University of Nevada, Reno

Use of Short Amino Acid Motifs in the Computational Analysis of Protein Diversity and Function.

A thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Science in Biochemistry and Molecular Biology, Bachelor of Science in Applied Mathematics, and the Honors Program

by

Alex Maurice Dussaq

Joseph J. Grzymski, Thesis Advisor

May, 2010

UNIVERSITY OF NEVADA RENO **THE HONORS PROGRAM**

We recommend that the thesis
prepared under our supervision by

Alex Maurice Dussaq

entitled

Use of Short Amino Acid Motifs in the Computational Analysis of Protein Diversity and Function.

be accepted in partial fulfillment of the
requirements for the degree of

BACHELOR OF SCIENCE IN BIOCHEMISTRY AND MOLECULAR BIOLOGY,
BACHELOR OF SCIENCE IN APPLIED MATHEMATICS,
AND THE HONORS PROGRAM

____
Joseph J. Grzymski, Thesis Advisor

____
Tamara Valentine, Ph. D., Director, **Honors Program**

May, 2010

Abstract

The explosion of whole genome sequence and environmental sequence data afford us the opportunity to explore protein diversity and protein function. This is particularly exciting given the nascent field of synthetic biology. A comprehensive computational analysis of extant proteins is needed in order to define the limitations on protein structure and diversity from a bioengineering perspective. This paper focuses on defining an upper limit for protein diversity using computational approaches derived from linguistic analyses. These methods are used to make a prediction on the upper limit of unique proteins and number of highly conserved motifs. Motifs deemed highly conserved will, more than likely represent important structural components of basic proteins. Results were gathered from two large data sets: all of the currently available microbial genome sequences available from NCBI and the Global Ocean Survey data set. There were 6.6 million unique proteins at 95% amino acid identity. The majority of unique motifs in these data sets were only found once. The motifs deemed highly conserved in lifestyle groupings of organisms and individual organisms were analyzed for function based on a conserved domain search. The importance between pathogenicity and cell motility and secretion related genes and proteins was observed. These motifs represent potential new drug targets or areas of future experimentation.

Acknowledgements

The following was supported by NIH Grant Number P20 RR-016464 from the INBRE Program of the National Center for Research Resources.

This paper’s contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH.

Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iii

Table Legends iv

Figure Legends v

Introduction 1

Methods 2

(1) Creation and Counting of Motifs 2
(2) Determination of motifs for analysis 5
(3) CDD blast technique 6
(4) Comparison technique 6
(5) Selection of microbial genomes 7
(6) Determination of 70% unique proteins 8

Computer Code 8

(1) p36-countAllMers.pl 8
(2) p-04compareGenomAtoGenomeB.pl 9

Results and Discussion 10

(1) Defining protein diversity on motif appearances. 10
(2) Species lifestyle as defined by motif usage. 13
(3) Selection of microbial genomes 15

Conclusions and Future Work 17

Literature Cited 20

Tables 23

Figures 30

Table Legends

Table 1: Data summary for four major datasets used in analysis of motif and protein diversity.

Table 2: Analysis of databases used for motif analysis. Data from the four major datasets was analyzed for motif usage as described in methods section 1.

Table 3: Table of predictions for motif saturation points. Based upon the six-parameter exponential saturation curve $\frac{M}{M_{\max}} = A \cdot e^{-\frac{a}{p}} + B \cdot e^{-\frac{b}{p}} + C \cdot e^{-\frac{c}{p}}$ and the data collected from methods section 1.

Table 4: Major functional categories in gene annotation.

Table 5: Genera distribution of original dataset.

Table 6: Genera distribution of equalized dataset.

Table 7: List of organisms used for motif based life style comparison. Generated as described in methods section 5.

Figure Legends

Figure 1: Number of protein clusters resulting from 70% clustering with cd-hit. Figure was generated as described in **Methods (6)**. The number of proteins sequenced is on the x-axis and the number of 70% unique protein clusters are on the y-axis.

Figure 2: Cumulative distribution of 6mers. Saturation curved of number of proteins (x-axis) and number of motifs occurring at least once (y-axis). The three databases used are indicated by the figure legend. The randomly generated database based on equal amino acid frequencies, assumes all amino acids have an equal probability of occurring, 5%.

Figure 3: Cumulative distribution of 7mers. As in figure 2 but for the motif length of 7mer.

Figure 4: T-score distribution for observed versus expected values of varying motif lengths for *Actinobacillus pleuropneumoniae serovar*. Values were derived as described in **Methods 2**. $pval \leq 0.25$ was used for determination of what is an overrepresented motif.

Figure 5: Comparison of functional protein categories between *Actinobacillus pleuropneumoniae serovar* and *Orientia tsutsugamushi Boryong* based on motif usage. Conserved motifs of 6-8 amino acids in length were matched to functional categories (**Methods 3**) and usage was compared (**Methods 4**).

Figure 6: Comparison of functional motif usage between several fundamental phylogenetic bacterial groups. Data are the percent of over occurring motifs dedicated to a given functional group according to a BLAST of the CDD

database. Values themselves are calculated from the average of the logarithm base two of the group on the right over the group on the left (**Methods 4**). Final values and errors are based on the average of motifs of six through eight amino acids in length.

Introduction

Proteins are on average 300 amino acids long— with 20 possible amino acids this represents more possible combinations than there are atoms in the observable universe. The future of synthetic protein design is dependant on parameterizing the possible combinations of sequence space. Proteins are not random assortment of short motifs; they are highly organized structurally with short random assortments of motifs. It has been known for over thirty years that amino acids follow non-random distributions (Black, Jarkins and Stenzel 1976). Despite this, the actual structural limitations of proteins are still largely unknown. As ordered as proteins are in three dimensions, amino acid usage follows basic thermodynamic principles and cost minimization. Organisms tend to use smaller amino acids (by mass) more frequently than larger ones (Barrai, Violina and Scapoli 1994) (Dufton 1997). The hypothesized reason for this is a combination of decreasing metabolic costs and increasing genetic stability (Dufton 1997). Looking for these sorts of correlations has driven the field of bioinformatics for the majority of the past 35 years; however, with the recent improvements in sequencing techniques and the explosion of data available more robust statistical analyses are possible. A moderate correlation between conserved motifs, or small segments of proteins, and overall protein sequence similarity indicated that motif analysis could be used in an effort to understand protein function (David, et al. 2003). Recent work correlated 3D domains with protein functionality and protein uniqueness and concluded that protein structural diversity had plateaued (Jaroszewski, et al. 2009). Despite these findings we still cannot answer the question of where protein diversity (based on sequence similarity or function) reaches its limit. The approach taken in these computational analyses attempts to measure uniqueness of sequence on the motif level. Projections are made for the number of unique motifs of varying sizes. The purpose of this is not only to describe protein diversity, but also to look at highly conserved motifs. Highly conserved motifs would logically lend themselves to overall protein functionality, whereas less conserved motifs are more likely to describe protein uniqueness. By discovering and quantifying the frequency of conserved motifs (words of N size) it is possible to define standard parts that could be used in the creation of overall protein structure, in order to maximize the efficiency of synthetic biology and bioengineering.

Methods

(1) Creation and Counting of Motifs

(a) Datasets

The following data sets were used: (1) 832 fully sequenced Bacteria and Archaea genomes (NCBI 2009) clustered with cd-hit (Li and Godzik 2009) to remove redundant sequences with greater than 95% similarity. (2) All 6.1 million protein fragments from the global ocean survey database (GOS) (The J. Craig Venter Institute 2009), also clustered at 95% similarity. (3) A combination of the two databases above clustered with a 95% similarity maximum. (4) All sequenced Bacteria and Archaea were grouped according to broad lifestyle categories (free-living/pathogenic gram-positive/gram-negative organisms); these databases came from the clustered individual organisms (1), but were not clustered once proteins were conglomerated, more in **Methods (5)**. (5) Randomly generated database to mimic (3), based on approximate amino acid frequencies.

(b) Randomization of protein order

When necessary in order to eliminate variable unique motif growth caused by grouping of similar proteins, proteins were placed into random order using BioPerl module: Bio::DB::Fasta (Stein 2001) to access all proteins without placing them into memory, then Perl module List::Util (Barr 1997) to randomize the order of the proteins themselves (**Computer Code 1b**).

(c) Creation of motifs

All proteins were broken into motifs. Motifs in these contexts are defined as short amino acid segments of a set length, an Nmer, where N=2, 3…. Motifs were formed by starting at every residue on the protein and going N amino acids over. All motifs with an unknown amino acid were ignored. (**Computer Code 1**)

(d) Counting of 2mers-6mers and small datasets

For all datasets small enough to be taken into memory, the motifs counted using Perl's hashing technique. As the number of proteins analyzed increased the number of unique motifs was counted. Unique motifs are defined as the different motifs that occur at least once in a given dataset. In addition at certain points in the data collection the distribution of motifs was recorded. (**Computer Code 1b**)

(e) Counting of 7mers-12mers for large datasets

A significant achievement in the analysis of large data sets was achieved by implementing a parallel procedure for counting very large lists of data. A technique was developed whereby lists of at least 1.2 billion motifs could be efficiently counted. Briefly the technique works as follows: for datasets too large to be taken into memory the counting was handled in pieces. This was done by creating files of approximately five million motifs. These lists were then counted and sorted in parallel. In order to add files together they were each opened and the motifs were added together based on alphabetical order. In order to make this a more efficient method the files were added in sets. In other words file1 was added to file2 at the same time file3 was added to file4. Then file1+2 can be added to file3+4. In this fashion $\lceil \log_2(n) \rceil \cdot \frac{n}{2c}$ iterations can be preformed instead of n; where n is the number of files. In this manner as long as the number of files is less than 65,536 for 16 cores or 256 for 8 cores the addition of files will be more rapid than adding them in sequence. In a similar method as *(d)* the number of unique motifs and the distribution data was recorded at certain key points in the data. (**Computer Code 1c**)

(f) Analysis of protein diversity

Protein diversity was estimated by fitting the count of unique motifs as a function of protein number to a six-parameter exponential saturation curve:

$$\frac{M}{M_{\max}} = A \cdot e^{-\frac{a}{p}} + B \cdot e^{-\frac{b}{p}} + C \cdot e^{-\frac{c}{p}}$$

Where M is the number of motifs found, $M_{\max}$ is the theoretical maximum number of motifs, 20^n (n being the motif length), A, a, B, b, C, c the six parameters fit by MATLAB, and p the number of proteins analyzed.

In order to estimate the number of possible proteins from the number of motifs the following formula was used:

$$x_i \leq x_j^{\,i-j+1}$$

Where x_i is the percent of possible motifs of length i. The reasoning behind this is as follows: The number of ways the motifs can be placed in a protein of length i is, ignoring overlap restrictions, $x(j)^{i-j+1}$ where $x(j)$ is the number of motif of length j in existence. This does not yield the number of possible proteins since overlap is excluded, however if each one of these places is limited to x_j of the total plausible jmers then the final set must be limited by a factor of x_j^{i-j+1} as well.

<u>(2) Determination of motifs for analysis</u>

Motifs were created from database input as described in **Methods (1c)** and counted as described in **Methods (1d)**. Expected motif frequencies were calculated in two ways based on probability theory, firstly by the multiplication of components:

$$P\left(x_1x_2x_3x_4\cdots x_n\right)=P\left(x_1\right)\cdot P\left(x_2\right)\cdot P\left(x_3\right)\cdot P\left(x_4\right)\cdots P\left(x_n\right)$$

Where x_i represents an amino acid and $x_1x_2x_3x_4\cdots x_n$ represents motifs in a set order. An obvious problem with this is that it relies heavily on the amino acids all being independent of one another. Since this is not the case a second method of calculation was used:

$$P\left(x_1x_2x_3x_4\cdots x_n\right)=\frac{P\left(x_1x_2\cdots x_{n-1}\right)\cdot P\left(x_2x_3\cdots x_n\right)}{P\left(x_2x_3\cdots x_{n-1}\right)}$$

Note that this formula still assumes some level of independence between component motifs; however the idea of expectation is based on motifs of slightly shorter length, not on the occurrence of individual amino acids. This causes the set of over-expected Nmers to be less identical to the over expected (N-1)mers, than would be created by the previous method. This is caused by the fact that an over-expected 4mer, $x_1x_1x_1x_1$, would more than likely indicate the over-expected 5mer, $x_1x_1x_1x_1x_i$ in the first method versus the second method where the expected value is based on the 4mer frequency.

Once expected values were calculated over-expected motifs were chosen for analysis based on a one-tailed Student's t-Test with a p-value≤ 0.25. Over-expected motifs will be referred to as conserved motifs. (**Computer Code 2**)

(3) CDD blast technique

Conserved motifs of length 6-8 as determined by **Methods(2)** were BLASTed against the CDD database (Marchler-Bauer, et al. 2009) using an rpsblast (Altschul, et al. 1990). The CDD database contains conserved protein domains from the following conserved domain databases: cd, pfam, smart, COG, KOG, PRK, TIGR, and LOAD_. Using this database allows a reduction in random matches since this database only uses domains that are considered functional which reduces the number of random coils and similar DNA portions that potentially skews results; this is important considering the large e-values, on average 300 for 6mers 250 for 7mers and 8mers, necessarily returned by short motifs. The BLAST returned results that contained functional categories describing general cellular functions detailed in table 4; these categories were parsed from all results that matched at least 80% of the motif length. The count associated with functional categories was determined by the number of occurrences of the corresponding motif in the original database. (**Computer Code 2**)

(4) Comparison technique

Functional categories were counted using the techniques described in **Methods (3)**. Following this the counts were translated into percentages and compared other data sets (e.g. gram positive free-living organisms were compared to gram-positive pathogens). The term $P_{i,A,n}$, is defined as the percent of functional category i dedicated to bacterial group A for motif length n. The following formulae were used for the analysis:

$$V_{i,n} = \log_2\left(\frac{P_{i,A,n}}{P_{i,B,n}}\right)$$

$$\bar{V} = \log_2\left[\left(V_{i,6} \cdot V_{i,7} \cdot V_{i,8}\right)^{1/3}\right]$$

$$s^2 = \left(2^{\sigma^2} - 1\right)2^{2\cdot\bar{V}+\sigma^2}$$

Where σ^2 is based on a normal distribution. These data were plotted on vertical bar graph in order to represent over-expression for certain functional categories among groups. (**Computer Code 2**)

<u>(5) Selection of microbial genomes</u>

The 649 sequenced bacterial genomes that could be split into lifestyle categories were obtained from NCBI then were split into four lifestyle categories, free-living/pathogenic gram-positive/gram-negative organisms. Inconsistencies between the four groups necessitated equalizing the groups. This was done in several stages. First all organisms belonging to the same species were removed, at random, until only one of each species remained. The same process was repeated for all genera with more than four species until each genus was represented by only four or fewer organisms. Seventeen genera were then chosen at random until the number of organisms was greater than or equal to thirty-four. If the number of organisms was less than thirty-four, a genus with only one organism was removed and another was chosen at random from the list. Following this thirty-four organisms were chosen from the list, making sure that all genera still only represented by one organism were selected. This was done in order to insure four lists with 34 organisms and 17 genera. Protein conservation was determined by counting the number of clusters found in a 70% similarity grouping by cd-hit (Li and Godzik 2009) before and after the grouping. The reduced dataset was used for all subsequent comparisons between groups.

(6) Determination of 70% unique proteins

Each data set was clustered at 70% identity in order to determine how many potentially unique proteins. This is a proxy to unique function. Proteins from combined Bacteria, Archaea and GOS dataset were randomly ordered as in **Method (1b)**. These proteins where then used to generate files with incrementally larger numbers of proteins. The files were clustered using cd-hit with a 70% similarity cut-off. These results were tabulated and graphed in comparison to a 1:1 protein to cluster line.

Computer Code

(1) p36-countAllMers.pl

Number of lines: 36

Description: Takes in a list of protein files in FASTA format and runs three programs on said files. The first program (a) is used to calculate amino acid frequency. The second program (b) is used to calculate the number of unique 2mers-6mers along with distribution of motifs. Finally the third program (c) is used to calculate the number of unique 7mers-12mers.

(a) p34-AAFreqForFile.pl

Number of lines: 117

Description: Takes in a protein file in FASTA format and iterates through the proteins counting the occurrences of each amino acid. The results are output to a new file.

(b) p35-count1t6Motifs.pl

Number of lines: 358

Description: Takes in a protein file in FASTA format, randomizes the order, and then proceeds to make a large file with all motifs. This file is then taken into memory using hashing, while at certain key points the distribution of motifs and the number of unique motifs found so far are output to files. This creates between 2 and 14 protein dictionaries and distributions.

(c) p32-countAllMers.pl

Number of lines: 1065

Description: Takes in a randomized protein file in FASTA format, splits the proteins into motifs and puts 5 million motifs into as many files as necessary to cover all motifs. Each file is counted and alphabetically sorted in parallel. Then files are added in groups using parallel processing (Methods1e) in order to decrease the number of necessary iterations.

(2) p-04compareGenomAtoGenomeB.pl

Number of lines: 1142

Description: Takes in a protein file in FASTA format, splits the proteins into motifs of size 4-8 (in separate files) then uses 4mers and 5mers to calculated expected values for 6mers (**Methods 4**). These values were compared to actual occurrences of motifs using a one-sampled Student's t-test. Motifs with a $pval \leq 0.25$ are then rpsblasted against the CDD database (**Methods 3**) this is repeated in parallel with another number of protein files for comparison as describe in **Methods 4**. These results were graphed using small python codes automatically.

Results and Discussion

(1) Defining protein diversity on motif appearances.

Using a relatively small computer cluster and the open source computer languages of Perl and python tools were developed to analyze very large: genome, protein and motif data sets. The combined Bacteria, Archaea and GOS data set had 6.6 million proteins of an average length of 247 amino acids. **Table 1** details many of the statistics for the GOS and the combined microorganism databases, as well as the weighted random dataset created to mimic a complete protein database. From **Table 1** the combined dataset of 15.2 million proteins yields only 6.6 million 95% unique proteins, this is retention of only 43.4% of sequenced proteins. This indicates a large amount of data redundancy, which is where the necessity to cluster the data comes from. If the same proteins have been sequenced more often then a motif may appear to be falsely important.

The 43% unique proteins at 95% is seen even more dramatically in **Figure 1** which represents the number of proteins that are 70% unique. At 0.1 million there are 0.08 million unique protein clusters (84%), where as at 15.2 million there are only 3.6 million unique protein clusters (24%). The unique protein curve digressed from nearly 1:1 to a 1:4 ratio of proteins sequenced to new proteins found. This curve is saturating at such a rate that sequencing will soon reach a point where very few novel proteins can be found from sequencing. According to these results and simple curve fitting to an exponential distribution indicates a 1:100 ratio of new proteins to sequenced proteins after 53 million proteins

Protein diversity was also computed by estimating the number of motifs possible for a given length and extending that to the number of proteins. In a perfect system there are approximately 20^{247} or approximately $2.26 \cdot 10^{321}$ possible proteins. This is an incredible number, and is not possible to reach given that there are approximately 10^{80} atoms in the observable universe (Villanueva 2009). However given some percent of Nmers it is possible to calculate the maximum according to that mathematical estimate using the following formula:

$$x_i \leq x_j^{\,i-j+1}$$

Where x_i is the percent of motifs found with length i. Data extension of the 6mer (**Methods (1f)**) dataset indicates approximately 98% (**Table 3**) of possible motifs will be found (**Figure 2**). This would correlate to 0.75% of possible 247mers being found. This correlates to the upper bound on the number of proteins in existence. Although this number is still $1.70 \cdot 10^{319}$ possible proteins this represents a reliable but high upper bound. Prediction of 7mers indicates an upper bound of 57%, which would bring the total number of possibilities down to $3.31 \cdot 10^{262}$. Even these significant reductions do not bring the limit down to reasonable levels. Barring more data it seems unreasonable to attempt to predict protein primary sequence diversity based solely on motif diversity.

This high upper bound does not make for reasonable predictions, however we can safely assume that not all motifs can occur in every location. Let us assume that in any given position 25% of the possible motifs can be found. In addition to the overlap of motifs must contain motifs from the set itself, with a minimum overlap of one amino acid there are 20^{j-1} possible combinations out of 20^j possible jmers correcting for the number of overlaps j we assume that only $j/_{20}$ can be found for any given overlap we can say:

$$x_i \approx \left(x_j \cdot 0.25 \cdot {}^{j}\!/_{20}\right)^{i-j+1}$$

Using this we can say that there are approximately $1.35 \cdot 10^{48}$ possible proteins based on the number of 6mers, better yet, however, we can say that there are only 35 million possible proteins based on the projected number of 7mers. 35 million is a far more reasonable limit for protein diversity, and given the current rate of protein discovery at least 70 million proteins would need to be sequenced to find this number of 95% unique proteins given the current 1:2 ratio of novel proteins to sequenced proteins. Since this ratio will increase as more proteins are sequenced, the number of proteins necessary for proteome completeness is most likely significantly larger than this.

An upper boundary to protein diversity can help to classify the primary sequence functional level and define possible ways for modular protein design in the field of synthetic biology. Discussed below is an attempt at classifying bacterial lifestyle based on motif usage, however when attempting to identify motifs of certain function purely numerically it may prove more useful to group motifs based on function. For instance the motif $x_1x_2x_3x_4x_5x_6x_7$ could be simplified to $x_1x_2x_ix_4x_ix_6x_7$, where x_i is any amino acid if all motifs of that form are linked to a certain function. It is well accepted that these sorts of mutations take place in the genome of organisms. This will significantly decrease the motif pool and help parameterize our estimates of protein diversity and novel function. Defining these parts will allow for direction in the synthesis of proteins by describing the potential building blocks.

A base set of key functional motifs is important for applications to synthetic biology, knowing without doubt what this set is could prove very useful. In order to determine approximately when all existent 7mers will be found a linear approximation was made. The approximation was the projected number of proteins necessary to reach

57% of 7mers was estimated using a linear extension of the data present since the data is largely linear (**Figure 3**). Based on this linearization the number of proteins needed is approximately 10 million. This is necessarily an under-estimation as the linear approximation quickly outgrows the actual curve.

(2) Species lifestyle as defined by motif usage.

Are there motifs that appear in specific groups of organisms, species or lifestyles that define a specific group? The relationship between conserved motifs that were over-observed in organisms and their relationship to functional categories was explored based on the methods described in section 2. Conserved motif are defined as occurring more often than expected with a corresponding $pval \leq 0.25$ based on an independent one sampled Student's t-test. This p-value was used based on the t-score distribution to ensure a significant number of data points above the cut-off while still creating some level of statistical significance. **Figure 4** represents the distribution of t-scores for motifs of 6-8 amino acids for *Actinobacillus pleuropneumoniae serovar*. The distribution of 6mers maintains the generally accepted normal distribution required for t-scores; however the bell curve loses shape for the larger motif sizes. This causes very few motifs to be considered conserved, with 386,782 conserved 6mers versus only 740 conserved 8mers. Considering the significantly larger number of possible 8mers with the corresponding low number of conserved motifs these data create large statistical variance. The conserved motifs, once determined, were analyzed to determine relations to protein functional categories (**Methods 3**). Although there are at least thirty different cog categories, the top twenty were chosen for analysis, the details of which can be seen in **Table 4**. These results were then compared to similar results for another randomly selected microorganism, *Orientia tsutsugamushi Boryong*. The comparison was described in methods section 4. Results are seen in **Figure 5**.

A. pleuropneumoniae is a pathogenic Gammaproteobacteria responsible for fibrinohaemorrhagic pneumonia (Sebunya and Saunders 1983). *O. tsutsugamushi* is a pathogenic Alphaproteobacteria that causes scrub typhus (Leelarasamee, et al. 2004). The large variance of the functional categories (**Figure 5**) causes most of the categories to return non-significant results. However, there were differences between *A. pleuropneumoniae* and *O. tsutsugamushi* with the secondary metabolites biosynthesis, transport and catabolism, post translational modification, nucleotide transport and metabolism, DNA replication, recombination and repair, and amino acid transport and metabolism functional categories all favoring *O. tsutsugamushi*; while the inorganic ion transport and metabolism and RNA processing and modification functional categories favor *A. pleuropneumoniae*. The two most significant differences based on values and variance were a 31% difference favoring post translational modification for *A. pleuropneumoniae* and a 20% difference favoring inorganic ion transport and metabolism for *O. tsutsugamushi*.

With this method of motif analysis and assignment to functional categories we can detail motifs and proteins under significant selection pressure compared to other groups. Closer inspection of motifs in the post translational modification functional category revealed several 6mers that matched multiple chaperone proteins in *A. pleuropneumoniae*. GIDLGT and KRLIGR each matched three proteins in *A. pleuropneumoniae* and one in *O. tsutsugamushi*. These motif matches to chaperones and similar examples in the broader category of posttranslational modification are favored in

A. pleuropneumoniae as compared to *O. tsutsugamushi*. The biological reasoning for this is unclear; however, these sorts of patterns can help to identify possible differences between certain groups and between organisms. We can speculate that posttranslational modification motifs are under greater conservative pressure in are under greater pressure in *A. pleuropneumoniae*. The importance of these motifs to *A. pleuropneumoniae* could be tested in a laboratory situation, however due to the over expected appearance in the chaperone proteins one would expect a significant loss of functionality with significant changes to these motifs.

Similar analysis of motifs from inorganic ion transport and metabolism revealed the motif ILVGLF to be important in three sodium ion pumps in *O. tsutsugamushi*, but not even found in *A. pleuropneumoniae*. Sodium ion pumps are inarguably important to bacterial homeostasis and as such this motif may represent a potential antibiotic target. In order to determine the validity of this argument the location and importance of the motif must be determined in 3D modeling, to determine the accessibility, in conjecture with laboratory experiments to determine the importance to function, nonetheless this does represent one of the potential uses for protein classification by motifs.

<u>(3) Selection of microbial genomes</u>

Analysis of the original 649 fully sequenced bacterial genomes indicated several inconstancies between lifestyle groups. The four groups, free-living/pathogenic gram-positive/gram-negative, had between 80 and 247 genomes per group, with the number of genera ranging from 17 to 114 (**Table 5**). In addition to this, analyses of the groups in the original data set indicated proteins in pathogenic gram positive organisms had more than twice as many redundant sequences than the free-living counterparts. In order to check if this was due to the unequal distribution of genomes and organisms, the groups were equalized as described in methods section 5. Once the datasets were equalized (**Table 6**) protein conservation levels became far less distinct, with pathogens having only 1.3x as many redundant sequences as compared to the free-living counterparts. This reduced dataset is used in all future data analyses.

Using the methods outlined in sections 2-4 motifs were compared among the groups of organisms detailed in **Table 7**. In the comparison between pathogenic and free-living organisms the functional category, cell motility and secretion, showed an average difference of 2.2% favoring the pathogens, and amino acid transport and metabolism showed an average difference of 2.7% favoring the free-living organisms. In the comparison between gram-positive and gram-negative organisms signal transduction showed an average difference of 6.6% favoring gram positive organisms, and a 5.1% difference favoring gram-negative organisms in carbohydrate transport and metabolism (**Figure 6**).

The overrepresentation of cell motility and secretion reflects the necessity of these pathways for pathogenicity. Many pathogens such as *Clostridium difficile* rely on the release of cytotoxins for pathogenicity (Borriello 1998) or the ability to burrow into the epithelial lining such as several gut based pathogens (Fischer, Prassl and Haas 2009).

Closer inspection of motifs for cell motility and secretion in pathogens returned the motif AAKTIDR which is linked to twitching motility proteins across all four groups. Although this does not differentiate groups it does identify an important motif for motility. The fact that it is not only conserved in pathogens, but in free-living organisms as well is indicative of the importance of this motif to overall motility. Given interest in cell motility useful data may result from the study the affects of mutations to this motif.

Although there are countless numbers of motifs available for similar analyses in order to extract sincerely useful information a hypothesis may be presented. For example if a protein was of interest it would be possible to compare its motifs to the databases and select the most conserved motifs for mutation experiments, allowing for more precision in mutation experiments. A full survey of protein motifs could yield interesting relationships between motif usage and lifestyle, however due to the vastness of this particular dataset this sort of analysis is difficult without testing a specific hypothesis.

These hypothesis all demand significant laboratory testing to validate, however, they could lead to practical applications. For instance the identification of protein domains that are important to pathogenic bacteria and not to host organisms or free-living organisms could lead to the development of more targeted antibiotics. This is becoming increasingly more important in modern medicine with the advent of antibiotic resistance and the significant block of research indicating the importance of free-living microorganisms in overall health (Savage 2001), especially in the gut.

Conclusions and Future Work

In 1995 The Institute for Genomic Research sequenced the first full bacterial genome, less than 15 years later we have over 800 fully sequenced bacterial genomes, well over 15 million bacterial protein sequences, several fully sequenced human genomes and countless other sources of genomic data. This explosion of data has changed biology from a field driven by laboratory breakthroughs into a field desperate for a way to analyze the incredible amounts of data that accumulate. The advent of new statistical methodology and computer science techniques into biology has aided greatly in this analysis, however the analytical potential of bioinformatics is still in its infancy.

In this thesis I analyzed 15 million proteins from two major databases: accrued Bacteria and Achaea genomes (NCBI 2009) and the global ocean survey database (The J. Craig Venter Institute 2009). This was accomplished in part by creating over 100 programs in the computer languages of Perl, python and MATLAB. My analysis focused on motifs: small portions of protein sequences. The goals of my thesis were to discover two major things, (1) a limitation to the number of proteins in existence and (2) a new link between primary sequence and protein function.

The significance of this lies in directing other fields of biology. Synthetic biology, which focuses on the creation of proteins, could be greatly benefited by knowing the limitations that nature has created in the forms of possible functional units and protein diversity. In order to create novel proteins intelligently there must be a toolkit of allowed combinations, something I started to define. In addition to this the common practice of mutating proteins, thought to be key to biochemical pathways or lifestyles such as pathogenicity, may be more directed in both selecting of the key proteins and in the selecting of mutation sites.

For example, in several recent papers type VI secretion systems have been shown to be key to pathogenicity (Mougous, et al. 2006) (Pukatzki, et al. 2006). Burtnick et. al. discovered that the type VI secretion system was essential to *Burkholderia mallei* by mutating the T6SS-1 gene (Burtnick, et al. 2010). This required the creation of several different mutant strains as well as the significant testing to determine the importance of this pathway to pathogenicity. Using the databases and methods created here we have shown again the importance of secretion systems to pathogenicity through over-representation of conserved motifs in this category. This comparison alone does not directly support the importance of this particular secretion pathway, however this places statistical emphasis on this direction.

In order to determine proteins of high importance in pathways, pathogenicity or other aspects of biology we could use these tools to develop a system based on number and conservation level of conserved motifs that could define a robust scoring systems. The proteins with the highest scores are hypothesized to hold high importance in the group chosen. Following this, mutations of the most important proteins could be selected based off a similar scoring system that ranks potentially important residues. This work cannot be done intelligently, however, without further study of motif diversity. We still need to define the motif size (if one exists) that is both linked to functionality and acts at the level of fitness and selection. However, the steps taken here provide both a theoretical and operational framework for future work.

Literature Cited

Altschul, S F, W Gish, W Miller, and E W Myers. "Basic local alignment search tool." *J. Mol. Biol.* 215 (1990): 403-410.

Barr, G. *List::Util - A selection of general-utility list subroutines.* 1997. http://search.cpan.org/~gbarr/Scalar-List-Utils-1.23/lib/List/Util.pm.

Barrai, I, S Violina, and C Scapoli. "The usage of oligopeptides in proteins correlates negatively with molecular weight." *Int J Pept Protein Res* 45 (1994): 326-331.

Black, J A, R N Jarkins, and P Stenzel. "Non-random relationships among amino acids in protein sequences." *Int J Pept Protein Res.* 8 (1976): 125-130.

Borriello, S P. "Pathogenesis of Clostridium difficile infection (Review)." *Journal of Antimicrobial Chemotherapy* 41 (1998): 13–19.

Burtnick, M N, D DeShazer, V Nair, FC Gherardini, and P J Brett. "Burkholderia mallei Cluster 1 Type VI Secretion Mutants Exhibit Growth and Actin Polymerization Defects in RAW 264.7 Murine Macrophages." *INFECTION AND IMMUNITY* 78 (January 2010): 88-99.

CAMERA. *CAMERA - Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis.* 2009. http://camera.calit2.net/ (accessed January 2009).

David, L, M Silver, R C Edgar, and D R Livesay. "Using Motif-Based Methods in Multiple Genome Analyses: A Case Study Comparing Orthologous Mesophilic and Thermophilic Proteins." *Biochemistry* 42 (2003): 8988-8998.

Dufton, M J. "Genetic Code Synonym Quotas and Amino Acid Complexity: Cutting the Cost of Proteins?" *J. theor. Biol.* 187 (1997): 165-173.

Fischer, W, S Prassl, and R Haas. "Virulence Mechanisms and Persistence Strategies of the Human Gastric Pathogen Helicobacter pylori." *Current Topics in Microbiology and Immunology*, 2009: 129-171.

Jaroszewski, L, et al. "Exploration of Uncharted Regions of the Protein Universe." *PLoS Biol* 9 (2009).

Leelarasamee, A, C Chupaprawan, M U Chenchittikul, and S Udompanthurat. "Etiologies of acute undifferentiated febrile illness in Thailand." *J Med Assoc* 87 (2004): 464–472.

Li, W, and A Godzik. *CD-HIT.* 2009. http://bioinformatics.ljcrf.edu/cd-hi/ (accessed 2009).

Marchler-Bauer, A, et al. "CDD: specific functional annotation with the Conserved Domain Database." Database (Jan 2009): D205-10.

Mougous, J D, et al. "A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus." *Science* 312 (June 2006): 1526-1530.

NCBI. *NCBI National Center for Biotechnology Information* . January 29, 2009. http://www.ncbi.nlm.nih.gov/ (accessed January 29, 2009).

Pukatzki, S, et al. "Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system." *Proceedings of the National Academy of Sciences of the United States of America* 103 (JAN 2006): 1528-1533.

Savage, D C. "Microbial biota of the human intestine:a tribute to some pioneering scientists." *Curr. Issues. Intest. Microbiol.*, 2 2001: 1–15.

Sebunya, T N, and J R Saunders. "Haemophilus pleuropneumoniae infection in swine: A review." *J Am Vet Med Assoc* 182 (1983): 1331–1337.

Stein, L. *Bio::DB::Fasta.* 2001. http://search.cpan.org/~birney/bioperl-1.4/Bio/DB/Fasta.pm (accessed June 1, 2009).

The J. Craig Venter Institute. *A collection of articles from The J. Craig Venter Institute's Global Ocean Sampling Expedition in PLoS Biology.* 2009. http://collections.plos.org/plosbiology/gos-2007.php (accessed 2009).

Villanueva, J C. *Atoms in the Universe.* July 30, 2009. http://www.universetoday.com/guide-to-space/the-universe/atoms-in-the-universe/.

Tables

Table 1: Data summary for four major datasets used in analysis of motif and protein diversity.

	GOS	**Prokaryote Genomes**	**Combined Dataset**	**Weighted Random**
Pre-Cluster at 95%				
Number Of Proteins	6,115,750	9,110,345	15,226,095	n/a
GC Content	39.082%	52.723%	n/a	n/a
GC Content Variance	0.005%	1.749%	n/a	n/a
Number Of Organisms	n/a	832	n/a	n/a
Number Of Collection Sites	57	n/a	n/a	n/a
Post-Cluster at 95%				
Number of Proteins	4,101,959	2,563,135	6,621,965	6,621,965
Average Protein Length	208.55	308.42	246.88	247
Protein Length Standard Deviation	117.12	259.78	191.91	0
Amino Acid Use				
Alanine	7.08%	9.99%	8.45%	8.00%
Arginine	4.57%	0.98%	5.19%	5.00%
Asparagine	5.52%	5.51%	4.64%	4.00%
Aspartic acid	5.62%	6.12%	5.56%	6.00%
Cysteine	1.08%	3.92%	1.03%	1.00%
Glutamic acid	6.03%	7.63%	6.07%	6.00%
Glutamine	3.16%	2.08%	3.39%	4.00%
Glycine	6.91%	5.95%	7.24%	8.00%
Histidine	1.90%	4.73%	1.98%	2.00%
Isoleucine	7.87%	10.17%	6.94%	7.00%
Leucine	9.48%	2.36%	9.79%	9.00%
Lysine	7.13%	3.69%	5.97%	6.00%
Methionine	2.11%	4.61%	2.23%	3.00%
Phenylalanine	4.68%	3.67%	4.31%	4.00%
Proline	3.92%	5.89%	4.24%	4.00%
Serine	7.14%	5.99%	6.58%	6.00%
Threonine	5.29%	5.41%	5.34%	6.00%
Tryptophan	1.17%	7.11%	1.21%	1.00%
Tyrosine	3.24%	1.27%	3.08%	3.00%
Valine	6.12%	2.92%	6.58%	7.00%

Table 3: Table of predictions for motif saturation points. Based upon the six-parameter exponential saturation curve $\frac{M}{M_{\max}} = A \cdot e^{-\frac{a}{p}} + B \cdot e^{-\frac{b}{p}} + C \cdot e^{-\frac{c}{p}}$ and the data collected from methods section 1.

Motif Length	Percent of possible actually found.	Percent of possible predicted to be found.
4	100%	100%
5	99.96%	100%
6	89.68%	98%
7	34.45%	57%

Table 4: Major functional categories in gene annotation.

Functional Category	**Abbreviation**
Translation	Translation
Transcription	Transcription
Signal transduction	Signal Transduction
Secondary metabolites biosynthesis, transport and catabolism	Secondary metabolites
RNA processing and modification	RNA
Post translational modification	Post trans mod
Nucleotide transport and metabolism	Nucleotide Meta/trans
Lipid transport and metabolism	Lipid meta
Intracellular trafficking and secretion	Intracellular
Inorganic ion transport and metabolism	Inorg ion trans/meta
Energy production and conversion	Energy prod/conv
Defense mechanisms	Defense
DNA replication, recombination and repair	DNA replication
Coenzyme transport and metabolism	Coenzyme meta
Cell motility and secretion	Cell motility/ secretion
Cell envelope biogenesis, outer membrane	Cell envelope
Cell division and chromosome partitioning	Cell division
Carbohydrate transport and metabolism	Carb trans/meta
Amino acid transport and metabolism	Amino acid meta

Table 5: Genera distribution of original dataset.

	Free-	Free+	Path-	Path+
Total Organisms	214	80	247	108
Number of genera	114	42	58	17
Average number of organisms per genus	1.88	1.90	4.26	6.35
Lowest number of organisms in a genus	1	1	1	1
Highest number of organisms in a genus	14	9	17	29

Table 6: Genera distribution of equalized dataset.

	Free-	Free+	Path-	Path+
Total Organisms	34	34	34	34
Number of genera	17	17	17	17
Average number of organisms per genus	2.00	2.00	2.00	2.00
Lowest number of organisms in a genus	1	1	1	1
Highest number of organisms in a genus	4	4	4	4

Table 7: List of organisms used for motif based life style comparison. Generated as described in methods section 5.

Free-Living Gram-Negative

- Candidatus_Azobacteroides_pseudotrichonymphae_genomovar__CFP2-PID29025
- Candidatus_Blochmannia_floridanus-PID443
- Candidatus_Desulfococcus_oleovorans_Hxd3-PID18007
- Candidatus_Pelagibacter_ubique_HTCC1062-PID13989
- Coprothermobacter_proteolyticus_DSM_5265-PID30729
- Desulfovibrio_desulfuricans_G20-PID329
- Desulfovibrio_vulgaris_Hildenborough-PID51
- Geobacter_bemidjiensis_Bem-PID17707
- Geobacter_lovleyi_SZ-PID17423
- Geobacter_metallireducens_GS-15-PID177
- Geobacter_uraniumreducens_Rf4-PID15768
- Hahella_chejuensis_KCTC_2396-PID16064
- Magnetospirillum_magneticum_AMB-1-PID16217
- Nitrosomonas_europaea-PID52
- Nitrosomonas_eutropha_C71-PID13913
- Pelodictyon_luteolum_DSM_273-PID13012
- Pelodictyon_phaeoclathratiforme_BU_1-PID13011
- Prosthecochloris_aestuarii_DSM_271-PID12749
- Prosthecochloris_vibrioformis_DSM_265-PID12607
- Rhodoferax_ferrireducens_T118-PID13908
- Rhodospirillum_centenum_SW-PID18307
- Rhodospirillum_rubrum_ATCC_11170-PID58
- Shewanella_ANA-3-PID13905
- Shewanella_frigidimarina_NCIMB_400-PID13391
- Shewanella_sediminis_HAW-EB3-PID18789
- Shewanella_woodyi_ATCC_51908-PID17455
- Solibacter_usitatus_Ellin6076-PID12638
- Synechococcus_CC9605-PID13643
- Synechococcus_PCC_7002-PID28247
- Synechococcus_WH_7803-PID13642
- Synechococcus_elongatus_PCC_6301-PID13282
- Thermodesulfovibrio_yellowstonii_DSM_11347-PID30733
- Thermosipho_africanus_TCF52B-PID27767
- Thiomicrospira_crunogena_XCL-2-PID13018

Free-Living Gram-Positive

- Acidothermus_cellulolyticus_11B-PID16097
- Anoxybacillus_flavithermus_WK1-PID28245
- Arthrobacter_aurescens_TC1-PID12512
- Arthrobacter_chlorophenolicus_A6-PID20011
- Candidatus_Desulforudis_audaxviator_MP104C-PID21047
- Clostridium_acetobutylicum-PID77
- Clostridium_cellulolyticum_H10-PID17419
- Clostridium_novyi_NT-PID16820
- Clostridium_thermocellum_ATCC_27405-PID314
- Corynebacterium_efficiens_YS-314-PID305
- Corynebacterium_glutamicum_R-PID19193
- Corynebacterium_jeikeium_K411-PID13967
- Dehalococcoides_BAV1-PID15770
- Dehalococcoides_ethenogenes_195-PID214
- Deinococcus_geothermalis_DSM_11300-PID13423
- Deinococcus_radiodurans-PID65
- Dictyoglomus_turgidum_DSM_6724-PID29175
- Geobacillus_kaustophilus_HTA426-PID13233
- Geobacillus_thermodenitrificans_NG80-2-PID18655
- Lactobacillus_acidophilus_NCFM-PID82
- Lactobacillus_delbrueckii_bulgaricus-PID16871
- Lactobacillus_fermentum_IFO_3956-PID18979
- Lactobacillus_sakei_23K-PID13435
- Listeria_innocua-PID86
- Listeria_welshimeri_serovar_6b_SLCC5334-PID13443
- Mycobacterium_JLS-PID16079
- Mycobacterium_KMS-PID16081
- Salinispora_arenicola_CNS-205-PID17109
- Salinispora_tropica_CNB-440-PID16342
- Streptomyces_avermitilis-PID189
- Streptomyces_coelicolor-PID242
- Streptomyces_griseus_NBRC_13350-PID20085
- Symbiobacterium_thermophilum_IAM14863-PID12994
- Thermoanaerobacter_pseudethanolicus_ATCC_33223-PID13901

Pathogenic Gram-Negative

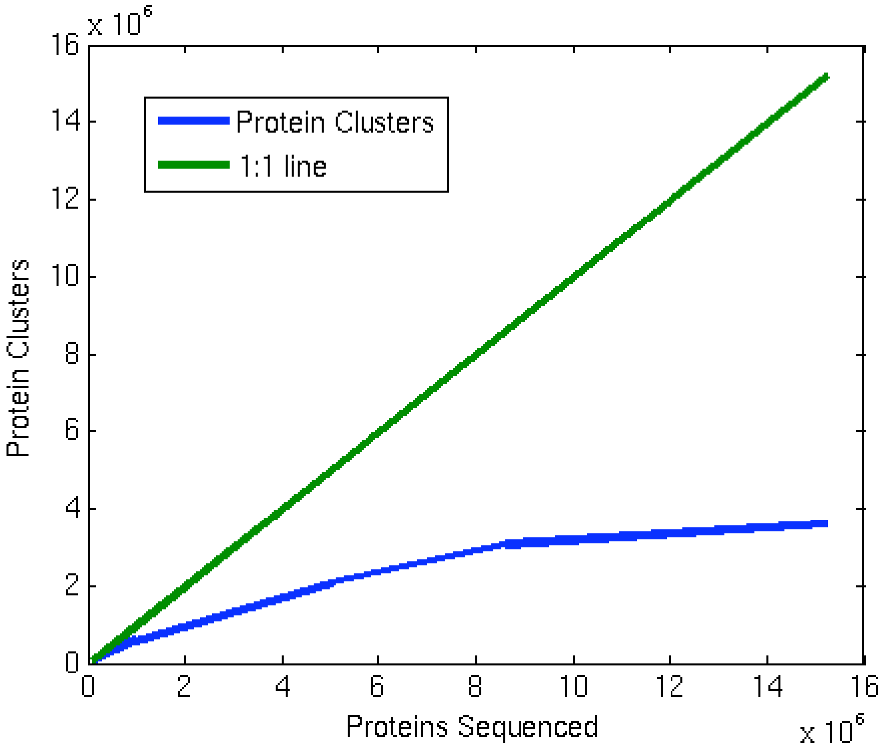
- Aeromonas_hydrophila_ATCC_7966-PID16697
- Aeromonas_salmonicida_A449-PID16723
- Agrobacterium_tumefaciens_C58_Cereon-PID283
- Agrobacterium_vitis_S4-PID13372
- Aliivibrio_salmonicida_LFI1238-PID30703
- Bordetella_bronchiseptica-PID24
- Bordetella_parapertussis-PID25
- Bordetella_pertussis-PID26
- Borrelia_garinii_PBi-PID12554
- Borrelia_recurrentis_A1-PID18233
- Borrelia_turicatae_91E135-PID13597
- Brucella_abortus_9-941-PID9619
- Brucella_canis_ATCC_23365-PID20243
- Brucella_melitensis-PID180
- Brucella_suis_ATCC_23445-PID20371
- Chlamydia_muridarum-PID229
- Helicobacter_acinonychis_Sheeba-PID17251
- Helicobacter_hepaticus-PID185
- Helicobacter_pylori_HPAG1-PID16183
- Leifsonia_xyli_xyli_CTCB0-PID212
- Parachlamydia_sp_UWE25-PID10700
- Porphyromonas_gingivalis_W83-PID48
- Proteus_mirabilis-PID12624
- Pseudomonas_aeruginosa_PA7-PID16720
- Pseudomonas_entomophila_L48-PID16800
- Pseudomonas_mendocina_ymp-PID17457
- Pseudomonas_syringae_phaseolicola_1448A-PID12416
- Rickettsia_conorii-PID42
- Stenotrophomonas_maltophilia_K279a-PID30351
- Xanthomonas_campestris_8004-PID15
- Xanthomonas_citri-PID297
- Xanthomonas_oryzae_MAFF_311018-PID16297
- Yersinia_pestis_Antiqua-PID16645
- Yersinia_pseudotuberculosis_YPIII-PID28743

Pathogenic Gram-Positive

- Bacillus_anthracis_Ames_0581-PID10784
- Bacillus_cereus_B4264-PID17731
- Bacillus_thuringiensis_Al_Hakam-PID18255
- Bacillus_weihenstephanensis_KBAB4-PID13623
- Bacteroides_vulgatus_ATCC_8482-PID13378
- Clavibacter_michiganensis_NCPPB_382-PID19643
- Clostridium_botulinum_E3_Alaska_E43-PID28855
- Clostridium_difficile_630-PID78
- Clostridium_perfringens-PID79
- Clostridium_tetani_E88-PID81
- Corynebacterium_diphtheriae-PID87
- Corynebacterium_urealyticum_DSM_7109-PID29211
- Enterococcus_faecalis_V583-PID70
- Listeria_monocytogenes_HCC23-PID29409
- Lysinibacillus_sphaericus_C3_41-PID19619
- Mycobacterium_abscessus_ATCC_19977T-PID15691
- Mycobacterium_bovis-PID89
- Mycobacterium_marinum_M-PID16725
- Mycobacterium_smegmatis_MC2_155-PID92
- Parabacteroides_distasonis_ATCC_8503-PID13485
- Propionibacterium_acnes_KPA171202-PID12460
- Renibacterium_salmoninarum_ATCC_33209-PID19227
- Staphylococcus_aureus_RF122-PID63
- Staphylococcus_epidermidis_RP62A-PID64
- Staphylococcus_haemolyticus-PID12508
- Staphylococcus_saprophyticus-PID15596
- Streptococcus_gordonii_Challis_substr_CH1-PID66
- Streptococcus_sanguinis_SK36-PID13942
- Streptococcus_suis_98HAH33-PID17155
- Streptococcus_uberis_0140J-PID353
- Thermobifida_fusca_YX-PID94
- Tropheryma_whipplei_Twist-PID95
- Ureaplasma_parvum_serovar_3_ATCC_27815-PID19087
- Ureaplasma_urealyticum-P

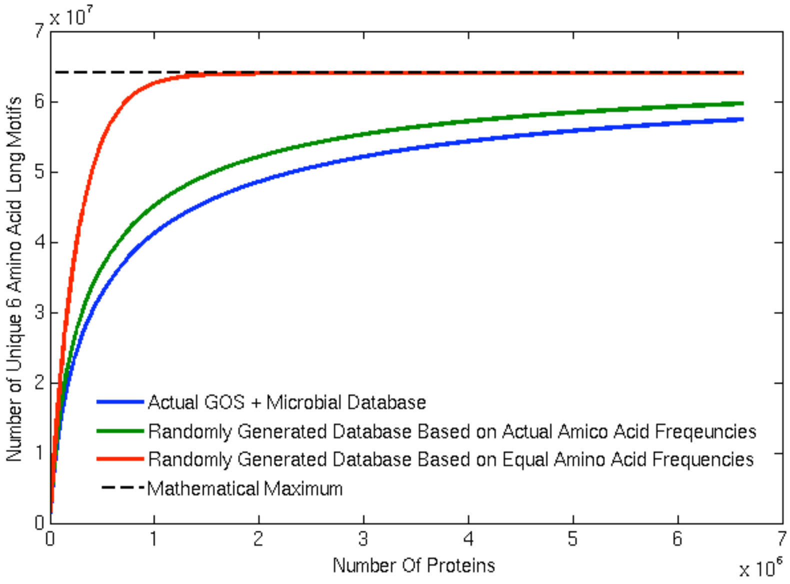
Figures

Figure 1



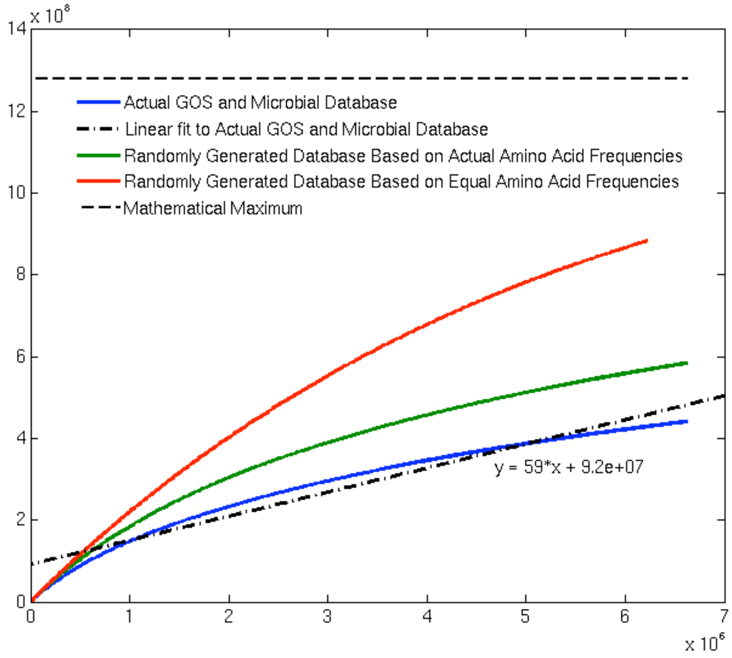
(a)

Figure 2


x 10^7
7
6
5
4
3
2
1
0
Number of Unique 6 Amino Acid Long Motifs
Actual GOS + Microbial Database
Randomly Generated Database Based on Actual Amico Acid Freqeuncies
Randomly Generated Database Based on Equal Amino Acid Frequencies
Mathematical Maximum
0
1
2
3
4
5
6
7
Number Of Proteins
x 10^6

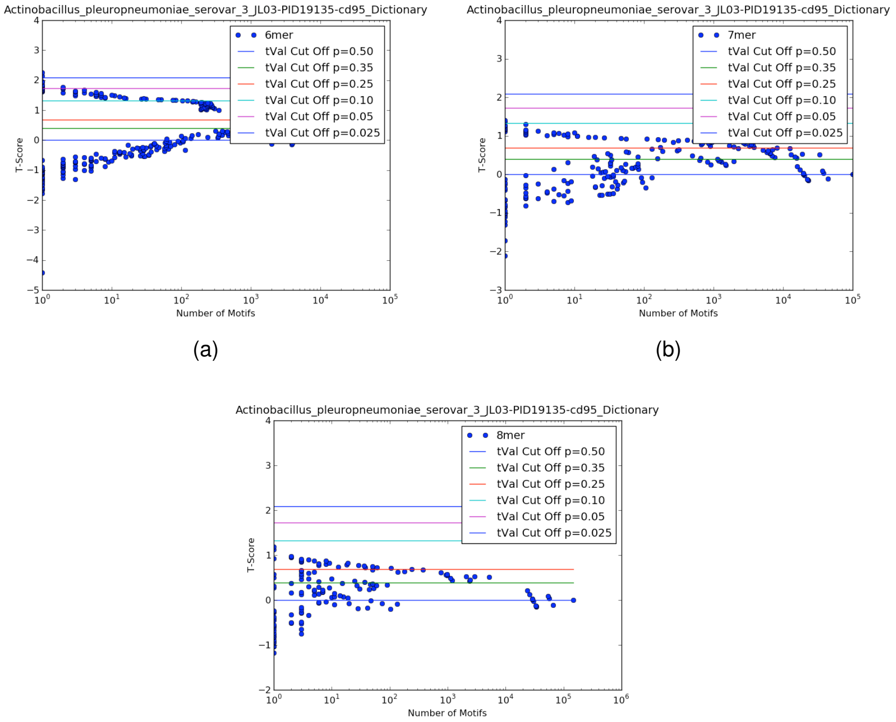
(a)

Figure 3


x 10^8
14
12
10
8
6
4
2
0
Actual GOS and Microbial Database
Linear fit to Actual GOS and Microbial Database
Randomly Generated Database Based on Actual Amino Acid Frequencies
Randomly Generated Database Based on Equal Amino Acid Frequencies
Mathematical Maximum
y = 59*x + 9.2e+07
0
1
2
3
4
5
6
7
x 10^6

(a)

Figure 4

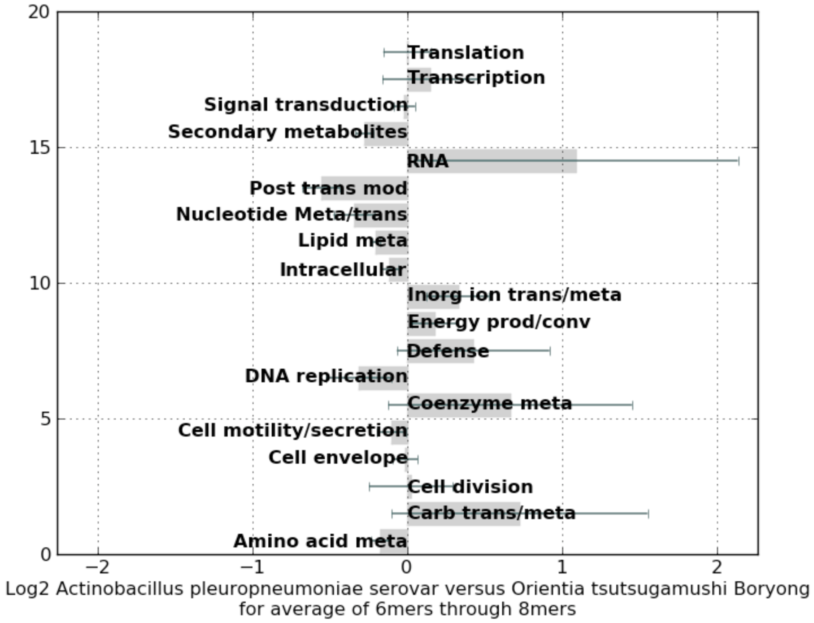

Actinobacillus_pleuropneumoniae_serovar_3_JL03-PID19135-cd95_Dictionary
T-Score
Number of Motifs
4
3
2
1
0
−1
−2
−3
−4
−5
10^0
10^1
10^2
10^3
10^4
10^5
6mer
tVal Cut Off p=0.50
tVal Cut Off p=0.35
tVal Cut Off p=0.25
tVal Cut Off p=0.10
tVal Cut Off p=0.05
tVal Cut Off p=0.025
Actinobacillus_pleuropneumoniae_serovar_3_JL03-PID19135-cd95_Dictionary
T-Score
Number of Motifs
4
3
2
1
0
−1
−2
−3
10^0
10^1
10^2
10^3
10^4
10^5
7mer
tVal Cut Off p=0.50
tVal Cut Off p=0.35
tVal Cut Off p=0.25
tVal Cut Off p=0.10
tVal Cut Off p=0.05
tVal Cut Off p=0.025
Actinobacillus_pleuropneumoniae_serovar_3_JL03-PID19135-cd95_Dictionary
T-Score
Number of Motifs
4
3
2
1
0
−1
−2
10^0
10^1
10^2
10^3
10^4
10^5
10^6
8mer
tVal Cut Off p=0.50
tVal Cut Off p=0.35
tVal Cut Off p=0.25
tVal Cut Off p=0.10
tVal Cut Off p=0.05
tVal Cut Off p=0.025

(a)

(b)

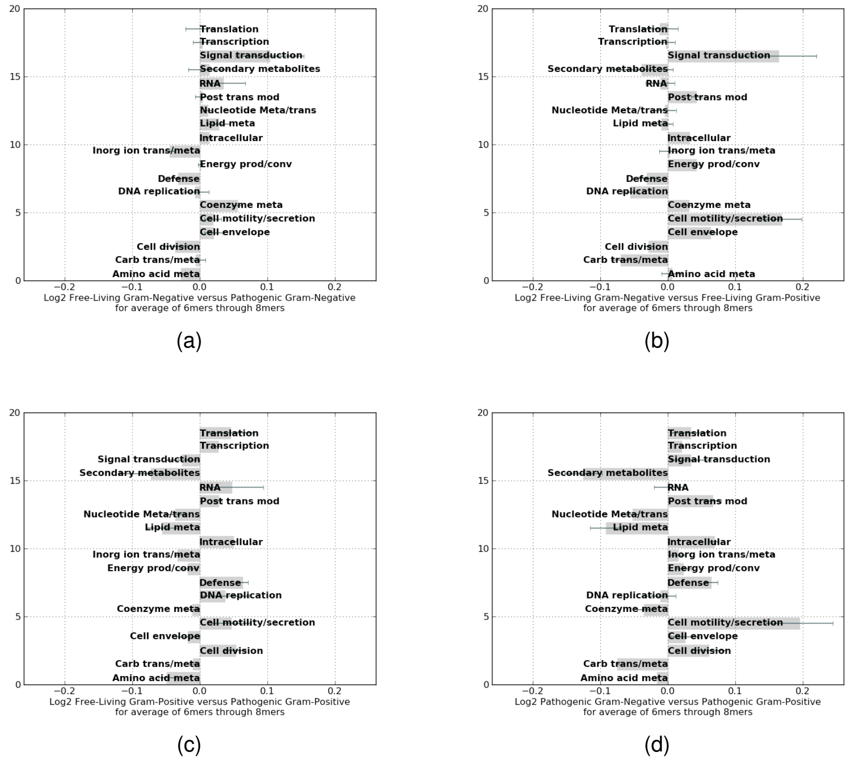
(c)

Figure 5


20
15
10
5
0
−2
−1
0
1
2
Translation
Transcription
Signal transduction
Secondary metabolites
RNA
Post trans mod
Nucleotide Meta/trans
Lipid meta
Intracellular
Inorg ion trans/meta
Energy prod/conv
Defense
DNA replication
Coenzyme meta
Cell motility/secretion
Cell envelope
Cell division
Carb trans/meta
Amino acid meta
Log2 Actinobacillus pleuropneumoniae serovar versus Orientia tsutsugamushi Boryong
for average of 6mers through 8mers

(a)

Figure 6



(a)

(b)

(c)

(d)