Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2019, *20*, 115-123

RESEARCH ARTICLE

Domain-based Comparative Analysis of Bacterial Proteomes: Uniqueness, Interactions, and the Dark Matter

Liang Wang[1,2,*], Jianye Yang[1], Yaping Xu[1], Xue Piao[1,3] and Jichang Lv[1]

[1]*Department of Bioinformatics, School of Medical Informatics, Xuzhou Medical University, Xuzhou, Jiangsu, 221000, P.R. China;* [2]*Key Laboratory of New Drug Research and Clinical Pharmacy of Jiangsu Province, Xuzhou Medical University, Xuzhou, Jiangsu, 221000, P.R. China;* [3]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, P.R. China*

ARTICLE HISTORY

Received: October 14, 2018
Revised: November 15, 2018
Accepted: January 10, 2019

DOI:
10.2174/1389202920666190320134438

Abstract: ***Background*:** Proteins may have none, single, double, or multiple domains, while a single domain may appear in multiple proteins. Their distribution patterns may have impacts on bacterial physiology and lifestyle.

***Objective*:** This study aims to understand how domains are distributed and duplicated in bacterial proteomes, in order to better understand bacterial physiology and lifestyles.

***Methods*:** In this study, we used 16712 Hidden Markov Models to screen 944 bacterial reference proteomes *versus* a threshold E-value<0.001. The number of non-redundant domains and duplication rates of redundant domains for each species were calculated. The unique domains, if any, were also identified for each species. In addition, the properties of no-domain proteins were investigated in terms of physicochemical properties.

***Results*:** The increasing number of non-redundant domains for a bacterial proteome follows the trend of an asymptotic function. The domain duplication rate is positively correlated with proteome size and increases more rapidly. The high percentage of single-domain proteins is more associated with small proteome size. For each proteome, unique domains were also obtained. Moreover, no-domain proteins show differences with the other three groups for several physicochemical properties analysed in this study.

***Conclusion*:** The study confirmed that a low domain duplication rate and a high percentage of single-domain proteins are more likely to be associated with bacterial host-dependent or restricted niche-adapted lifestyle. In addition, the unique lifestyle and physiology were revealed based on the analysis of species-specific domains and core domain interactions or co-occurrences.

Keywords: Bacterial proteome, Hidden markov model, Pfam, Bacterial lifestyle, Domain interaction, Domain redundancy.

1. INTRODUCTION

Proteins are a group of large and complex molecules that play essential roles in the activities of organisms. As the largest experimentally-validated three-dimensional (3D) structural database of proteins, PDB is currently holding 45146 distinct protein sequences (https://www.rcsb.org/), which provides both accurate functional information and atomic coordinates for each entry item [1]. However, determination of a protein 3D structure is an expensive and time-consuming work and insights into many proteins can be derived based on sequence comparison and domain analysis automatically [2]. Proteins can be considered as consisting of domains, which are functionally and structurally independent. Through insertion, deletion, transfer, duplication, and recombination, new proteins with distinct functions are continuously formed [3, 4]. It has been previously proposed that the entire protein domain set of the proteome of a specific organism is termed as domainome, which can be used as measures of relative biological complexity of the organisms [5]. Thus, domain analysis also provides us an insight into the functional analysis of bacterial proteomes.

Currently, there are a variety of online databases focusing on the prediction and classification of protein families and domains, such as InterPro, Pfam, PROSITE, and SMART, *etc* [6]. Among the databases, Pfam hosts a collection of profile Hidden Markov Models (16712 in Pfam 31.0 release) based on UniProt reference proteomes, which can be easily downloaded from the corresponding FTP site (ftp://ftp.ebi.ac.uk/pub/databases/Pfam) and used locally to automatically screen a large set of bacterial protein sequences *via* the HMMER package [7]. It was shown that a protein sequence could have none, single, double or multiple domains through lineage-based protein domain architecture content

*Address correspondence to this author at the Department of Bioinformatics, School of Medical Informatics, Xuzhou Medical University, Xuzhou, Jiangsu, 221000, P.R. China; Tel: + 86 13921750542;
E-mail: leonwang@xzhmu.edu.cn

analysis *via* Pfam domain prediction in 14 completed green plant genomes [8]. Another study focusing on prokaryotic proteins revealed that multi-domain proteins were mainly formed through gene fusion and/or fission processes, accounting for up to 64% cases [4]. Domains with known functions can be used for predicting protein functions [9]. However, for domains with unknown functions (DUFs) and proteins without domains, domain detection and function assignment are not helpful. The difference is that DUFs are comparatively easier to locate in proteins due to their consistent appearance, while proteins with no domains exist more like dark matter in bacterial proteomes, in terms of domain-centric analysis.

Driven by the motivation of a domain-based understanding of protein evolution, we sourced 944 completely and manually annotated bacterial reference proteomes from UniProt database and systematically analysed the global landscape of domain distributions in these bacteria [10]. All proteins were classified into four groups, defined in a previous study with modifications, which are: no-domain proteins (Domain0), single-domain proteins (Domain1), two-domain proteins (Domain2), and proteins with three or more domains (Domain3) [8]. Since the number of non-redundant domains and corresponding distinct architectures are able to reflect an organism's complexity, we studied both the distribution of the four protein groups in proteomes and the number of non-redundant domains in bacterial species, wishing to get insights into domain distributions [11]. By domain redundancy, we mean that a domain is present in a proteome more than once. During the study of domain redundancy and duplication rates, it was found that some bacteria have comparatively fewer non-redundant domain numbers while domain duplication rates are abnormally high, and *vice versa*. Thus, we dived into these bacterial proteomes and identified all the co-occurrent domains with corresponding frequencies, which were then visualized through CytoScape 3.6.1 in order to understand how domains interact with each other and how the core domains differ in terms of bacterial physiology and lifestyle variations [12]. Moreover, species-specific domains in all 944 species were also identified, which may shed light on bacterial distinct physiological activities. As for the four groups of proteins, those without detectable domains are normally annotated as uncharacterized or hypothetical sequences. In order to understand whether these proteins share properties in bacterial proteomes, we analysed their physicochemical properties such as stability, aromaticity, hydrophobicity, and amino acid compositions, *etc*., through ProtParam package in Biopython [13]. General patterns were observed through our study, which may help us better understand these proteins.

2. MATERIALS AND METHODS

2.1. Collection of Bacterial Reference Proteomes and Pfam HMM Models

Bacterial proteomes were sourced from UniProt database in February 2016 by using filters of Bacteria and Reference Proteomes. Initially, 3442 bacterial proteomes were collected. Unclassified or inappropriately named bacteria were excluded from the collection, such as Acetobacteraceae bacterium and Candidatus, *etc*. In addition, only one representative bacterium for each genus was randomly selected in order to make the analysis less biased. A set of 944 bacterial proteomes was used for all domain analysis in this study. A complete list of the bacteria is available in the Supplementary Table **S1** with incorporated domain information.

2.2. Classification and Analysis of Proteins in Terms of Domains Numbers

Average domain lengths of 16712 Pfam HMM models were first compared with that of all non-redundant domains present in 944 bacterial proteomes. Sequences in each proteome were then classified into different categories in terms of domain counts, which are domain0 (no detectable domain), domain1 (only 1 detectable domain), domain2 (2 detectable domains) and domain3 (3 or more domains), based on e-value threshold of 0.001. By incorporating the factor of proteome size, we visualized the distribution patterns of the four protein categories across 944 bacterial species. We then focused on the domain level to assess how the counts of all domains, non-redundant domains, and domain duplications is distributed along bacterial proteome sizes in all the species. By domain redundancy, we mean that the presence of the same domain in a proteome is more than once. We then pooled all the protein sequences and divided them into four groups as previously defined based on domain numbers. Average lengths of the four protein groups were calculated.

2.3. Uniqueness and Interactions of Domains in Bacterial Species

Based on previous analysis of domain numbers and duplication rates in bacterial species, we searched proteomes for unique domains and calculated the number of domain occurrences in each bacterial species. After that, several representative bacterial species were selected for further study of domain interactions due to their uncommon features, such as comparatively low domain redundancy or very high number of duplicated domains, such as *Hodgkinia cicadicola* and *Actinospica robiniae*, *etc*. Bacteria with different lifestyles were also analysed in terms of domain uniqueness and interactions, such as *Deinococcus deserti*, *Yersinia pestis*, *Helicobacter pylori*, and *Anaplasma phagocytophilum*, which fell into the categories of free-living bacteria, sit-and-wait pathogens, exclusively host-associated bacteria, and vector-borne pathogens, respectively [14]. Domain associations in these bacterial proteomes were then mined *via* a Python script (available on request) in order to find out 1) which domains are co-occurring and 2) how strong these linkages are. Results were then visualized in CytoScape 3.6.1 [12]. Both core domains that are widely present across proteins and high-frequently co-occurrent domains in a proteome were identified, which were explored further for biological explanations in terms of bacterial physiology and lifestyle.

2.4. Physiochemical Analysis of Bacterial Protein Sequences

One of the protein categories, Domain0 (619297 sequences in total from 944 bacterial species), has no identified domains *via* Pfam HMM model search, most of which are annotated as hypothetical or uncharacterized sequences. In this study, we analysed these proteins on batch based on Bio.Seq and Bio.Sequtils packages from Biopython [13].

ProtParam tools were used for calculating proteins' physicochemical properties, such as stability, hydrophobicity, isoelectric points, aromaticity, and amino acid compositions, *etc*. [15]. In particular, the stability index provides an estimate of the stability of a protein in a test tube, which is calculated based on weight value of instability of 400 different dipeptides. A protein with the value smaller than 40 is predicted as stable while a value above 40 predicts that the protein may be unstable. Aromaticity is simply the relative frequency of Phenylalanine (Phe), Tryptophan (Trp), and Tyrosine (Tyr) in a protein sequence. Aromatic amino acids are normally used to quantify the concentration of proteins in an unknown sample. Grand Average of Hydropathy (GRAVY) is calculated as the sum of hydropathy values of all the amino acids that are divided by the number of residues in the sequence. A positive GRAVY value indicates that the protein is hydrophobic, and a negative value indicates that it is hydrophilic. The results were then visualized through R programming and compared with the other three groups (Domain1, Domain2, and Domain3) in order to identify significant differences in biological features.

2.5. Data Mining and Statistical Analysis

A set of Python scripts were used to dissect bacterial proteomes through HMM models, classify proteins based on domain numbers and redundancy, and analyse co occurrence of domains and physiochemical properties of protein sequences, which are all available under request. All statistical analyses were performed by using Student *t*-test.

3. RESULTS

In general, our studies showed that the average length of a protein domain is 174 aa across 16712 Pfam domains, while the average length of domains in prokaryotic species is 176 aa. Average lengths of proteins with 0, 1, 2, and 3 or more domains tend to have a length difference of 100 aa between each group and the preceding one. Thus, proteins with more domains should be statistically longer. The study also found that the number of all domains is linearly increasing with proteome size. In addition, the analysis also indicated that domain duplication might play an important role in protein evolution and is genome-size dependent. It was revealed that the probability of domain duplication in small-size bacterial genomes (<2000 proteins/proteome) is generally less than 40% while those with comparatively larger bacteria genomes (>2000 proteins/proteome) have duplication rates between roughly 40-60%, though exceptions exist (*P*-value<0.001). Domain-centric views of bacterial physiology and lifestyle were also reported. More details are presented below.

3.1. Distribution of Four Groups of Proteins in Bacterial Proteomes

All protein sequences in each proteome were screened against 16712 Pfam HMM sequence models automatically *via* the HMMER package. The e-value threshold was set to 0.001 for a domain hit to be considered statistically significant. For each proteome, four groups of proteins were identified as previously mentioned, that is, proteins with zero (Group 1), single (Group 2), double (Group 3), and multiple domain(s) (Group 4) (Fig. **1**). There was a clear correlation between the number of proteins in each group and bacterial proteome size. Through the analysis of polynomial regression, all correlations follow the formula of $y=ax^2+bx$, where y represents bacterial proteome size while x represents number of proteins. The parameters of a and b are different in each group to reflect group differences. According to the regression analysis, it was observed that the number of no-domain proteins is positively correlated with proteome sizes ($a>0$) while the other three groups are negatively correlated ($a<0$). It was also revealed that single-domain and multiple-domain proteins are comparatively more abundant in bacteria than no- and double-domain proteins. Moreover, a couple of the bacterial species were notable in Fig. **1** due to the extremely low number of multi-domain proteins: *Chlamydia psittaci* (102 out of 2863, 3.56%) and *Richelia intracellularis* (703 out of 7610, 9.24%). Interestingly, both of the species have a large number of single domain proteins. In contrast, other bacterial species with a similar or the same proteome size have many more multi-domain proteins, such as *Listeria monocytogenes* (812 out of 2844, 28.55%) and *Myxococcus xanthus* (1802 out of 7314, 24.64%). Interestingly, both *Chlamydia psittaci* and *Richelia intracellularis* adopt an intracellular lifestyle while *Listeria monocytogenes* and *Myxococcus xanthus* are environmental bacteria with versatile metabolic abilities and resistance to hash conditions.

3.2. Non-redundant Domains and their Duplication Rates in Bacterial Proteomes

Proteome size was previously confirmed to be linked with organism complexity and bacterial lifestyle, although complexity is a multi-parametric trait [16, 17]. In addition, domains are independent units for protein structure and function. Thus, the number of domains should also reflect bacterial complexity to some degree. Our initial analysis of the relationship between bacterial proteome sizes and domain numbers generated a linear regression between the two factors with a coefficient of determination equal to 0.939 (purple dots in Fig. **2**). However, proteomes are made up a set of proteins evolved through recombination and duplication of a limited number of domains [18]. Simply comparing total protein numbers or domain numbers cannot reflect real bacterial complexity and the number of non-redundant domains should be utilized for the measurement (red dots in Fig. **2**). Based on this principle, it was observed that maximal number of non-redundant domains (2818 domains) exist when proteome size is at 6309 proteins by using polynomial regression with power of 2. However, the curve seemed to be underfitting. After adjustment, we adopted an asymptotic function to simulate the trend by following power law, which is better to reflect the continuous evolution of bacterial domains and proteins.

Further analysis calculated the ratio of duplicated and non-redundant domains and identified a proteome size dependent relationship (yellow dots in Fig. **2**). That is, with the increment of proteome size, the increasing rate of non-redundant domains tends to reach a maximal value while the

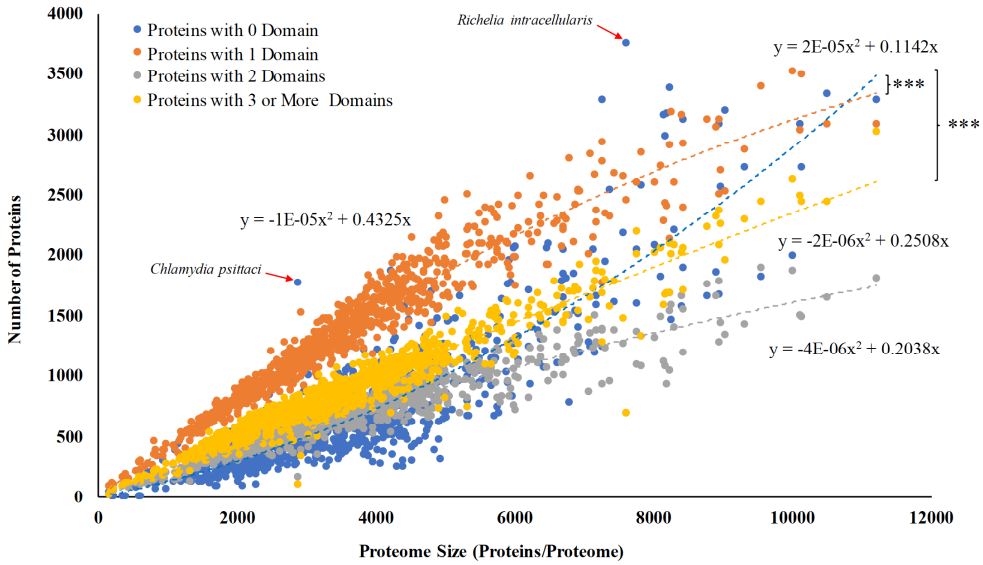

Fig. (1). Classification of bacterial protein sequences sourced from 944 bacterial reference proteomes into four categories according to the number of Pfam-detectable domains that each protein possesses, which include proteins with 0 (Group1, blue dots), 1 (Group 2, orange dots), 2 (Group 3, grey dots), and 3 or more (Group 4, yellow dots) domain(s). Each coloured dot corresponds to a single bacterial species while each species corresponds to four dots in total as specified above. For the distribution of each category, polynomial regression analysis was used to fit each set of scattered plots into a curve. All curves follow the formula $y=ax^2+bx$, although a and b are different in each group. For details, please see the specific formula in the figure. The number of proteins with 0 domains increases with bacterial proteome size ($a>0$) while the other three groups are not ($a<0$). Differences among Group 1, Group 2, and Group 4 are statistically significant (annotated as triple stars ***, *P*-value<0.001) while Group1 and Group 3 are not statistically different. Two representative bacterial species were pointed out with red arrows due to their comparatively low number of multi-domain proteins. All statistical analysis was performed based on two-tailed un-equal-variance Student's *t*-test. (*The color version of the figure is available in the electronic copy of the article*).

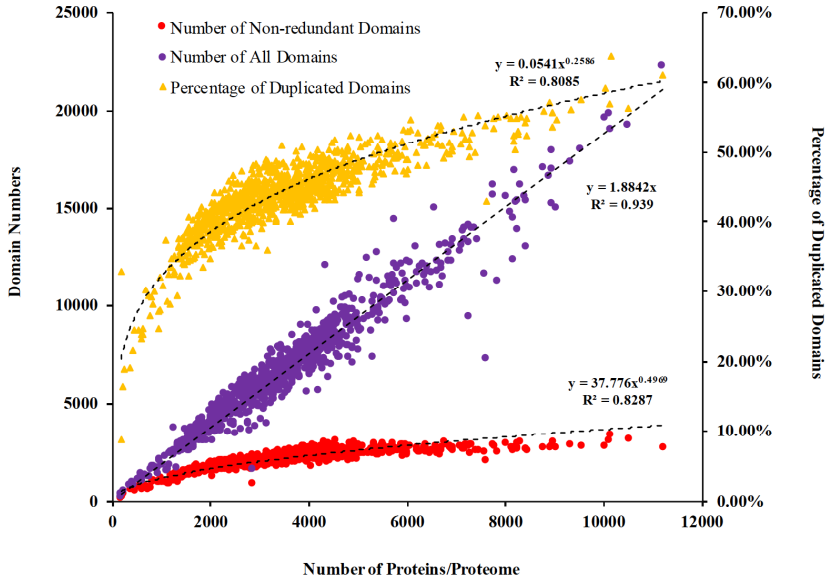

Fig. (2). Relationships between bacterial proteome size and the number of all domains (purple dots), the number of non-redundant domains (red dots), and the rate of domain duplications (yellow dots) at species level. For each distribution pattern, both a regression equation and a corresponding coefficient of determination R^2 were given. The primary vertical axis indicates the number of all (purple dots) and non-redundant (red dots) domains, respectively. The secondary vertical axis indicates the number of duplicated domains as a percentage of all non-redundant domains (yellow dots) in a single proteome. Comparison of the three groups of domains revealed that the number of all domains in a proteome is linearly correlated with the proteome size. In addition, the number of non-redundant domains seems to be constant after proteome size surpass a certain value (6309 proteins/proteome) with a quadratic equation simulation, although further analysis using 4^{th} degree polynomial equation confirmed that the curve is asymptotic but could reach a maximal increasing rate at the proteome size of 16962 proteins. Correspondingly, percentage of domain duplication tends to increase along with the increase of proteome size with a higher rate. Thus, bacteria with comparatively large proteomes tend to have more redundant domains through domain repetitions and combinations rather than novel domains to achieve multiple functions. (*The color version of the figure is available in the electronic copy of the article*).

number of all domains and non-redundant domain keep increasing with a much higher rate, which leads to the duplication rates going up from 8.79% to 63.76%. Interestingly, some bacterial species show comparatively abnormal domain redundancy, such as *Hodgkinia cicadicola* (169 proteins/proteome) and *Actinospica robiniae* (191 proteins/proteome) with domain duplication rates of 8.79% and 32.89%, respectively, which may shed light on their specific lifestyles.

3.3. Domain Interactions in Representative Bacterial Species

In order to understand the importance of domain associations in bacterial physiology and lifestyle, we selected four representative bacterial species, *Deinococcus deserti*, *Yersinia pestis, Helicobacter pylori*, and *Anaplasma phagocytophilum* , which fall within the categories of free-living bacteria, sit-and-wait pathogens, exclusively host-associated bacteria, and vector-borne pathogens, as previously classified [14]. Through the visualization analysis *via* CytoScape, core domains, high-frequently co-occurrent domains, and their interactions were systematically identified, which were then interpreted for their contributions to bacterial specific living features [12]. Core domains are defined as those belonging to the largest partitioned cluster of concurrent domains in a proteome while high-frequently co-occurrent domains have connections with equal to or more than 20 other domains. In particular, for each bacterium, core domain interactions were displayed and high-frequently co-occurrent domains are marked in yellow in Supplementary Fig. (**1**). Names of highly associated Pfam domains corresponding to the yellow dots in Supplementary Fig. (**1**) are presented in Supplementary Table (**2**) for each species. Domains such as AAA, ABC_Tran, GTP_EFTU, and MMR_HSR1 were identified in all four species that are essential for bacterial physiology. On the hand, domains like MarR, HisKA, Res III, and FeoB_N were also found to be present with a bias distribution, and could reflect bacterial unique lifestyle to a certain degree. It is also clear that domain associations are more intensive for free-living and sit-and-wait bacteria, while domain interaction in obligate-intracellular and vector-borne bacteria are more scattered. In addition, for sit-and-wait pathogen *Yersinia pestis* and obligate intracellular parasite *Helicobacter pylori*, unique domains, such as Antig_Caf1and CagA, respectively, at species level by comparing with other 943 bacteria were also identified, which are able to provide further insight into their unique lifestyles. For unique domains in each bacterium, please refer to Supplementary Table **3** for details.

3.4. Physicochemical Properties of No-domain Proteins

Although protein domain analysis at species level could provide insights into bacterial physiology and lifestyle, it is not helpful to explain how sequences are intrinsically different in terms of domain organizations for those with no detectable domains. Thus, in order to better understand the features of no-domain proteins, we pooled all the sequences from the 944 bacterial proteomes together and divided them into four groups as previously described. Protein stability, aromaticity, hydrophobicity, and isoelectric points (pI index) were calculated based on ProtParam package from Biopython (Fig. **3** and **4**) [13]. In addition, amino acid compositions were also studied for the four-group proteins (Fig. **5**).

Distributions of the three indexes in Fig. **3** are statistically different (P-value<0.001) and no-domain proteins have the widest range. As for protein stability, the minimal value for no-domain group is -67.06 while the maximum is 226.42. Overall, mean stability of no-domain proteins is 40.95, that is considered as not stable while the other three groups have stability indexes of less than 40 and are thought to be stable. As for aromaticity index, no-domain proteins showed statistically higher mean value and wider range than the other three groups. GRAVY index showed that mean values of the four groups of proteins are negative, indicating biased hydrophilic distribution in all four groups.

Since the isoelectric point (pI) value not only reflects whether a protein is negatively or positively charged in the environment, but also indicated when a protein is less soluble at specific pH, we calculated the pI value of all proteins in the four groups. The results showed that no-domain proteins tend to be more basic with average pI greater than 7 (Fig. **4A**). In addition, density plot of pI value in the four groups presents a clearer view about how pI values are distributed, in which no-domain proteins tend to be comparatively dominant at higher pI values with peaks at around 10 and 12. When pI is less than 7, the other three groups are more abundant. As for the amino acid compositions, no-domain proteins tend to have a larger variety of amino acid composition. However, the specific biological meaning is hard to explain and requires further exploration.

4. DISCUSSION

A proteome is a set of protein sequences that are derived by translation of all protein coding genes of a completely sequenced genome [10]. Pair-wise and multiple sequence alignments like BLAST and MUSCLE provide powerful ways to understand protein functions simply based on the protein primary structure of amino acid composition. On the other hand, protein domains are functionally and structurally independent genetic elements. Through Pfam HMM models, it is convenient to identify more remotely homologous domains in protein sequences automatically than sequence alignment tools [19]. In addition, all protein domains in a specific organism is called domainome that is relative to the organism's biological complexity [5]. Thus, we used 16712 Pfam HMM models in this study and screened a set of reference proteomes that belong to species representing 944 bacterial genera, in order to gain insights into the domain distribution patterns and their impacts on physiology in prokaryotes.

4.1. Domain Distributions and Bacterial Lifestyle

Consistent with previous studies, different types of proteins exist in bacterial proteomes in terms of domain organizations, such as single, double, and multiple domain(s) [8, 20]. In addition, a special group of proteins with no detectable domains was also identified, which is rarely mentioned and present as a dark matter in proteomes. Systematic study of all proteins from 944 bacterial species shows that no-domain proteins in bacterial proteomes account for 19.06% of proteins on average, while the percentage of the sum of

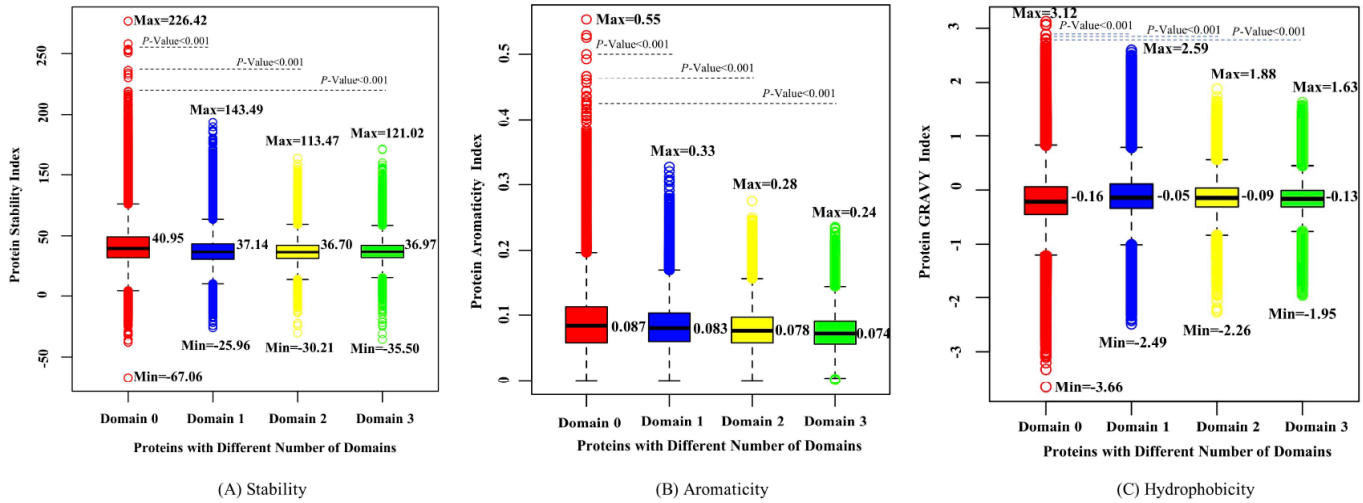

Fig. (3). Calculation of physicochemical features of four groups of bacterial proteins that have no- (red), single- (blue), double- (yellow), or multiple (green) domain(s). (**A**) Bacterial protein stability index. (**B**) Bacterial aromaticity index. (**C**) Bacterial grand average of hydrophobicity (GRAVY) index. No-domain proteins are statistically less stable (stability index>40) than the other three protein groups. Aromaticity of no-domain proteins is also significant higher. As for hydrophobicity, all protein groups show bias toward hydrophilicity (GRAVY index<0). *P*-values were calculated based on two-tail unequal-variance Student's *t*-test. (*The color version of the figure is available in the electronic copy of the article*).

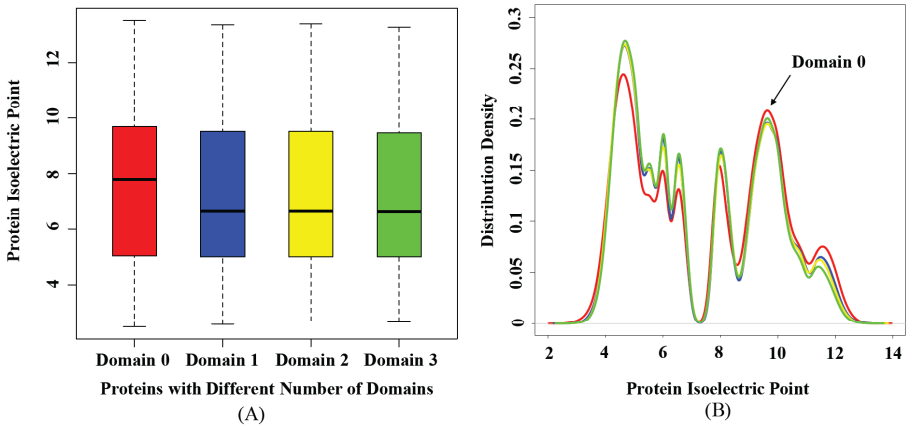

Fig. (4). Comparisons of isoelectric points among four-group proteins that have no- (red), single- (blue), double- (yellow), or multiple (green) domain(s). (**A**) Boxplot analysis showed that no-domain proteins tend to be more basic, while other three groups of proteins are more likely to be acidic. (**B**) Density plot of protein isoelectric points showed that distribution of no-domain proteins are skewed towards higher pI value with peaks at 10 and 12, approximately. (*The color version of the figure is available in the electronic copy of the article*).

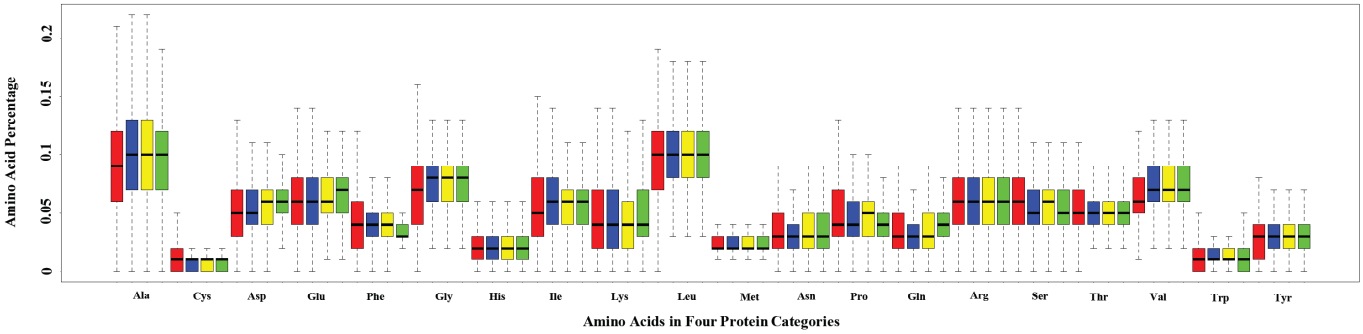

Fig. (5). Comparison of the compositions of amino acids among four-group bacterial proteins that have no- (red), single- (blue), double- (yellow), or multiple (green) domain(s). The 20 amino acids are abbreviated by following IUPAC nomenclature standards. A set of amino acids have identical distribution patterns in the four groups, such as His, Met, and Arg while other amino acids have different level of variations, especially for Ala, Gly, Ile, Thr, Val, and Tyr. Both Ala and Leu are comparatively abundant while Cys and Trp are statistically scarce in bacterial proteins. (*The color version of the figure is available in the electronic copy of the article*).

double- and multiple-domain proteins is around 42.97%. Single domain proteins, on the other hand, are the most abundant protein type with a percentage of 37.98%. Among these distribution, two species, *Chlamydia psittaci* and *Richelia intracellularis* are distinct due to their extremely high proportion of no-domain and single domain proteins and very low percentage of multi-domain proteins (Fig. **1**). Both of the species are obligate intracellular organisms and have experienced differential levels of genome reduction [21, 22]. It has been reported that changes of redundancy and diversity in the evolution of reductive genomes is reflected at several levels, such as genes and proteins [23, 24]. Due to the niche-specific adaptation, paralogous genes or proteins tend to be lost, while diversity of genes or protein families is preserved. Thus, although the robustness of genomes is compromised, there is still sufficient protein functional complexity for maintaining host-bacterial interactions [24, 25]. Accordingly, it was found that multi-domain proteins tend to be stripped down to the bones by losing terminal domains in genome-reduced *Serratia symbiotica*, which is consistent with our initial observation that *Chlamydia psittaci* and *Richelia intracellularis* are abundant in single domain proteins, with an extremely low percentage of multi-domain proteins, probably due to genome reduction or stream-lining evolution [23]. It was also found through modelling that proteins with two domains are twice as likely to be lost as proteins with a single domains [24]. Our analysis confirmed that the average proteome size of bacterial species with the highest percentage of single domain proteins (top 50 out of 944 bacteria) is smaller, with 2781 proteins/proteome, and is statistically significantly different from those with lowest percentage of single domain proteins (bottom 50), proteome size of which is 5974 proteins/proteome on average (p-value<0.001 on a two-tail unequal variance Student's t-test). Among the top 50 species include endosymbiont *Buchnera aphidicola* (359 proteins/proteome), obligate intracellular parasite *Rickettsia prowazekii* (834 proteins/proteome), and the *Plautia stali* symbiont (5007 proteins/proteome), *etc*. Thus, abundance of single-domain proteins could serve as an indicator for bacterial host-dependent or restricted niche-adapted lifestyle to a certain degree.

4.2. Domain Redundancy and Bacterial Lifestyle

Further analysis looked into the domain redundancy *versus* bacterial proteome size, and reinforced the finding that domain duplication rates are positively linked with bacterial proteome size, and also possibly with bacterial lifestyle (Fig. **2**). That is, for bacteria with tightly host-associated features, low domain redundancy is more preferred. For example, *Hodgkinia cicadicola* (169 proteins/proteome) has and extremely small genome and its domain redundancy is 8.9%, the lowest of all 944 bacteria. In contrast, domain redundancy of a similar small genome bacteria *Actinospica robiniae* (191 proteins/proteome) is 32.89%. The vast difference between the two organisms is that the former is an endosymbiont while the latter live freely in an acidophilic environment in forest soil [26, 27]. Thus, it was partially confirmed that lifestyle could play important roles in domain redundancy. Several other bacterial species show comparatively higher domain duplication rates, such as *Conexibacter woesei* isolated from forest soil and *Verrucosispora maris* isolated from deep-sea sediments, further study of which may shed light on their unique physiological activities and lifestyles. In addition, we also selected 17 representative bacterial species according to the classification of Wang *et al.* [14] based on their lifestyles and higher domain duplication rate is obviously associated with bacteria capable of environmental survival such as free-living and sit-and-wait bacteria according to the comparison of averaged duplication rates. Statistical significance in terms of domain duplication rates was also identified between free-living and parasitic groups with p-value less than 0.005 (Supplementary Table **4**).

4.3. Domain Interactions and Bacterial Physiology

Considering that the domainome could reflect bacterial metabolic complexity [5], we also attempted to confirm this hypothesis *via* domain interaction networks, that is, linkage of all non-redundant co-occurrent domains in a proteome. Domain interactions in *Deinococcus deserti*, *Yersinia* pestis, *Helicobacter pylori*, and *Anaplasma phagocytophilum* that fall into categories of free-living bacteria, sit-and-wait pathogens, exclusively host-associated bacteria, and vector-borne pathogens were visualized (Supplementary Fig. **1**). *Y. pestis* showed the most sophisticated domain interaction network, which is also a reflection of its large number of non-redundant domains and multi-faceted living environments. It also holds true for the free-living *D. deserti*. Core proteins such as AAA (PF00004), ABC_Tran (PF00005), GTP_EFTU (PF00009), and MMR_HSR1 (PF01926) were identified and widely distributed in bacterial species, which contribute to variety of essential cellular activities like DNA replication, compound transportation, GTP hydrolysis, and ribosome interactions. On the other hand, some core domains could reflect bacterial unique lifestyle in these representative bacteria. For example, MarR (PF01047) identified in *Deinococcus deserti* is involved in a non-systematic multiple antibiotic resistance, the abundance of which typically correlates with a free-living lifestyle and large genome size [28]. HisKA (PF00512) found in *Yersinia pestis* is a two-component regulatory system that serves as a basic stimulus-response coupling mechanism to allow organisms to sense and respond to changes in many different environmental conditions [29]. This is consistent with the lifestyle of *Y. pestis* that needs to survive both within and outside hosts. ResIII (PF04851) from *Helicobacter pylori* is normally found in enzymes belonging to Restriction-Modification (R-M) system that are used for protecting the bacterium against invading foreign DNA [30]. Finally, FeoB_N (PF02421) from tick-borne intracellular pathogen *Anaplasma phagocytophilum* is the N-terminus of ferrous iron transport protein B. Nearly all bacteria require iron as a metabolic co-factor to grow normally, and infection by an intracellular pathogen is characterised by an enhanced level of ferritin protein that acts as an iron storage, which has been confirmed in *A. phagocytophilum* [31]. Consequently, abundance of FeoB_N could be an indicator for the intracellular lifestyle of *A. phagocytophilum*. In addition, unique domains for each of 944 bacterial species were also identified by comparing with other 943 bacteria, which could be used for inferring bacterial unique physiology, although some bacteria do not return any results from the comparison (Supple-

mentary Table **3**). For example, unique domains in *Y. pestis* and *H. pylori* are Antig_Caf1 (Caf1 capsule antigen) [32] and CagA (cytotoxin-associated gene A) [33], respectively, which are tightly correlated with their pathogenic phenotypes and are able to provide further insight into their unique physiological activities and lifestyles. Thus, through the analysis of domain interactions networks and species-centric domains, we could get insights into the complexity of bacterial physiology and identify core domains that may be responsible for their unique lifestyles.

4.4. Physiochemical Properties of No-domain Proteins

We tried to analyse the physicochemical features of the these "theoretically invisible" sequences in order to obtain as much information as possible for this special group of proteins from statistics point of view. Initial BLAST analysis found that no-domain proteins are normally distributed within limited number of species or genera (unpublished data). Recent study also identified that recently evolved proteins tend to have no annotated domains, which suggestes that no-domain proteins could still be young and have not spread across bacterial species [34]. As for physicochemical properties, no-domain proteins are statistically different from other groups, which also suggested that this group is distinct, especially for stability and isoelectric point. However, since the availability of limited studies, no much biological explanation could be drawn from these results. Thus, more experimental efforts should be focused on this group of proteins in order to better understand bacterial physiology and metabolism.

CONCLUSION

In this study, we systematically investigated 944 bacterial species based on manually curated proteomes. Domain duplication rate is positively correlated with bacterial proteome size and possibly linked with bacterial lifestyle. Comparison of domain redundancy of two species with similar proteome sizes, symbioint *Hodgkinia cicadicola* and free-living *Actinospica robiniae*, supported the claim. Further analysis revealed that domain redundancy is more associated with bacteria that interact with environment and/or host intensively, which is beneficial to bacterial robustness [24]. In contrast, for obligate intracellular bacteria, single-domain proteins are dominant while protein interactions are strengthened probably for compensating the loss of multi-domain proteins [24, 25]. Thus, high percentage of single domain proteins in a proteome may be associated with bacterial host-associated lifestyle. Domain co-occurrence based on network analysis *via* CytoScape revealed that bacteria free-living or sit-and-wait bacteria tend to have more sophisticated interaction modes while domain interaction networks for obligate intracellular and vector-borne bacteria are less intensive. In addition, species-specific domains were also identified that may contribute to bacterial unique physiology and metabolism, which provided a novel method for studying bacterial lifestyle. Finally, four groups of proteins from 944 bacteria were all screened for physicochemical properties. No-domain proteins was distinctly different from other three groups in terms of stability, aromaticity, hydrophobicity, and isoelectric point based on statistical analysis. However, no apparent differences were observed for amino acid compositions among the four protein groups. Specific biological explanations for these differences require more experimental studies and are left for further analysis. In addition, our future work will focus on the construction of an integrated online platform for systematic analysis of bacterial proteomes based on Pfam HMM domain models so as to improve the analysis and data visualization of bacterial proteomes and also obtain a better understanding of bacterial physiology.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

For the protein domain study, a complete set of 16712 HMM-based protein domains was downloaded from the Pfam database http://pfam.xfam.org/_(31.0 Release), which was then used to scan protein sequences in the collected bacterial proteomes.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We appreciate Associate Professor Michael J. Wise from the University of Western Australia for the assistance of checking and improving English language. The work was supported by the Startup Foundation for Excellent Researchers at Xuzhou Medical University (No. D2016007), The Natural Science Foundation for the Jiangsu Higher Education Institutions of China (No. 16KJB180028), Innovative and Entrepreneurial Talent Scheme of Jiangsu Province (2017), and Natural Science Foundation of Jiangsu Province (BK20180997).

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.,* **2000**, *28*(1), 235-242.

[2] Goodacre, N.F.; Gerloff, D.L.; Uetz, P. Protein domains of unknown function are essential in bacteria. *MBio,* **2013**, *5*(1), e00744-13, (DOI: 10.1128/mBio.00744-13).

[3] Belshaw, R.; Yang, S.; Bourne, P. E. The evolutionary history of protein domains viewed by species phylogeny. *PLoS ONE,* **2009**, *4* (12), e8378.

[4] Pasek, S.; Risler, J. L.; Brezellec, P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics,* **2006**, *22*(12), 1418-1423.

[5] Kuznetsov, V.A.; Pickalov, V.V.; Kanapin, A.A. Proteome com-

plexity measures based on counting of domain-to-protein links for replicative and non-replicative domains. In: *Bioinformatics of Genome Regulation and Structure II*, **2006**; pp 329-341.

[6] Chen, C.; Huang, H.; Wu, C.H. Protein bioinformatics databases and resources. *Method Mol. Biol.,* **2017***;*1558(1)*,* 3-39.

[7] Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*(1), 279-285.

[8] Zhang, X.C.; Wang, Z.; Zhang, X.; Le, M.H.; Sun, J.; Xu, D.; Cheng, J.; Stacey, G. Evolutionary dynamics of protein domain architecture in plants. *BMC Evol. Biol.,* **2012**, *12*(1), 6.

[9] Rentzsch, R.; Orengo, C.A. Protein function prediction using domain families. *BMC Bioinform.* **2013**, *14* (Suppl 3).

[10] Apweiler, R., UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.,* **2004**, *32*(90001), 115-119.

[11] Babushok, D.V.; Ostertag, E.M.; Kazazian, H.H., Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cell. Mol. Life Sci.,* **2006**, *64*(5), 542-554.

[12] Shannon, P., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.,* **2003**, *13*(11), 2498-2504.

[13] Cock, P.J. A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* **2009**, *25*(11), 1422-1423.

[14] Wang, L.; Liu, Z.; Dai, S.; Yan, J.; Wise, M.J. The sit-and-wait hypothesis in bacterial pathogens: A theoretical study of durability and virulence. *Front. Microbiol.,* **2017**, *8*(2167), (doi: 10.3389/fmicb.2017.02167).

[15] Walker, J.M. *The Proteomics Protocols Handbook,* Humana Press: New York, **2005**.

[16] Schad, E.; Tompa, P.; Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol.,* **2011**, *12*(12), R120, (doi: 10.1186/gb-2011-12-12-r120).

[17] Wang, L.; Yan, J.; Wise, M. J.; Liu, Q.; Asenso, J.; Huang, Y.; Dai, S.; Liu, Z.; Du, Y.; Tang, D. Distribution patterns of polyphosphate metabolism pathway and its relationships with bacterial durability and virulence. *Front. Microbiol.,* **2018,** *9,* 782.

[18] P. Bagowski, C.; Bruins, W.; J.W. te Velthuis, A. The nature of protein domain evolution: Shaping the Interaction Network. *Curr. Genom.,* **2010,** *11*(5), 368-376.

[19] Sonnhammer, E., Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.,* **1998**, *26*(1), 320-322.

[20] Buljan, M. Mechanisms of change in protein architecture. University of Cambridge, Cambridge, **2011**.

[21] Hilton, J.A.; Foster, R.A.; James Tripp, H.; Carter, B.J.; Zehr, J.P.; Villareal, T.A. Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nature Comm.,* **2013**, *4*(1767), (doi: 10.1038/ncomms2748).

[22] Ojcius, D.M.; Voigt, A.; Schöfl, G.; Saluz, H.P., The *Chlamydia psittaci* Genome: A comparative analysis of intracellular pathogens. *PLoS ONE,* **2012**, *7*(4).

[23] Manzano-Marín, A.; Latorre, A. Snapshots of a shrinking partner: Genome reduction in Serratia symbiotica. *Scientific Reports* **2016**, *6*(32590), (doi: 10.1038/srep32590).

[24] Pilpel, Y.; Mendonça, A.G.; Alves, R.J.; Pereira-Leal, J.B., Loss of genetic redundancy in reductive genome evolution. *PLoS Comput. Biol.,* **2011**, *7*(2), e1001082).

[25] Kelkar, Y. D.; Ochman, H. Genome reduction promotes increase in protein functional complexity in bacteria. *Genetics,* **2012**, *193*(1), 303-307, (doi: 10.1534/genetics.112.145656).

[26] Cavaletti, L. Actinospica robiniae gen. nov., sp. nov. and *Actinospica acidiphila* sp. nov.: Proposal for Actinospicaceae fam. nov. and Catenulisporinae subord. nov. in the order Actinomycetales. *Int. J. Sys. Evol. Micro.,* **2006**, *56*(8), 1747-1753.

[27] Molloy, S. A tiny alternative. *Nature Rev. Micro.,* **2009**, *7*(9), 620-620.

[28] Grove, A. MarR family transcription factors. *Curr. Biol.,* **2013**, *23* (4), 142-143.

[29] Viollier, P.H.; Willett, J.W.; Kirby, J.R. Genetic and biochemical dissection of a hiska domain identifies residues required exclusively for kinase and phosphatase activities. *PLoS Genetics,* **2012**, *8* (11), e1003084.

[30] Donahue, J.P.; Peek, J.; R. M. *Helicobacter pylori*: Physiology and Genetics. In *Helicobacter pylori: Physiology and Genetics*, Mobley, H.L.T.; Mendz, G.L.; Hazell, S.L., Eds.; ASM Press: Washington (DC), **2001**.

[31] Carlyon, J.A.; Ryan, D.; Archer, K.; Fikrig, E. Effects of anaplasma phagocytophilum on host cell ferritin mrna and protein levels. *Infect. Immun.,* **2005**, *73*(11), 7629-7636.

[32] Du, Y. Role of fraction 1 antigen of yersinia pestis in inhibition of phagocytosis. *Infect. Immun.,* **2002**, *70*(3), 1453-1460.

[33] Hatakeyama, M. Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer. *Proc. Japn. Acad. Ser. B. Phys. Biol. Sci.,* **2017**, *93*(4), 196-219.

[34] Toll-Riera, M.; Albà, M.M., Emergence of novel domains in proteins. *BMC Evol. Biol.,* **2013**, *13*(47), (DOI: 10.1186/1471-2148-13-47).