ANALYSIS OF PROTEOME COMPLEXITY
BASED ON COUNTING DOMAIN-TO-PROTEIN LINKS

Kuznetsov V.A.*1, Pickalov V.V.2, Knott G.D.3, Kanapin A.A.4

1 CIT/NIH & SRA International, Inc. Bethesda, MD, 20892, USA, e-mail: vk28u@nih.gov; 2 Institute of Theoretical and Applied Mechanics SB RAS, Novosibirsk, Russia, e-mail: pickalov@itam.nsc.ru; 3 Civilized Software, Inc., Silver Spring, MD, 20906, USA; 4 EMBL-EBI Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, * Corresponding author: e-mail: alex@ebi.ac.uk

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

Summary

We define the list of domains to proteins, together with the numbers of their occurrences (links to proteins) found in the proteome of an organism to be the domain-to-protein linkage profile (DPLP) of the proteome. We estimated the DPLP for the 156 fully-sequenced genome organisms represented in the InterPro database. This work presents several quantitative measures of the complexity of a proteome based on the DPLP. For each of the 156 studied organisms, we found two large subsets of domains: the domains which occur two or more times in at least one protein, and the domains which are not duplicated within any protein of the proteome. The latter set of domains well reflects the increasing trend of biological complexity due to evolution. The statistical distributions of the number of domain-to-protein links in the proteome and the estimates of the differences between the DPLPs for pairs of organisms are used as measures of relative biological complexity of the organisms. In particular, we show quantitatively the greater complexity of the human proteome, relative to that of a mouse or a rat. These differences are only partially reflected in the number of protein-coding genes estimated for these species by the sequencing genome projects.

Introduction

There are no consensus quantitative measures of the relative or absolute complexity of an organism. The biological complexity of an organism should be characterized by the list of all proteins in a given proteome and their dynamic interactions among themselves (protein network), the other molecules (protein-DNA network, etc.). However, the total number of putative protein sequences based on complete genome sequence data of a given organism can be predicted only approximately, and our knowledge on dynamic interactions of proteins in an organism is also essentially incomplete. Thus, we need a more tractable and practical approaches to describing the biological complexity.

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. The entire number of domains in nature is probably a very limited number. The estimated number of such classes of homologous sequences of motifs and domains in nature ranges from 6,000 to 10,000 or so [1–3]. In this work, all such protein blocks will be called, for brevity, domains. The domains are essential to the biological function(s) of the protein in which they occur and serve as evolutionarily conserved building blocks for the forming of proteins. The domains are corresponding 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. We attempt to determine the measures of the biological complexity based on a limited set of domains.
If a domain $D_i$ occurs in the protein $P_j$, 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 DPLP should allow us to characterize the domain-to-protein link network for each organism. Currently, we do not know all domains and all proteins in a given organism. We can study the DPLP of organism by observing sample DPLPs in representative proteomes (i.e. protein sets which were specified for organisms). However, several protein domain/motif databases provide large enough samples of DPLPs for a variety of organisms. We can analyze the samples of domain-to-protein linkage information in fully-sequenced genome organisms to categorize their proteome complexities. This work presents several quantitative measures of the complexity of a proteome based on observed DPLPs that appear to be appropriate and consistent.

**Protein Domain Database Analyzer**

Our information about DPLPs of the 156 fully-sequenced genome archaeal, bacterial and eukaryotic organisms was obtained from the InterPro database (www.ebi.ac.uk/interpro, April, 2004) [5], the protein sequences for the organisms were obtained from the Proteome Analysis Database (www.ebi.ac.uk/proteome, April 2004).

We developed a Protein Domain Database Analyzer (PDDA) program which we used to access the InterPro database and download data into a local MySQL relational database [1]. The MySQL database consists of a $N_{tot} \times L$ table, where $N_{tot} = 9609$ domains; and $L = 156$ organisms. Each row corresponds to an InterPro domain and each column corresponds to an organism. The $(i, S)$-th entry of the table is the number of occurrences of the domain $i$ in the proteome of organism $S$. Information of this table was analyzed by PDDA data mining tools, which include the statistical and graphical functions, logical functions, descriptive statistics, and correlation analysis.

**Definitions**

Let us focus on a specific organism $S$. 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 $S$. We define the $N \times P$ adjacency matrix $C := [c'_{ij}]$ $(i = 1, 2, \ldots, N; j = 1, 2, \ldots, P)$ for the organism $S$ as follows: the value $c'_{ij} = k$, if protein $b_j$ contains domain $a_i$ exactly $k$ times. Note that $k \in \{0,1,2,\ldots\}$.

The value $c'_{ij}$ is the number of domain-to-protein links for the (domain, protein) pair $\{a_i, b_j\}$.

Let $m'_i := c'_{i*} = \sum_{j=1}^{P} c'_{ij}$ denote the number of occurrences of the domain $a_i$ in all proteins of the representative proteome of $S$; $i = 1, 2, \ldots, N$. We call the value $m'_i$ the number of redundant domain-to-protein links of the domain $a_i$ of $S$. Let $M'$ denote the total number of domain-to-protein links in the representative proteome of the organism $S$. $M' = \sum_{i=1}^{N} \sum_{j=1}^{P} c'_{ij}$. The value $M'$ is
called the redundant connectivity number of the (domain, protein) network. Note that all occurrences of a domain in the proteins of \( S \) are counted.

We also define the adjacency matrix \( C \), where \( c_{ij} = 1 \), if protein \( b_j \) contains domain \( a_i \) at least once, and we define \( c_{ij} = 0 \) otherwise. This matrix reflects the non-redundant structure of the domain-to-protein links in the (domain, protein) network. This matrix counts each (domain, protein) link only once. Let \( m_i = \sum_{j=1}^{P} c_{ij} \). We call the value \( m_i \) the number of non-redundant domain-to-protein links of the domain \( a_i \) of \( S \). Let \( M = \sum_{i=1}^{N} \sum_{j=1}^{P} c_{ij} \). Note \( M \leq M' \). We call the matrix \( C \) the non-redundant adjacency matrix of \( S \), and we call \( M \) the non-redundant connectivity number of the (domain, protein) network of \( S \). The values \( M \) and \( M' \) have been used as the non redundant and redundant measures of the proteome complexity of an organism [1, 4].

To quantify the complexity associated with multiple occurrences of the domain \( a_i \) in an organism, we define the redundancy value

\[
\hat{\delta}m_i = m'_i - m_i,
\]

(1) To quantify the relative complexity of the domain \( a_i \) which redundantly occurs \( m_i^{(s)} \) times in the organism \( S \) and which redundantly occurs \( m_i^{(k)} \) times in the organism \( K \), we can define the difference of redundant complexity value for the domain \( a_i \)

\[
\hat{\delta}m_i^{(s,k)} = m_i^{(s)} - m_i^{(k)}.
\]

By omitting the prim symbol in above notations, we can also introduce the difference of non redundant complexity value of the domain \( a_i \)

\[
\hat{\delta}m_i^{(s,k)} = m_i^{(s)} - m_i^{(k)}.
\]

To characterize the relative redundant complexity of the organism \( S \) versus the organism \( K \), we use the empirical bivariate distributions of the random values \( (\mu_i^{(s,k)}, \hat{\delta}m_i^{(s,k)}) \), where \( \mu_i^{(s,k)} = (m_i^{(s,k)} + m_i^{(s,k)})/2 \) is a mean value of the numbers of the links and \( i = 1, 2, ..., N^{(s,k)} \); \( N^{(s,k)} \) is the total number of domains occurred in domain-to-protein profiles of the organisms \( S \) and \( K \). By omitting the prim symbol in above notations, we obtain the empirical bivariate distribution of the random values \( (\mu_i^{(s,k)}, \hat{\delta}m_i^{(s,k)}) \), which we can use as the relative non redundant complexity measure of the organisms \( S \) and \( K \).

**Results and Discussion**

There are at least two distinguish processes of domain’s spread in nature: integration of two or more different domains in a new protein (forming non-redundant domain-to-protein links) and multiplication of a domain within the same protein (forming redundant domain-to-protein links). Panels A-C in Figures 1 show the relationships between the numbers of non-redundant and the numbers of redundant domain-to-protein links counted for the \( T.\ volcanium, E.\ coli \) and human
representative proteomes. These panels demonstrate typical patterns of the bivariate frequency distributions of the random values \( (m_i, m'_i) = \{(m_i, m'_i), \ldots, (m_{N^i}, m'_{N^i})\} \) corresponding to the numbers of redundant and non-redundant occurrences of domains in the organism S. All panels in Figure 1 show a preferential distribution of points along the diagonal; the spread of points in the orthogonal direction to the diagonal is smaller. The extreme cases (T. volcanium and human proteomes) clearly display the increased complexity of the human proteome due to the increased number of proteins which preferentially combining different domains together. Panel D in Fig. 1 shows the observed values \( (m, m') \) for the 156 organisms is a factor of 10 larger then any specific organism. We found that these distribution patterns are typical for the studied archaeal, bacterial and eukaryotic organisms. Figures 1A–D imply that increasing biological complexity in nature is mostly associated with formation of new multi-domain proteins combining different domains rater that with the increase of domain-to-protein links occurring due to increasing of multiplication of domains within proteins in which domains have been already occurred.

Fig. 1. The empirical distributions of the numbers of redundant links versus the numbers of non redundant links counted for the (A) T. volcanium, (B) E. coli and (C) human representative proteomes, and for (D) pooled data of 156 organisms. Discontinue line: linear regression; solid line: diagonal.

Fig. 2. The relative complexity of proteome of a human versus a mouse, a rat (C) and A. thalina, respectively. A: the distribution of \( (\mu, \hat{\alpha}_i^{(r,k)}); \) symbol s indicates the human organism, k = 1 indicates the mouse organism. B, C, D: the distributions of \( (\mu^{(r,k)}, \hat{\alpha}_i^{(r,k)}); \) k=2, 3, 4 indicate the mouse, the rat and A. thalina, respectively. \( i = 1,2,\ldots, N^{(r,k)} \).
For all of the 156 studied organisms, we also found two disjointed subsets of domains: the domains which occurred two or more times in at least one protein of a proteome, and the domains which was not multiplied within any protein of the proteome (see examples on Fig. 1). 4541 (47 %) of the 9609 domains occurred two or more times in a same protein of the 156 organisms. We also found that if a domain is more common in the entire proteome world, then that domain has a bigger chance to appear multiple times in a protein of any organism (see Fig. 1).

Figure 2 demonstrates how the proteome complexity of pairs of organisms can be compared. This figure shows the relative proteome complexity measure for a human versus a mouse, a rat and A. thaliana, respectively. For example, the panel A shows the bivariate distributions of differences of the number of redundant links for the human and mouse organisms, multiplied by factor 0.5 with respect to average number of the redundant links of the human and mouse organisms. This distribution is highly asymmetric about axes $x$: the number of positive differences (abundant for a human) is bigger than the number of negative differences (abundant for a mouse), and the positive highly-abundant values occur more often in the human sample than in the mouse sample. Similar asymmetric trend we observed for our human-mouse comparison in the case of redundant links (Fig. 2B). This indicates that the human organism reuses domains more frequently in same protein and in different proteins, and invents more diverse multi-domain proteins than the mouse organism. This analysis suggests that even though the numbers of non redundant protein-coding genes in a human (~33,600 genes [1, 4]) and in a mouse (~32,000 genes, by our current estimate based on the method in [1, 4]) are approximately similar, and that even the total number of domains are also similar numbers in these organisms (~5,600 for a human and ~5,100 for a mouse [4]), the proteome complexity and diversity of a significant fraction of multi-domain proteins in a human is higher than in a mouse.

Large asymmetry in the bivariate distribution of the values $(\mu_{s, k}, \delta_{s, k})$ around axes $x$ we observed for a human versus a rat (Fig. 2C). However, the human and A. thaliana proteomes show similar proteome complexity by our criteria (Fig. 2D). This is not a surprise, because even though the number of protein-coding genes in human is larger than the number of protein-coding genes in A. thaliana (~26,000-27,000 [1, 4]), relatively recent massive (and imperfect) genome duplication events in A. thaliana might dramatically increased of protein repertoire due to recombination events, which perhaps increased of the domain shuffling in proteins, but did not increase of the number of domains (~5,300 domains [4]) presented in the A. thaliana proteome.

References