VALIDATION OF RANDOM BIRTH-DEATH MODEL OF EVOLUTION OF PROTEOME COMPLEXITY

Kuznetsov V.A.
CIT/NIH & SRA International, Inc. Bethesda, MD, 20892, USA, e-mail: vk28u@nih.gov

Summary
We have shown that counting the domain-to-protein links observed in the protein and protein domain/motifs data sets and analysis of statistical distributions of these counts lead to common probabilistic model of evolution of proteome complexity of the archael, bacterial and eukaryotic organisms (Kuznetsov et al., 2002; Kuznetsov, 2003a, b). In this work, using InterPro data sets, we test several basic assumptions and predictions of our model.

Keywords: proteome complexity, evolution, domain-to-protein links, skew distributions

Introduction
Proteins contain short structural and/or functional ‘blocks’ (sequences of amino acids) called structural motifs (of length 5–35 nt) and structural domains (of length 30–250 nt) that are seen repeatedly in many proteins of all species whose genome have been examined (Kuznetsov et al., 2002; Kuznetsov, 2003a). In this work, all such protein blocks will be called, for brevity, domains. The domains are correspond to specific sequences of DNA within genes which have been evolutionarily conserved. DNA corresponding to a domain may occur multiple times in a given protein-coding gene and/or in many different protein-coding genes within a given genome, and in the genomes of many species.

If a domain \( a_i \) occurs in the protein \( b_j \), we say this constitutes a domain-to-protein link. The list of domains together with the numbers of occurrences of each domain (links) found in the proteome of an organism is called the domain-to-protein linkage profile (DPLP) of the proteome. The DPLPs allow us to characterize the domain-to-protein link network (and proteome complexity) for each organism \( S \).

Let us focus on a specific organism \( S \). Let the random variable \( X \) denote the number of occurrences of the domain \( a_i \) in all proteins of the representative proteome of \( S \). We call this number as the number of redundant domain-to-protein links of a random domain of the organism \( S \). We can define the domain occurrence probability function (DOPF) \( p'_l := \Pr(X = l) \), where \( l = 0,1,2,\ldots \). The value \( p'_l \) is the probability that a random domain occurs exactly \( l \) times within the proteome of the organism.

If protein \( b_j \) contains domain \( a_i \) at least once, we can count the number of such non-redundant occurrences (non-redundant domain-to-protein links) in the (domain, protein) network for the given domain. Let the random variable \( X \) denote the number of non-redundant domain-to-protein links of a random domain of \( S \). We also define the DOPF \( p_h := \Pr(X = h) \), where \( h = 0,1,2,\ldots \). The value \( p_h \) is the probability that a random domain occurs exactly \( h \) times within the proteome of \( S \).

Let \( A = \{ a_1, a_2, \ldots, a_N \} \) be the set of observed domains contained in the \( P \) proteins of the organism \( S \). Let \( B = \{ b_1, b_2, \ldots, b_P \} \) be the set of proteins in the representative proteome (sample of observed
proteins) for the organism $S$. The value $\hat{p}_l = \hat{h}_l / N$ ($l = 1, 2, ...$) is an empirical estimate of the probability function $p_l$, where $\hat{h}_l$ is the number of observed domains occurring exactly $l$ times in the representative proteome of $S$ and $N$ is the total number of observed domains in the representative proteome. Let $M'$ denote the total number of domain-to-protein links in the representative proteome of the organism $S$. We call the value $M'$ the redundant connectivity number of the (domain, protein) network. Note that all occurrences of a domain in the proteins of $S$ are counted.

As before we can obtain an estimate of $p_h$. The value $\hat{p}_h = \hat{h}_h / N$ ($h = 1, 2, ...$) is an empirical estimate of the probability function $p_h$, where $\hat{h}_h$ is the number of non-redundant domain-to-protein links occurring exactly $h$ times in the representative proteome of $S$. Let $M$ the total number of non-redundant domain-to-protein links in the representative proteome of the organism $S$, and we call $M$ the non-redundant connectivity number of the (domain, protein) network of $S$. Note $M \leq M'$. The values $M$ and $M'$ can be used as the non-redundant and redundant measures of the proteome complexity of an organism (Kuznetsov et al., 2002; Kuznetsov, 2003a).

Let $X^{(s)}$ be the number of non-redundant links for a random domain of the organism $S$. Let $X^{(k)}$ be the number of non-redundant links for a random domain of the other organisms. To characterize the relative non-redundant complexity of the organism $S$ versus the organism $K$, we can plot the bivariate distributions of the random variables $(X^{(s)}, X^{(k)})$.

In this work, we test the assumptions and predictions of our probabilistic model of evolution of proteome complexity based on the analysis of statistical distributions of the redundant and non-redundant domain-to-protein link networks of variety organisms. For this purpose, data sets of the 156 fully-sequenced genome organisms was downloaded from the InterPro database (www.ebi.ac.uk/interpro) and from the Proteome Analysis Database (www.ebi.ac.uk/proteome).

Hypergeometric Model of Evolution

Let the random variable $D_s(a, P)$ be the number of domain-to-protein links of the domain $a$ in the organism occurring at time $t$ in the evolutionary path $P$ of some end-point organism. The $D_s(a, P)$ is a realization of a continuous-time stochastic process $\{D_s, t \geq 0\}$. This process can be considered as a birth-death random process that has a net change across an infinitesimal time interval. Note for redundant links $\sum_{i=0}^{J} m'_i = M'$ and for non-redundant links $\sum_{i=0}^{J'} m_i = M$. Let $J$ denote the number of non-redundant links for most abundant domains found in the proteome and let $J'$ denote the number of occurrences of the most abundant domains found in the proteome. Note $J' \geq J$ and the both values $J$ and $J'$ are increasing if $M$ becomes bigger.

Let consider the case of non-redundant domain-to-protein links. Let $p_m(t) = P(D_s = m)$ ($m = 0, 1, 2, ..., J$) denote the probability function associated with the random process $\{D_s, t \geq 0\}$. Then the rate of the probability function $p_m$ ($m = 0, 1, 2, ..., J$) can be described by finite system of the forward Kolmogorov equations. For our application purposes, we consider the intensity rates of the random process $\{D_s, t \geq 0\}$ be the linear functions of the value $m$: 299
\[ \lambda_i = \lambda_i^* + \mu_i^* m \]

and
\[ \mu_i = \mu_i^* + \mu_i^* m \]

where \( \lambda_i^* > 0, \lambda_i^* > 0, \mu_i^* > 0, \mu_i^* > 0 \). Hence, during an interval \((t, t + \tau)\) where \( \tau \) is small, we assume four independent processes: the spontaneous “birth” and “death” of a domain, with constant intensities \( \lambda_i^* \) and \( \mu_i^* \), respectively, and the “flows” of the domains with the intensities proportional to the number of the links already counted for that domain \( \lambda_i^* m \) and \( \mu_i^* m \). The parameters \( \lambda_i^* \) might be associated with purification and positive selection of already used links for a random domain and \( \mu_i^* \) might be associated with loss of the links due to random constraints and negative selection forces.

We assumed that in the most evolving near end-point organisms, the random birth and death processes of protein-encoding sequences are keeping near equilibrium. This equilibrium can be parametrically described with the following recursive probability formula which we called the Kolmogorov-Waring (KW) probability function (Kuznetsov, 2003a, b).

\[ p_{m+1} = \theta \frac{(a + m)}{b + m + 1} p_m^*, \]

where \( m = 0, 1, \ldots \), and \( a, b, \) and \( \theta \) are the positive parameters of our model;

\[ p_m = \frac{1}{\binom{a + b + 1}{m} F(a; b + 1; \theta)} > 0, \text{ where } \binom{a + b + 1}{m} F(a; b + 1; \theta) \text{ is the hypergeometric Gauss function.} \]

\[ a = \lambda_i^* / \lambda_i^*; \theta = \lambda_i^* / \mu_i^*; b = \mu_i^* / \mu_i^*. \]

In order to fitting the probability function Eq.1 to the empirical DOPFs, we re-normalized this probability function due to the random values of empirical DOPFs are changed between 1 to \( J \) and estimated of the parameters \( a, b, \theta \), as described in (Kuznetsov, 2003a, b).

**Results**

Even with their large differences in proteome complexity of archael, bacterial and eukaryotic organisms all demonstrate similar skewed long-tail (Pareto-like) DOPFs (see three examples on Fig. 1A) and all data sets fit accurately by Eq.1. The shape of the empirical DOPFs, the parameter \( J \) and estimated parameters of the best-fit probability function by Eq. 1 correlate on the total number of domain-to-protein links \( M \). Note that for most of the studied organisms (with several exceptions among bacteria and archaea) estimated values of the parameter \( \theta \) equal around 1, which means that \( \lambda_i^* \approx \mu_i^* \). This results specifies our assumption that rates of the birth and death processes should be near equilibrium. However, the best-fit model predicts that for many bacterial and arcaheal organisms the equilibrium in domain-to-protein link network might be unstable and the dispersion approaches infinity.

To characterize the relative redundant complexity of the organisms \( S \) versus the organism \( K \), we used the empirical bivariate distributions of the random variables \( (X^{(s)} = m^{(s)}, X^{(i)} = m^{(i)}) \), where \( m^{(s)} = 0, 1, 2, \ldots, J^{(s)} = 0, 1, \ldots, J^{(i)} \). Figure 1B shows typical pattern of such distribution when more complex organism (Drosophila melanogaster) compares to the less complex organism (yeast): a random domain which occurs in more complex (multi-domain) proteins of
less complex organism has a higher than average chance to appear in many more complex proteins in a more complex organism, even over widely evolutionary paths. Figure 1B shows that domains represented by a larger number of non-redundant links in the yeast proteome tend to also more often in the fly proteome. We found that ‘yeast’ domains (evolutionarily “older” domains) repre-
sented in the fly proteome have, on an average, more domain-to-protein links than domains represented only in the fly proteome (domains not presented in the yeast proteome; perhaps, evolutionarily “younger” domains). These results indicate that the probability of acquisition of new non-redundant links to proteins is roughly proportional to the number of occurrences of this domain (in particular in the “younger” proteomes). This conclusion qualitatively agrees to the basic assumptions lead to Eq.1.

Figure 2A shows strong positive correlation (r=0.88; p<0.001) between the random increment of domain-to-protein links for the domains of the organism (\( \delta m = m' - m \)) and the number of non-redundant domain-to-protein links (\( m \)) in that organism. This data supports the analysis presented in Figure 1B.

Figure 2B shows the empirical conditional probability of random variable \( m | \delta m = 0 \) which is the number of links of a random domain among the domains which has no multiple occurrences \( \delta m = 0 \) in any protein. 2195 (47%) of the 4705 human domains have no multiple concurrencies. The form of histogram in Figure 2B fits well by Eq.1 (not presented), and allows us to assume that redundant occurrence of rare and/or recent domains within proteome of an organism is much less than the redundant occurrence of more abundant domains. This result is qualitatively support our simple probabilistic model of evolution of proteome complexity.

**Conclusion**

We might assume that the domain occurrence counts in a proteome during evolution is a random birth-death quasi-steady state process such that new domains are rarely appeared as a singletons and lost at constant rates, and domains are reused and lost at rates proportional to their current use. The occurrences of a given domain in the proteome is essentially random process, determined by the intrinsic properties of domain (i.e. hydrophobic group, etc.) and also the evolution history of the domain (i.e. evolution age). We need to better understand the non-random mechanisms lead to changes of the skewed distribution of domain-to-protein links and to specific domain-domain interactions in the course of evolution.

**References**

