տարբեր ընտանիքների կանոնավոր փոփոխման պայմաններ:

Я. Астола, Э. А. Даниелян, С. К. Арзуманян.

Частотные распределения в биоинформатике: развитие

Математическое изучение биомолекулярных последовательностей больших размеров осуществляется анализом свойств частотных распределений событий, возникающих в таких последовательностях. Обзор посвящен обсуждению результатов в этой области. На основе общих эмпирических фактов, полученных для всех частотных распределений, обсуждается предложенная Я. Астолой и Э.А. Даниеляном аксиоматика.

Аксиоматика постулирует правильное изменение частотного распределения с асимптотически постоянной медленно меняющейся компонентой, форму его графика и устойчивость по параметрам. Проверяется выполнимость аксиоматики для применяемых на практике частотных распределений.

Описаны результаты построения новых параметрических семейств частотных распределений следующими методами: использование стационарных распределений процесса гибели и размножения, специальных функций, устойчивых плотностей и т.д.

Сформулирована задача устойчивости по параметрам, приведены результаты установления устойчивости в терминах различных классических метрик. Для различных семейств частотных распределений даны условия правильного изменения.