Bioinformatics Advance Access published May 22, 2015

*Genome analysis*

# A Novel Essential Domain Perspective for Exploring Gene Essentiality

Yao Lu [1#], Yulan Lu [2#], Jingyuan Deng [3], Hai Peng [4], and Hui Lu [1, 5], Long Jason Lu[3, *],

[1] Shanghai Institute of Medical Genetics, Children's Hospital of Shanghai Shanghai Jiao Tong University, 24/1400 Beijing (W) Road, Shanghai 200040, P. R. China

[2] State Key Laboratory of Genetic Engineering Institute of Biostatistics, School of Life Science, Fudan University, 220 Handan Road, Shanghai 200433, P. R. China

[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, U.S.A.

[4] Institute for Systems Biology, Jianghan University, Wuhan, Hubei, China.

[5] Department of Bioengineering (MC 063), University of Illinois at Chicago, 851 S Morgan St, 218 SEO, Chicago, IL 60607-7052, U.S.A.

[#] Contribute equally

Associate Editor: Dr. John Hancock

**ABSTRACT**

**Motivation:** Genes with indispensable functions are identified as essential; however, the traditional gene-level studies of essentiality have several limitations. In this study, we characterized gene essentiality from a new perspective of protein domains, the independent structural or functional units of a polypeptide chain.

**Results:** To identify such essential domains, we have developed an Expectation-Maximization (EM) algorithm-based Essential Domain Prediction (EDP) Model. With simulated datasets, the model provided convergent results given different initial values and offered accurate predictions even with noise. We then applied the EDP model to six microbial species and predicted 1,879 domains to be essential in at least one species, ranging 10-23% in each species. The predicted essential domains were more conserved than either non-essential domains or essential genes. Comparing essential domains in prokaryotes and eukaryotes revealed an evolutionary distance consistent with that inferred from ribosomal RNA. When utilizing these essential domains to reproduce the annotation of essential genes, we received accurate results that suggest protein domains are more basic units for the essentiality of genes. Furthermore, we presented several examples to illustrate how the combination of essential and non-essential domains can lead to genes with divergent essentiality. In summary, we have described the first systematic analysis on gene essentiality on the level of domains.

**Contact:** Long.Lu@cchmc.org

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Genes are widely regarded as the basic units of a cell, a complex system made up of a large number of components and reactions. Therefore, a fundamental question in synthetic biology is, what is the minimal gene set that is necessary and sufficient to sustain life (Mushegian, 1999)? The individual genes that constitute a minimal gene set are called essential genes. Experimentally, essential genes are defined as those that when disrupted, confer a lethal phenotype to organisms under defined conditions. Therefore, the essentiality of a gene is the indispensability of this gene's product to the survival of a microorganism. Systematic genome-wide interrogations of essential genes have been conducted by single-gene knockouts (Baba, et al., 2006; de Berardinis, et al., 2008; Kato and Hashimoto, 2007; Kobayashi, et al., 2006), transposon mutagenesis (Akerley, et al., 2002; Gallagher, et al., 2007; Gerdes, et al., 2003; Glass, et al., 2006; Hutchison, et al., 1999; Jacobs, et al., 2003; Liberati, et al., 2006), or antisense RNA inhibitions (Forsyth, et al., 2002; Ji, et al., 2001). These experiments have provided a tremendous amount of resources to further our understanding on gene essentiality, one important step closer to unraveling the complex relationship between genotype and phenotype (Dowell, et al., 2010).

Recent comparative research on the available essential gene datasets has shown surprising results and has challenged many early assumptions in genomics. Firstly, it was discovered that microorganisms share a limited set of essential genes. For example, a comparison of the essential gene sets of four bacteria revealed that only 12-72% of essential genes are shared between any pair, indicating that a large number of unique essential genes exist (**Table S1**). This is somewhat expected as independent studies have firmly established that bacterial species share a very limited number of orthologs (e.g., 19-53% in **Table S1**), regardless of which ortholog detection method is used (Bruccoleri, et al., 1998). This may reflect the physiology of diverse microorganisms. Increasing the number of organisms in the list will further decrease the number of orthologs and thus common essential genes. For example, merely 265 orthologs and 34 essential genes are common to all four bacteria (**Fig. S1**). It is highly unlikely that these 34 essential genes would be capable of executing all functions of a cell.

Secondly, and more surprisingly, a recent study showed that, when tested experimentally in model bacteria, less than a quarter of the highly conserved genes were essential (Arigoni, et al., 1998; Freiberg, et al., 2001; Song, et al., 2005; Zalacain, et al., 2003). This suggests that evolutionary conservation of a gene does not necessarily imply that it is essential for microbial survival. Consider a hypothetical bacterium that utilizes both solar and chemical energy. Although the genes for converting solar and chemical energy have been highly conserved and optimized (thus allowing them to be conserved through evolutionary time), neither is essential for survival if both solar and chemical energy are present.

*To whom correspondence should be addressed.

Finally, orthologs are often observed to be essential in one organism but not another. For example, the dapE gene is essential in *E. coli* but non-essential in *P. aeruginosa* (Gerdes, et al., 2006). It is also possible for orthologs to have different functions in different organisms (Liao and Zhang, 2008), even though it is a fundamental assumption of genomics that most orthologs perform a similar function. This suggests that differences in genetic regulation, genetic redundancy, and divergence in cellular pathways or processes between organisms may all affect gene essentiality; their combined effects result in the discrepancy in essentiality between orthologs.

Given the above evidence, a theoretical approach that attempts to identify a universal minimal gene set by searching for conserved orthologs across multiple organisms is unlikely to succeed. Since relying on orthologs, or conservation on the gene level, does not adequately explain essentiality, are there other characteristics among different organisms that are more consistent with essentiality? Observing that genes from different evolutionary origins perform the same function (i.e., non-orthologous gene displacements), a group of researchers switched their focus to defining the set of essential functions (Delaye and Moya, 2010; Gil, et al., 2004). However, knowing a set of essential functions does not necessarily mean that the genes responsible for these functions can be easily identified, especially in understudied organisms.

In this study, we have reexamined gene essentiality from a novel essential protein domain point of view. Our results suggest that this new perspective may offer unique insights into the mechanistic basis of gene essentiality and help to resolve the controversy regarding this phenomenon.

Protein domains with known function, such as those of the ATPase enzyme class (Jaroszewski, et al., 2009), are amino acid sequence patterns with an associated activity. They often correspond to structural domains, the independent globular components of the polypeptide chain found in three-dimensional protein structures, although the correspondence is not exact (Zhang, et al., 2005). Independent of neighboring sequences, a protein domain folds into a distinct structure and mediates the protein's biological functionality (Kanaan, et al., 2009) (**Fig. S2**). Protein domains can be detected from gene/protein sequences by well-established algorithms, such as HMMER (Finn, et al., 2011). Currently, the Pfam v.27.0 database contains 14,831 domains detected by HMMER, many of which have unknown functions (Finn, et al., 2010).

We hypothesized that gene essentiality is likely preserved through the function of protein domains or domain combinations, rather than through the conservation of the entire genes. Numerous examples can be found in literature to support this postulation. For instance, there is only a single copy of DNA polymerase III subunits τ and γ domain III (PF12169) (Jergic, et al., 2007) in *E. coli, P. aeruginosa* and *B. subtilis*. The host gene of this domain, dnaX, is essential in all three species; however, the sequence identity between them is low (**Fig. S2**). In further support of modularity within essential genes, previous studies have discovered that although a gene as a whole may be essential, not every domain within the gene is required for the essential function. For example, *E. coli* ftsK (b0890) is an essential gene consisting of two domains: the N-terminal (a.a. 1-780) and the C-terminal (a.a. 781-1329). Only the N-terminal domain of this gene is required for its role in cell division and viability (Wang and Lutkenhaus, 1998). This study, among others, provides direct evidence that protein domains may be responsible for the essentiality of a gene.

The structure of this paper is as follows: In the **Methods** section, we describe the data sources and give the details of the Essential Domain Prediction (EDP) model. In the **Results** section, we first assess the performance of the EDP model on simulated data sets. We then use the EDP model to predict essential domains in six microbes, detailed domain information can be found in **Table S2**. Next, we investigate the properties of these predicted essential domains, such as their conservation and functions. We also use these essential domains to reproduce the annotation of essential genes as a validation of our predicted essential domains and present several cases where literature provides support for our predictions. In the **Discussion** section, we discuss the significance of our research, the advantages of explaining gene essentiality from a domain perspective, and possible improvements that can be applied to our EDP model.

## 2 METHODS

### 2.1 Essential Gene Data Sets

*Escherichia coli* K-12 sequence data were downloaded from Comprehensive Microbial Resource (CMR) (http://cmr.jcvi.org/tigrscripts/CMR/GenomePage.cgi?database=ntec01). This database contains 4,289 protein sequences in total (Hashimoto, et al., 2005). The essential genes of E. coli K-12 were downloaded from the PEC database (Kato and Hashimoto, 2007). The Kato data set contains 302 essential genes from gene deletion experiments.

*Pseudomonas aeruginosa* PAO1 sequence data were downloaded from the Pseudomonas Genome Database (http://www.pseudomonas.com/) (Pseudomonas_aeruginosa_PAO1.faa, revision 2009-07-17). PA essential genes were adopted from (Winsor, et al., 2011). The Jacobs data set contains 678 essential genes from transposon mutagenesis experiments in PAO1.

*Acinetobacter baylyi* ADP1 sequences were collected from the Magnifying Genomes Database (http://www.genoscope.cns.fr/agc/mage). Out of a total of 3,308 genes, 499 essential genes were acquired from (de Berardinis, et al., 2008).

*Bacillus subtilis* sequence data were downloaded from Microbial Genome Database (http://mbgd.genome.ad.jp/). This data set contains 4,117 ORFs. The essential gene list was acquired from (Uchiyama, et al., 2010) and consists of three data sets: 150 essential genes determined by Kobayashi's mutant genesis experiments, 42 known essential genes from previous studies, and 79 essential genes by homology mapping to other bacteria, most of which encode proteins involved in ribosome or synthesis.

*Saccharomyces cerevisiae* sequences were downloaded from Saccharomyces Genome Database (http://downloads.yeastgenome.org/sequence/genomic_sequence/). This resource contains 5,885 ORFs. The essential gene list was obtained from (Giaever, et al., 2002). This data set contains 1,049 essential genes from targeted mutagenesis experiments.

*Neurospora crassa* ORFs were downloaded from *Neurospora crassa* database at Broad Institute (http://www.broadinstitute.org/annotation/ genome/neurospora/MultiDownloads.html). Dubious ORFs and pseudo genes were excluded from this list. Essential gene data was kindly provided by K. Borkovich from the systematic genome deletion project in *N. crassa* at UC Riverside. This list contains 1,251 essential genes.

Gene ontology (GO) annotations for domains were downloaded from the Mappings of External Classification Systems to GO (http://www.geneontology.org/GO.indices.shtml).

### 2.2 The domain data set and data filtering

We used Pfam (http://pfam.sanger.ac.uk/) (Punta, et al., 2012) to derive domain information from protein sequences, and Pfam-A entries are combined with Pfam-B entries to construct our domain data set. In total, 3,629 Pfam-A and 5,098 Pfam-B domains were included in our analysis. Genes that had no domain annotation were excluded, leaving a total of 23,009 genes. Specific numbers of genes and domains for each species are shown in **Table 1**.

| **Essential**/non-essential | Domain | | Gene |
|---|---|---|---|
| | PfamA | PfamB | |
| *A.baylyi* | 283/1058 | 67/577 | 448/2185 |
| *E.coli* | 191/1558 | 56/678 | 265/3279 |
| *P.aeruginosa* | 236/1477 | 78/1043 | 550/4104 |
| *B.subtilis* | 107/1300 | 34/759 | 178/2978 |
| *N.crassa* | 424/1332 | 585/2024 | 1035/3871 |
| *S.cerevisiae* | 370/1388 | 343/1090 | 866/3250 |

**Table 1** Details for gene and domain datasets. The number of domains and genes in each dataset are shown, as well as the essential count of each type.

## 2.3 The Essential Domain Predictor (EDP) Model

Assuming the genome contains n different genes, we defined $G = \{g_1, g_2, \dots, g_n\}$. For $g_i$, $i \in (1, n)$, let $g_i=1$ if the i-th gene is essential, and 0 otherwise. The vector G was obtained from the experiments, and thus was treated as observed data in our model. Suppose gene $g_i$ contains $n_i$ different domains which form the set $D(g_i) = \{D_i^1, D_i^2, \dots, D_i^{n_i}\}$. Here variable $D_i^j = 1$ if this domain is essential, and 0 otherwise, where (i, j) denotes the j-th domain of the i-th gene. These $D_i^j$ values are unobserved from the experiments, need to be predicted from the model, and are treated as missing values. Each protein may include several distinct domains and each domain may occur in different proteins. Suppose that a genome contains a total of m unique domains denoted $D = \{D_1, D_2, \dots, D_m\}$, where $D_k=1$, $k \in (1, m)$ if the k-th unique domain is essential, and 0 otherwise. We also define $S = \{S_1, S_2, \dots, S_m\}$, where $S_k$, $k \in (1, m)$ is the set of domains $D_m^n$s that are equal to the k-th unique domain Dk. We use $|S_k|$ to denote the size of the set $S_k$, and we further define $\delta_k$ as the probability that domain $D_k$ is essential.

We also described two kinds of errors that may exist in the prediction process: falsely predicted essential rate (FER) and falsely predicted non-essential rate (FNR). These can be defined as follows:

$$FER = Pr(p_i = 1 | g_i = 0), FNR = Pr(p_i = 0 | g_i = 1), \quad (1)$$

where $p_i=1$ if the i-th gene is predicted to be essential and 0 otherwise.

The goal of this model is to estimate the parameters set θ to maximize the likelihood of observed essential genes. Because L(G|θ) is difficult to optimize directly, we augmented the observed likelihood L(G|θ) with missing data D, and the complete data likelihood is thus defined as: L(G,D|θ)=L(D|θ)L(G|D,θ). We further derived the formula as:

$$L(D|\theta) = \prod_{k=1}^{m} L_k \quad \text{where}, L_k = \begin{cases} \delta_{k,} & D_k = 1 \\ 1 - \delta_{k,} & D_k = 0 \end{cases}$$

$$L(G|D, \theta) = \prod_{i=1}^{n} L_i, \quad (2)$$

$$\text{where } L_i = \begin{cases} (1 - FNR): g_i = 1 \quad and\ one\ D_i^j \in D(g_i) = 1 \\ FER: g_i = 1 \quad and\ all\ D_i^j \in D(g_i) = 0 \\ FNR: g_i = 0\ and\ one\ D_i^j \in D(g_i) = 1 \\ 1 - FER: g_i = 0\ and\ all\ D_i^j \in D(g_i) = 0 \end{cases}, \quad (3)$$

where FER and FNR are defined as before and $\theta = (\delta_k, FER, FNR)$. Under this framework, we adopted the conventional Expectation-Maximization (EM) algorithm (Hastie, et al., 2001) to compute the optimal θ that maximizes L(G|θ). We derived the EM algorithm as follows:

E-step: during the E step of the t-th iteration, D is updated by the conditional expectation given the estimated θ from last iteration $\theta_{t-1}$ and G, that is:

$$p_i^j(t) = E\left[D_i^j | G, \theta^{(t-1)}\right] = Pr\left(D_i^j = 1 | G, \theta^{(t-1)}\right) = \frac{Pr\left(D_i^j = 1 | \theta^{(t-1)}\right) Pr\left(g_i | D_i^j = 1, \theta^{(t-1)}\right)}{Pr\left(g_i | \theta^{(t-1)}\right)}$$

$$= \frac{\delta_i^j(t-1)\, Pr\left(g_i | D_i^j = 1, \theta^{(t-1)}\right)}{Pr\left(g_i | \theta^{(t-1)}\right)}, i \in (1, n), D_i^j \in D(g_i) \quad (4)$$

M step: update θ using maximum likelihood estimation approach.

$$\theta^{(t)}(\delta_k, fer, fnr) = \max_{\theta} L(\theta; G, D)$$

$$\delta_k(t) = \frac{\sum_{D_m^n \in S_k} p_m^n(t)}{|S_k|}, \forall k \in (1, m)$$

$$FER(t) = \frac{\sum_{i=1}^{n}(1-g_i)\left[1 - \prod_{D_i^j \in D(g_i)}(1 - D_i^j(t))\right]}{\sum_{i=1}^{n}\left[1 - \prod_{D_i^j \in D(g_i)}(1 - D_i^j(t))\right]}$$

$$FNR(t) = \frac{\sum_{i=1}^{n} g_i \prod_{D_i^j \in D(g_i)}(1 - D_i^j(t))}{\sum_{i=1}^{n} \prod_{D_i^j \in D(g_i)}(1 - D_i^j(t))} \quad (5)$$

Each domain receives a probability score $\delta_j$ indicating its likelihood of being essential. In this study those with $\delta_j \geq 0.9$ were classified as essential. The cutoff value was obtained by minimizing the sum of false positive and false negative essential gene predictions. The perl script of EDP model is provided in **Supplementary files**.

## 2.4 Test EDP Model with different initial values

The results of the EM-algorithm may be different if the initial values of the model are changed. For the EDP Model, three parameters are given by initial values: the probability that one domain is essential $D = \{D_1, D_2, \dots, D_m\}$, the falsely predicted essential rate (FER) and falsely predicted non-essential rate (FNR). The initial value of $D_k$ is set equal to the percentage of essential genes in which domain k can be found. We fixed the initial value of D and tested the influence of different FERs and FNRs on the results of the EDP Model. Six pairs of FER and FNR values were tested, and the final result of the EDP Model was convergent as long as the data set remained constant.

## 2.5 Building the Domain Clustering Matrix

The matrix describes the existence and essentiality of all domains in the six species. The values for the matrix elements are as follows: 1 for an essential domain, -1 for a non-essential domain, and 0 if no domains were found. Here we evaluate the distance between species based on the Mutual Information of the domains' essentiality. For species X and Y, the Mutual Information I(X;Y) is:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \ln\left(\frac{p(x,y)}{p(x)p(y)}\right),$$

and the distance between X and Y is D(X,Y)=1-I(X;Y). The bootstrapping values of each branch are evaluated by randomly selected 400~500 domains to rebuild the evolution tree for 1,000 times.

# 3 RESULTS

## 3.1 Testing the Essential Domain Prediction (EDP) Model using simulated data sets

Based on the Expectation-Maximization (EM) algorithm, we designed the EDP Model (see **Methods** for details) to predict the essentiality of each domain. To test the performance of the EDP Model, we generated 20 independent simulated data sets, each of which contains 1,200 genes and 1,000 domains (**Fig. S3**) and are provided in **Supplementary files**. Previous study has shown that the number of domains for each gene follows a power-law-like distribution (Karev, et al., 2002), and we found the same distribution in the combined gene-domain annotation for the six microbes (**Fig. S4A**). Therefore, we required that the degree of genes and domains in the simulated data sets follow a power-law distribution (**Fig. S4B**). For each data set, we randomly assigned a certain number of essential domains, and then assigned essential genes if and only if the gene contains at least one essential domain. Among the 20 simulated data sets, the number of essential domains ranged from 50 to 83, while the number of essential genes ranged from 120 to 164.

We then applied the EDP Model to these simulated data sets, revealing only the essential gene labels and gene-domain association.

When given different initial values (discussed in **Methods**), the EDP Model produced convergent results for each simulated data set. The results from the 20 simulated data sets were then compared to the original assignment of essential domains (**Fig. S5A**). All pre-assigned essential domains were correctly predicted, i.e., no false negative (FN) predictions, while the false positive rate (FPR) is ~0.4. All false positive (FP) assignments were the results of the same scenario: a domain appeared in only one gene and that gene was essential. Additionally, we annotated the essentiality of genes reciprocally based on the predicted essential domains, and no false predictions were made (**Fig. S5A**).

Next, we added noise to the simulated data and repeated the prediction process. The noise data included 10 genes, five were annotated as essential but contained no essential domains, and the remaining five were annotated as non-essential but contained at least one essential domain. When adding "noise genes" to the data set, the number of domains that a noise gene contained was randomly determined and followed the same degree distribution as the "real genes" in the data set. The errors for repeated predictions are also shown in **Fig. S5A**. For essential domain prediction, the number of FP errors remained almost the same and even decreased in some cases (due to the influence of noise genes). The FPR increased slightly because of the loss of true positives, which became FN errors. Each of these was caused by the addition of noise genes. As for the re-annotation of genes based on predicted essential domains, 16 simulated data sets had no incorrect essential gene assignments and four of them had only one, each of which were noise genes. The number of FN predictions ranged from 8 to 17, including 3-5 noise genes (**Fig. S5B**). The test on simulated data shows that the EDP model offers an accurate prediction of essential domains, even with substantial noise.

## 3.2 Predicting microbial essential domains using the EDP model

After testing the predictive capability of the EDP Model on simulated data sets, we applied it to predict essential domains in real data sets, which included six microbes: *E. coli* (EC), *A. baylyi* (AB), *P. aeruginosa* (PA), *B. subtilis* (BS), *S. cerevisiae* (SC), and *N. crassa* (NC). Essential gene annotation for these species were collected and filtered, excluding genes that do not have annotated domains in PfamA or PfamB (see **Methods**). After filtering, we obtained 23,009 genes and 8,727 domains (3,629 from PfamA and 5,098 from PfamB) in total. The number for each species is shown in **Table 1**. Considering that gene essentiality differs across species, we applied the EDP Model to each organism separately.

We first tested the influence of initial parameters of the iterative EM algorithm on its ability to converge. There are three parameters in the EDP Model: domain essentiality, falsely predicted essential rate (FER) and falsely predicted non-essential rate (FNR) (defined in **Methods**). Specifically, we obtained the initial essentiality of domains by computing the percentage of essential genes associated with each domain. To test the influence of changes in the FER and FNR, we generated six pairs of FERs and FNRs, and compared their convergence process to the final results within each species. As shown in **Fig. S6**, all processes converged within 50 steps, including the pair with an FER of 0.9 and an FNR of 0.8. This result indicated that the iterative process produces stable results for all data sets despite wide variation in the initial FER and FNR values.

Based on the distribution of domain essentiality scores, we set the cutoff for essentiality to 0.9 for the final prediction. Thus, when we generated the set of essential domains, we accounted for 10~23% of the total number of domains in different species (**Table 1**). We found that the number of essential domains was much larger in eukaryotes than prokaryotes. We identified 713 essential domains in SC and 1,009 in NC, while the numbers were 350, 247, 141 and 314 for AB, EC, BS, and PA, respectively. This difference was not caused by the variation in gene numbers; the number of genes in our prokaryotic dataset was similar to that of the eukaryotic dataset. For example, PA and SC have 4,654 and 4,116 genes, respectively. We interpreted this phenomenon to be a consequence of the increased complexity of eukaryotic genomes. In order to increase the number of essential functions in a genome without increasing the size (gene count), there must be an increase in functionally essential "core" components, i.e., essential domains.

## 3.3 The conservation of essential domains

The microbes used in our study differ in both the number and type of their associated genes and domains, as well as the essentiality of these components. To further analyze conservation patterns, we built a phylogenetic tree based on ribosomal RNAs (16S rRNAs for prokaryotes and 18S rRNAs for eukaryotes) (**Fig. S7**) and used this as the gold standard for assessing evolutionary distance among species. Based on this phylogenetic tree, we see that the six microbes form three groups: NC and SC in the fungi group, AB, PA and EC in the Gram-negative group, and BS in the Gram-positive group.

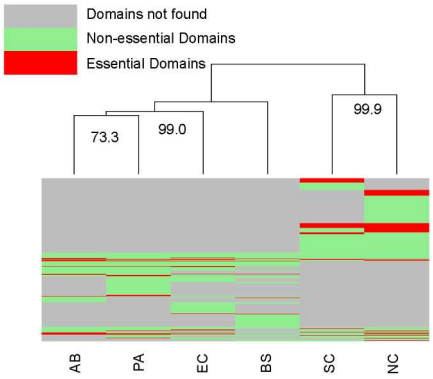


**Figure 1**. Species clusters based on the essentiality and conservation of domains. For each species, a vector for the essentiality of all domains is constructed. Red represents essential domains, green represents non-essential domains, and gray represents domains not found. The Euclidian distance between every two species is calculated based on values discussed in **Methods**.

Concurrently, we clustered these species based on domain essentiality. We used a matrix to represent the association between domains and species, where rows are domains and columns are microbes. As shown in **Fig. 1**, domains are assigned different colors according to their existence and essentiality (see **Methods**). We then applied hierarchical clustering on the columns to group the microbes. The clustering result closely resembled the rRNA-based phylogenetic trees, i.e., AB, EC and PA are closest to each other, BS is the outlier of the prokaryotes, and the fungi group is separated from the prokaryotes. This result suggests that the conservation pattern of domains is highly indicative of the evolutionary distance across species.

The conservation of non-essential domains also contained evolutionary information. We further compared the conservation level between essential and non-essential domains. In this comparison, domain D would be considered as conserved between two species

if it is present in both species without limit on which gene domain D is found. The comparisons were made between every pair of the six species (**Fig. 2A**). In general, essential domains were more conserved than non-essential domains. The odds ratio of the percentage of conserved domains is also correlated with the evolutionary distance among species, that is, more closely related species tend to have a higher odds ratio. For example, the three highest odds ratios were shared among AB, PA and EC, all members of the Gram-negative group. The lowest three odds ratios were found between SC and BS (1.187), NC and BS (1.312), SC and PA (1.47), each comparison between a eukaryote and a prokaryote. This result suggests that essential domains are highly correlated with the species specificity – the more closely related two species are in evolutionary distance, the more essential domains they tend to share rather than non-essential domains.

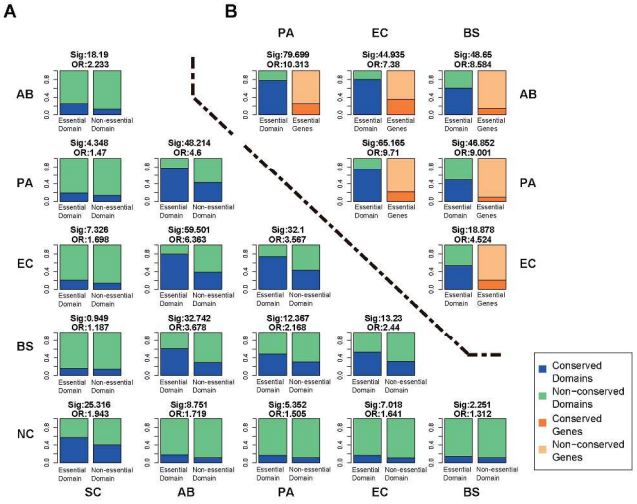


**Figure 2.** Essential domains are more conserved than non-essential domains or essential genes. Comparisons between species pairs are made. The percentage of essential domains conservation is compared with non-essential domains in (A), and with essential genes in (B). Fisher's exact test was used to assess the significance, and test results are labeled in each of the subfigures.

We further show the difference in domain copy numbers between essential and non-essential domains in the six genomes as follows. According to the UniProt database, there are multiple copies of most domains in each of the six species. The number of copies of a domain in a genome is indicative of its prevalence in that species. We compared the detected frequency between essential and non-essential domains in each species using a Wilcoxon Rank Sum test (**Fig. 3**). In all six species, the non-essential domains were present in a significantly higher frequency than essential ones. Furthermore, the occurrence of non-essential domains tended to be more frequent in the eukaryotes (SC and NC). This result provides additional support for the conservation of essential domains. If one considers domains to be the building blocks of proteins, essential domains are the key components that constitute an organism's core functions and would therefore be more likely to be conserved during evolution. On the other hand, non-essential domains are more like decorative blocks and evolve much faster to enable the organisms to diversify and to better adapt to a complex environment, which makes them less conserved across species and present in higher copy numbers within a genome than their essential counterparts.

In our analysis, essential domains were identified based on essential genes, and we found that essential domains are more conserved than essential genes. Here we focused on prokaryotic genes and domains to maintain the conservation of essential genes. Two genes from different bacteria are considered "conserved genes" if they are mutual best hit in reciprocal BLAST and have a sequence similarity higher than 30%. Between any two bacteria, the percentage (intersection to union) of conserved essential domains was higher than essential genes (**Fig. 2B**). The odds ratios were between 4.52 (AB and EC) and 10.31 (PA and BS). Like the relationship between essential and non-essential domains, the difference between essential domains and essential genes were also correlated to the species similarity: the more closely related two species were, the more significant the difference between the number of conserved essential domains and conserved essential genes would be. This result provides additional support to the idea that essential domains are more conserved than essential genes, and are therefore likely more basic units responsible for the essentiality of genes.

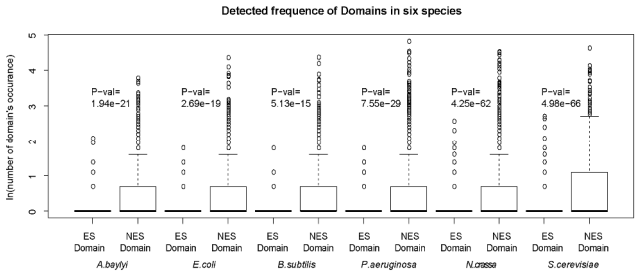


**Figure 3.** Frequency of non-essential domains and essential domains in different species. The detection frequency of each domain in six microorganisms is shown in boxplot. Frequency of essential domains (ES Domain) is compared with non-essential domains (NES Domain) in each species. The Wilcoxon signed-rank test is used to specify the significant difference between essential and non-essential domains in the same species.

## 3.4 Results of single domain genes

From **Table S3** we can see that in all the six species, the result of Fisher's exact test suggests that when we evaluate a gene's essentiality, the single-domain genes would not influence the process. We also compared with a simpler and straightforward method: select all single-domain essential genes and assign these domains as essential. Then we check the percentage of non-essential genes in multi-domain genes which include essential domains. Compared with EDP model (shown in **Table S3**), this simpler method would generate many more false predictions. **Table S4** shows results of paralogous genes in different species. In all six species, although there are only a small number of genes with a differently annotated paralogous gene, they made up the majority of false evaluations on the essentiality of genes.

## 3.5 Functions of essential domains

Compared with non-essential domains, would essential domains be more frequently related with certain biological functions? We performed GO enrichment analysis for domains that are essential for at least one species. In total there are 48 GO terms significantly highly enriched in essential domains (adjusted p-value $< 0.05$, **Table S5**). We grouped these functions in larger branches and show their percentage in essential domains in **Fig. 4**. GO terms which are also enriched in domains that are essential for a given species are shown in the figure. These enriched functions are mainly related with metabolic process, including heterocycle metabolic and organic substance biosynthetic process.

Given that essential domains are correlated with species relatedness, we subsequently investigated whether the functions are related with species. To simplify the analysis and also because these three species have many essential domains in common (**Fig. 2**), we combined the domains in PA, AB, and EC together into Gram-negative group (GN). All domains were then further grouped according to their essentiality in the following species (or groups): GN, BS, NC and SC. For some of these groups, GO enrichment analysis also showed significantly enriched functions related to the corresponding species' feature. For example, the lipid A biosynthetic process (GO:0009245) was enriched in GN-specific essential domains, while the GPI anchor biosynthetic process (GO:0006506) was enriched in NC- and SC-specific essential domains (**Table S6**). In domains that were found non-essential in NC but essential in SC, the transcription factor TFIIA complex (GO:0005672) was enriched. This TF is necessary for the initiation of gene transcription in eukaryotes, but only got one copy in SC and had at least two copies in NC. The DNA polymerase III complex (GO:0009360) was enriched in domains that were only essential in prokaryotes, while the RNA biosynthetic process (GO:0032774) was enriched in conserved essential domains in all groups. Further study of the enriched functions of domains with various conservation levels and essentiality patterns will reveal the characteristics of each species and their unique requirement of essential functions.

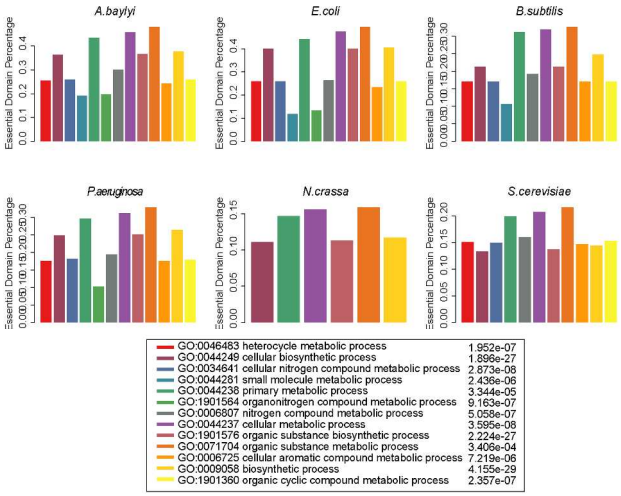


**Figure 4**. The percentage of essential domains and associated GO terms. Domains in each species are mapped to GO terms, and the most frequent GO terms are shown in the figure.

## 3.6 Annotate gene essentiality using essential domain

We identified each domain's essentiality designation using that of the gene(s) to which it belongs. Reciprocally, we tried to reproduce the annotation of essential genes based on these predicted essential domains. The rationale is that if we can accurately reproduce the essential gene annotations based on essential domains, there are two implications: First, the EDP model has made accurate predictions of essential domains. Second, gene essentiality can be adequately explained from a protein domain perspective. For each species, we annotated the essential genes based on essential domains: that is, we assigned to a gene the label "essential" if and only if it contained at least one essential domain. Those genes that did not contain any essential domains were labeled "non-essential". We then compared the assigned essential genes with the observed essential genes in each species. As shown in **Fig. 5**, the "reproduced essentiality" of genes is highly consistent with the observed cases. As a control, we randomly assigned the essential labels to genes in the real data set (labeled "random" in **Fig. 5**), and used the EDP model to produce a set of "essential domains" based on this randomized data set. When comparing the area under the ROC curves (AUCs) of the essential gene estimates using real data versus random data, we found that the essential domains based on the real data set produced significantly more accurate annotation for essential genes (Student's t-test). We also combined all six datasets and tested the prediction on essential genes with a nested cross validation. We randomly take 10% of data out, train the EDP model on the rest 90% with a 10 fold cross-validation and then test the trained model on the previous 10% data. The test is repeated for 10 times, and the result is shown in **Table S7.** This result suggests that gene essentiality can be adequately explained from a protein domain perspective, and thus protein domains are likely more basic units that are responsible for gene essentiality.

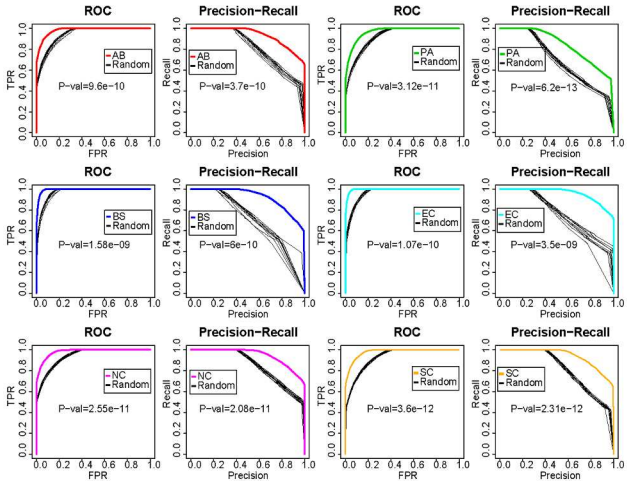


**Figure 5.** Estimates of gene essentiality based on essential domains. A ROC curve and precision-recall curve are shown for both the estimation based on real data (colored) and random data (black) from which the essentiality of genes are randomly assigned. Also shown is t-test that was run between the AUC of real and random data and p-values.

## 3.7 Literature support of the predicted prokaryotic essential domains

We investigated further some of the essential domains we predicted. Below we will present several cases to illustrate how these domains could be used to explain gene essentiality (**Fig. 6**).

The first case includes two domains, PF00294 and PB003571. The domain PB003571 is annotated by PfamB with unknown function but is predicted as an essential domain. All genes contains this domain are essential genes, including BSU35740, PA4996, YGR277C and NCU05311, and all these genes contain only this domain except for PA4996. This result is somehow consistent with a recent report, in which Norman etc. found that many essential genes in microbes contain only one domain without known function(Goodacre, et al., 2014). Also, this domain is conserved among four microbes which differ a lot in taxonomy, indicating a significant important function it may have. At the same time, the other domain PF00294 which is also found in essential gene PA4996 is predicted as non-essential, for its function of pfkB family carbohydrate kinase is not that important, and could be found in

many non-essential genes including b1723, ACIAD1992 and BSU14390.

The second case includes four domains. PF03120 and PF00533 are predicted as essential domains while PF00633 and PB002244 as non-essential. PF03120 is the DNA ligase OB domain, which is necessary for the ligation of DNA fragments during DNA duplication, repair and recombination. These functions are necessary for organisms to survive, and genes contain this domain are all essential genes, such as ACIAD0848 PA1529, b2411 and BSU06620. As for PF00633, the BRCT domain that is conserved in proteins involved in cell cycle checkpoint function is also predicted as essential domain. On the other hand, PF00633 is the helix-hairpin-helix DNA-binding motif, which is related with the interaction between protein and DNA. The function of that domain is not indispensable, since ACIAD0848 could also function in DNA repair process without PF00633's present. And even contains this domain, like BSU25590, won't make it essential as well.

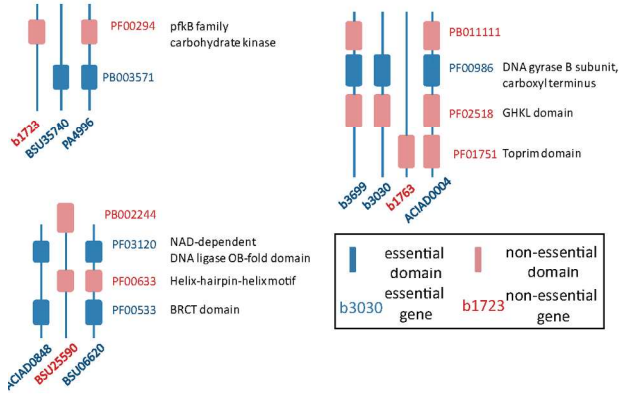


**Figure 6.** Examples that essential domains could explain gene essentiality. The annotation of domains for each gene is shown in lines and boxes, while essential genes and domains are filled with blue.

The last case also includes four domains: PF00986 is the essential one, and PF02518, PF01751 dna PB011111are the non-essential domains. PF00986 is the carboxyl terminus of DNA gyrase B subunit, and is responsible for the ATP-independent relaxation during the complexation of DNA gyrate process. This function is important during DNA duplication, and many genes contains this domains could be inhibited by antibiotics(Engle, et al., 1982). Many genes contains this domain are essential, such as ACIAD004, b3030 and b3699. It's worth noting that in *E.coli* there are two copies of that domain, however the evaluation of its essentiality has not been effected.

Most interestingly, the EDP model seems to be able to tolerate mistakes in the original essential annotations. We found that the PA3987 gene contained the same domain as PA3834, PA4560, b0026, b0642, ACIAD3106 and ACIAD0022. There other genes are all essential for the organisms, and it's odd that PA3987 alone is a non-essential gene. The non-essential annotation was from the data set produced by (Jacobs, et al., 2003). In their experiment, only one transposon insertion, which did not change the reading frame, was observed within PA3987. It is well recognized that a single transposon insertion is often insufficient to disrupt a gene's function, and this can result in false annotation of essential genes (Deng, et al., 2013). Therefore, the annotation based on this transposon mutagenesis is highly likely to be a false-negative error. Our EDP model correctly predicts the PF00133 domain to be essential despite the error in this annotation based on the transposon mutagenesis.

## 4 DISCUSSION

Although protein domains have been sought in determining the mechanistic basis of some essential genes, to our knowledge, large-scale analysis of essentiality from a protein domain perspective has not been attempted. In this study, using well-defined essential gene data sets in six microorganisms, we have developed an EM algorithm-based EDP model to evaluate the gene essentiality from a protein domain perspective.

We demonstrated that, when tested with simulated data, the EDP model was capable of identifying essential domains and could tolerate the disturbance of substantial noise. Further application of this model to both prokaryotic and eukaryotic data sets led to the identification of hundreds of essential domains in each organism. There are several interesting findings based on the analysis of these domains. First, essential domains are highly conserved among species, not only more so than non-essential domains, but to a higher degree than essential genes as well. This result suggests that essential domains may play the role of "functional bricks" during the evolution of genes, which would be constrained through evolutionary pressure and undergo few changes. Second, conserved essential domains perform species-related functions, such as the lipid A biosynthetic process shared among gram-negative bacteria. Third, domain essentiality can be used to accurately reproduce the annotations of essential genes, and can even be used to test the results of the transposon mutagenesis data sets.

The EDP model could be influenced by the following factors and may be further improved for use in our future studies. First, the cross talk between domains may affect the EDP model's accuracy. According to the study of Vogel et al, certain domain pairs might be highly associated to occur in genes together(Vogel, et al., 2004). In these cases, these protein domains are more likely to be estimated as non-essential if their related genes were not all essential. In some extreme situations, the combination of domains might even generate novel functions (Bashton and Chothia, 2007; Dessailly, et al., 2009) and thus result in different essentialities. This would impact the conservation of domain's essentiality when species were analyzed separately, or bring false predictions of essential domains when using combined datasets from multiple genomes.

Second, the estimation of domain essentiality can be improved. For example, in the EDP Model, FER and FNR are estimated based on the entire data set (both essential genes and domains) (Eq. (1)). However, FER and FNR influence each domain's essentiality score when re-estimating the probability for each domain to be essential (Eq. (4)). As a result, applying the EDP Model on two separated gene-domain associations may produce different results when this data is combined and the model is run again.

Third, the performance of the EDP Model is dependent upon the quality of the data set. Although we have shown that it is capable of tolerating noise to some degree, the quality of the input data will inevitably affect the accuracy of the predictions.

Despite some limitations, there are several advantages of an essential domain-centric versus gene-centric approach to gene essentiality: First, essential domains are more basic functional units than essential genes. Many genes consist of multiple domains that carry out distinct functions. Employing an essential domain concept will allow us to pinpoint which function is essential. The seemingly diverse mechanistic basis of essential genes and higher-order essentiality, e.g., synthetic lethality, may be revealed and united. Second, essential domains are more transferable between organisms. EC and PA share <20% of their genes (orthologs), but they share more than 50% of their domains. Third, essential domains are more scalable than genes. Recently, it was discovered that alt-

hough the number of multi-domain architecture families (MDAs) is growing rapidly over time, the number of single-domain architecture families (SDAs) appears to be reaching a saturation point (Levitt, 2009). Nearly all novelty comes from the arrangement of known SDA domains along an MDA sequence (Levitt, 2009).

In summary, with these new insights gleaned from the study of essential domains, it will become possible to discover the fundamental principles that govern gene essentiality, which will lead to the eventual deciphering of the complex relationship between genotype and phenotype. Furthermore, unraveling the mechanistic basis of gene essentiality will provide crucial information on cell physiology and lead to drastic improvements in our ability to efficiently re-engineer microorganisms. This will have applications on many fronts, including energy, biodefense, pharmaceuticals and bioremediation.

## ACKNOWLEDGEMENTS

*Funding*: This work was supported by the exchange program fund of doctoral student under the Fudan University Graduate School (to Yulan Lu).

## REFERENCES

Akerley, B.J., *et al.* A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *Proc Natl Acad Sci U S A* 2002;99(2):966-971.

Arigoni, F., *et al.* A genome-based approach for the identification of essential bacterial genes. *Nature biotechnology* 1998;16(9):851-856.

Baba, T., *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;2:2006 0008.

Bashton, M. and Chothia, C. The generation of new protein functions by the combination of domains. *Structure* 2007;15(1):85-99.

Bruccoleri, R.E., Dougherty, T.J. and Davison, D.B. Concordance analysis of microbial genomes. *Nucleic Acids Res* 1998;26(19):4482-4486.

de Berardinis, V., *et al.* A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Molecular systems biology* 2008;4:174.

Delaye, L. and Moya, A. Evolution of reduced prokaryotic genomes and the minimal cell concept: variations on a theme. *Bioessays* 2010;32(4):281-287.

Deng, J., *et al.* A statistical framework for improving genomic annotations of prokaryotic essential genes. *PloS one* 2013;8(3):e58178.

Dessailly, B.H., *et al.* Exploiting structural classifications for function prediction: towards a domain grammar for protein function. *Current opinion in structural biology* 2009;19(3):349-356.

Dowell, R.D., *et al.* Genotype to phenotype: a complex problem. *Science* 2010;328(5977):469.

Engle, E.C., Manes, S.H. and Drlica, K. Differential effects of antibiotics inhibiting gyrase. *Journal of bacteriology* 1982;149(1):92-98.

Finn, R.D., Clements, J. and Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 2011;39(Web Server issue):W29-37.

Finn, R.D., *et al.* The Pfam protein families database. *Nucleic acids research* 2010;38(Database issue):D211-222.

Forsyth, R.A., *et al.* A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Mol Microbiol* 2002;43(6):1387-1400.

Freiberg, C., *et al.* Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. *J Mol Microbiol Biotechnol* 2001;3(3):483-489.

Gallagher, L.A., *et al.* A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(3):1009-1014.

Gerdes, S., *et al.* Essential genes on metabolic maps. *Curr Opin Biotechnol* 2006;17(5):448-456.

Gerdes, S.Y., *et al.* Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol* 2003;185(19):5673-5684.

Giaever, G., *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 2002;418(6896):387-391.

Gil, R., *et al.* Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* 2004;68(3):518-537.

Glass, J.I., *et al.* Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 2006;103(2):425-430.

Goodacre, N.F., Gerloff, D.L. and Uetz, P. Protein domains of unknown function are essential in bacteria. *mBio* 2014;5(1):e00744-00713.

Hashimoto, M., *et al.* Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Molecular microbiology* 2005;55(1):137-149.

Hastie, T., Tibshirani, R. and Friedman, J.J.H. The elements of statistical learning. Springer New York; 2001.

Hutchison, C.A., *et al.* Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* 1999;286(5447):2165-2169.

Jacobs, M.A., *et al.* Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100(24):14339-14344.

Jaroszewski, L., *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol* 2009;7(9):e1000205.

Jergic, S., *et al.* The unstructured C-terminus of the tau subunit of Escherichia coli DNA polymerase III holoenzyme is the site of interaction with the alpha subunit. *Nucleic Acids Res* 2007;35(9):2813-2824.

Ji, Y., *et al.* Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 2001;293(5538):2266-2269.

Kanaan, S.P., *et al.* Inferring protein-protein interactions from multiple protein domain combinations. *Methods Mol Biol* 2009;541:43-59.

Karev, G.P., *et al.* Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC evolutionary biology* 2002;2:18.

Kato, J. and Hashimoto, M. Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol* 2007;3:132.

Kobayashi, M., *et al.* Bacterial sepsis and chemokines. *Curr Drug Targets* 2006;7(1):119-134.

Levitt, M. Nature of the protein universe. *Proc Natl Acad Sci U S A* 2009;106(27):11079-11084.

Liao, B.Y. and Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* 2008;105(19):6987-6992.

Liberati, N.T., *et al.* An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proc Natl Acad Sci U S A* 2006;103(8):2833-2838.

Mushegian, A. The minimal genome concept. *Curr Opin Genet Dev* 1999;9(6):709-714.

Punta, M., *et al.* The Pfam protein families database. *Nucleic acids research* 2012;40(Database issue):D290-301.

Song, J.H., *et al.* Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. *Mol Cells* 2005;19(3):365-374.

Uchiyama, I., Higuchi, T. and Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic acids research* 2010;38(Database issue):D361-365.

Vogel, C., *et al.* Structure, function and evolution of multidomain proteins. *Current opinion in structural biology* 2004;14(2):208-216.

Wang, L. and Lutkenhaus, J. FtsK is an essential cell division protein that is localized to the septum and induced as part of the SOS response. *Mol Microbiol* 1998;29(3):731-740.

Winsor, G.L., *et al.* Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic acids research* 2011;39(Database issue):D596-600.

Zalacain, M., *et al.* A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. *J Mol Microbiol Biotechnol* 2003;6(2):109-126.

Zhang, Y., *et al.* Comparative mapping of sequence-based and structure-based protein domains. *BMC Bioinformatics* 2005;6:77.