Inter and Intra Organism Patterns in Proteomic Sequences

Ankit Kr. Pathak#1, Pragya Jaiswal#2, Prashant Sinha#3, Simmi#4

#Cluster Innovation Centre, University of Delhi
Delhi, India
1ankitkrpathak@yahoo.com, 2pragya.jswl@gmail.com,
3prashant@ducic.ac.in, 4simmimourya@ymail.com

Abstract— **This paper presented here analyses the inherent patterns that arise between various classes of the organisms. We use several statistical and machine learning tools on the proteomic sequences of various organisms and viruses to re-establish known results that were obtained through other experiments. We also investigate if there exists some relation between viral proteome sequences and their hosts. Furthermore, we investigate the significance of very small proteomic sequences.**

Keywords: Proteomes, Taxonomy, Viral Targets, Mycobacterium Tuberculosis, kMean, Principle Component Analysis, Correlation Coefficient, Parallel Plot techniques, Unsupervised Learning.

I. Introduction

The proteome is the entire set of proteins expressed by a genome, cell, tissue or an organism at a certain time under defined conditions. Proteins are a linear chain of amino acids called the polypeptide chain. There are 20 such known amino acids, which can code for various proteins. The sequence of these amino acids plays a vital role in deciding the structure and function of a protein. The work presented here takes into account the proteomic sequences and uses the statistical tools tho establish our hypothesis.

II. Hypothesis

To investigate the relevance of the Proteomic Sequences, we define following three hypotheses:

A. Reestablishment of Evolutionary Taxonomy

Evolutionary taxonomy [1] has classified organisms on the basis of phylogenetic relationship, progenitor-descendant relationship, and degree of evolutionary change. In our first hypothesis, we aim to re-establish the eukaryotic classification using the organism's proteome sequences. Further, observing the bacterial and viral proteomic sequences may validate this result.

B. Prediction of Virus targets through Proteomic Sequences

As there must be a reason to why certain organisms are hosts to only certain viruses, we aim here to see if this reason could be a relation in their amino acid content. The hypothesis set here is that, the reason certain viruses have a particular subspecies or phyla as their hosts can be attributed, to some extent, to the similarities in the amino acid content of those set of viruses and their host set.

C. Significance of small Proteomic Sequences

Biological events do not occur at random. Thus, the hypothesis tends to check whether some sequences of five amino acids have multiple occurrences in the proteome of Mycobacterium Tuberculosis, and if found, their significance is queried for the known motifs of the organism and elsewhere.

III. Methodology

We employed different methodologies for the sub-hypothesis that were defined in previous section.

A. (Hypothesis 1) Reestablishment of Evolutionary Taxonomy

We obtained a sample of 59 organism's (Eukaryotes) proteome sequences from `Ensembl Genome Browser` [2], 17 viral[1] proteome sequences and 18[2] bacterial proteome sequences from `Uniprot Catalog`. The resulting sequence was pre-processed and the amino acid counts were calculated. We used several statistical tools to establish our hypothesis.

1) Pre-processing:

The proteome sequences for the sample were processed to find out the respective count of each of the 20 (and undetermined 21st) amino acids. We define a count vector V_C that stores this respective count in a 21 dimensional vector.

$$V_C = [a_1 \quad \cdots \quad a_{21}]$$

$$a = \{x: x_i = n(sequence) \forall i \in [1, 21]\}$$

Further, we scaled the vector V_C to get the relative ratios of each amino acid. This was done due to the fact that the counts of individual amino acids change drastically between different organisms (on the basis of their biological complexity). We define the scaled count vector V_{SC} as:

$$V_{SC} = [v_1 \quad \cdots \quad v_{21}]$$

$$v = \left\{x: x_i = \left(\frac{a_i}{\sum_{j=1}^{21} a_j}\right) \forall i \in [1, 21]\right\}$$

[1] 17 viruses sampled randomly from 429 Viruses.

[2] 18 bacteria, sampled randomly from 629 Bacteria.

We use vector V_{SC} for all the statistics under this hypothesis.

2) Plotting the Vectors:

Due to the large dimension (21) of the vector V_{SC} we used the parallel plot technique [3] to obtain the line charts. This was done for a visualization of the relatively high dimensional vectors. The parallel axes were set to a common scale.

3) Pearson product-moment correlation coefficient:

To quantize the observation in the plots obtained we calculated four sets of correlation coefficients [4]. The sets are:

- Correlation of Eukaryotic Proteome against itself,
- Correlation of Bacterial Proteome against itself,
- Correlation of Viral Proteome against itself, and
- Correlation of Eukaryotic Proteome against Viral and Bacterial Proteome.

Further, the mean of the correlation coefficients obtained was calculated. A higher value of this mean would indicate a greater closeness between the vectors. A lower value of the correlation coefficient, on the other hand, would indicate the inherently distinct behaviour of the data.

4) k-Mean Clustering:

To further strengthen the closeness between the vectors, we ran a k-mean partitioning [5] with all vector samples. A k-Mean object was trained using the concatenation of Eukaryotic Proteomes. By re-running the samples on the k-mean fitted object and counting the number of respective partitions obtained, we were able to judge the clustering property of the vectors. Lower number of partitions would indicate greater cluster formation.

5) Principle Component Analysis:

We calculated the first two Principal Components [6] of each vector classes separately. Different components for different classes of vectors indicated the data's dissimilarity or similarity based on the coordinates they clustered at.

B. (Hypothesis 2) Prediction of Virus targets through Proteomic Sequences

Viruses cannot survive in isolation; they need a host in which they can survive. The above sample of 429 viruses, obtained for our previous hypothesis, was further tagged according to their hosts. The two subsamples considered were the viruses that attack Bacteria and those that attack *Homo sapiens* (Humans). Alongside, the proteome sequence of *Homo sapiens* and bacteria was taken for comparison. The sample length of bacteria was 250, obtained after random sampling from 692 bacteria.

The pre-processing of the data follows from the previous section.

1) Plot:

The transformed data of a sample of 250 bacteria, 49 viruses and *Homo sapiens*, 117 viruses was plotted and the mean of each was calculated for an initial visualization of a general trend in each of the samples. Parallel plots were obtained with 20 amino acids (and the undetermined 21st) as X-axis and the ratio of the amino acid in the organism as the Y-axis. For comparison, the line plot of the mean of bacteria was plotted alongside the mean of the viruses that attack bacteria. A similar operation was performed for *Homo sapiens* and the viruses that attack *Homo sapiens*.

2) k-Mean Clustering:

To quantitatively analyse the closeness or similarities in the plots, we applied the k-mean clustering algorithm. Three different k-Mean objects were fitted with the concatenation of the following transformed vectors.

- Bacteria and the viruses that attack bacteria
- Homo sapiens and the viruses that attack homo sapiens
- Viruses that attack Bacteria and those that attack Homo sapiens.

Predicting the cluster of the respective samples with their trained k-mean object and then counting the number of respective partitions obtained would determine the similarities. A partition indicates the similarities in that data. This was done for all the three k-mean objects above.

In this case, the favourable result is the attainment of no clean clusters from the data that would imply that there is enough similarity in the data for the algorithm to completely tag it in a particular set.

C. (Hypothesis 3) Significance of small Proteomic Sequences

Past studies have established the generality of this central principle of biochemistry that sequence of amino acids in protein specifies conformation. The dependence of conformation on sequence is significant because of the connection between conformation and function of a protein.

TABLE I

Amino acid	ala	Cys	Leu	Met	Glu	Gln	His	lys	Val	Ile	Phe	Tyr	Trp	Thr	Gly	Ser	Asp	Asn	Pro	Arg
ɑ̣ helix	**1.29**	**1.11**	**1.30**	**1.47**	**1.44**	**1.27**	**1.22**	**1.23**	0.91	0.97	1.07	0.72	0.99	0.82	0.56	0.82	1.04	0.90	0.52	0.96
ß sheet	0.90	0.74	1.02	0.97	0.75	0.80	1.08	0.77	**1.49**	**1.45**	**1.32**	**1.25**	**1.14**	**1.21**	0.92	0.95	0.72	0.76	0.64	0.99
Turn	0.78	0.80	0.59	0.39	1.00	0.97	0.69	0.96	0.47	0.51	0.58	1.05	0.75	1.03	**1.64**	**1.33**	**1.41**	**1.28**	**1.91**	**0.88**

It is seen that residues such as alanine, glutamate, and leucine tend to be present in a helices, whereas valine and isoleucine tend to be present in ß strands. Glycine, asparagine, and proline have a propensity for being in turns.

The proteome data of Mycobacterium Tuberculosis was taken in the FASTA format from `UniProt database` of proteomes. All the possible amino acids sequences of length five and their respective number of occurrences in the proteome were processed. The sequences were ordered in the decreasing order of their frequencies.

The resulting sequences with highest frequencies were analysed and then compared to known motifs of *Mycobacterium Tuberculosis*.

IV. Results and Discussion

The following sections summarize the results obtained for the three hypotheses.

A. (Hypothesis 1) Reestablishment of Evolutionary Taxonomy

1) Plots:

The plots show a clear pattern formation for the Eukaryotes, Bacteria, and Viruses and it can be clearly observed that the three classes of data are following a respective pattern (Refer to Figure 1 and Figure 2).

2) Pearson product-moment correlation coefficient:

The value of correlation coefficient is higher for Eukaryotic proteome than the Bacterial and Viral proteome. The following table summarises the result of PPMCC test. An important result here is that the values obtained here conform to the arguments put up from the plots.

TABLE II
Correlation Coefficients and Interpretation

Candidate	Value	Interpretation	Comments
Eukaryotic Proteome against itself	`0.9737`	High correlation (closeness) between the vectors. (Rank 1)	Explains the closeness as obtained in the data.
Bacterial Proteome against itself	`0.8813`	Less correlation between the vectors. (Rank 2)	Explains observable irregularities in the plotted vectors.
Viral Proteome against itself	`0.8283`	Lesser correlation between the vectors. (Rank 4)	Explains the high irregularities in the vectors.
Eukaryotes Proteome against Viral and Bacterial Proteome	`0.8802`	Less correlation between the vectors. (Rank 3)	Explains the irregularities. Strengthens the hypothesis that the vectors are deviating.

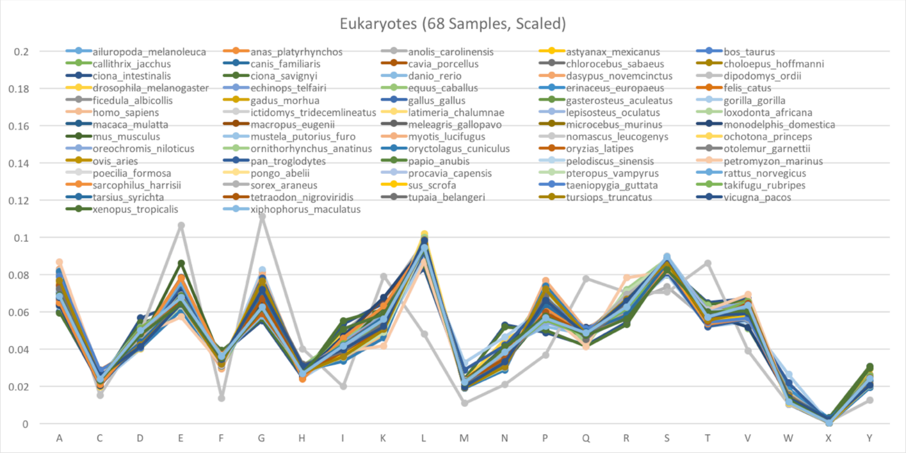

Fig. 1- Parallel plot of the Eukaryotic samples

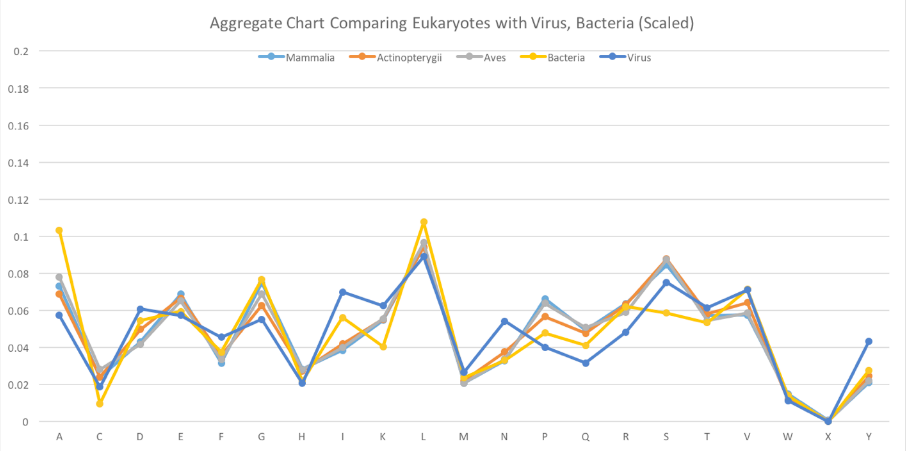

Fig. 2- Parallel plot of the Eukaryotic samples against Viral and Bacterial samples

3) *k-Mean Clustering:*

This result indicates that the samples cluster in three distinct partitions.

TABLE III
STATISTICS FROM K-MEAN CLUSTERING

Sample	Clustered	Total	Ratio	Interpretations
Eukaryotic Proteome	59	59	100%	All the sampled Eukaryotic Proteomes clustered in same partition. (Rank 1)
Bacterial Proteome	15	17	88.2%	Most of the samples clustered in same partition. (Rank 2)
Viral Proteome	14	18	77.7%	Quite high number of samples clustered. (Rank 3)

4) *Principle Component Analysis:*

This result indicates that the samples cluster in three distinct partitions.

TABLE IV
CLUSTER STATISTICS FROM PRINCIPLE COMPONENT ANALYSIS

Sample	Principal Component	Interpretation
Eukaryotic Proteome	[0.48020417, 0.29809942]	Three distinct Principal Components imply three distinct classifications of the vectors.
Bacterial Proteome	[0.80249911, 0.08815054]	
Viral Proteome	[0.43344298, 0.15875654]	

5) *Conclusion for Hypothesis 1:*

The results obtained through the said analysis all agrees to our initial hypothesis that the Evolutionary taxonomy, which is based on phylogenetic relationship, can be revalidated through the proteomic sequences. This comes from the result that the proteomes of different classes of organisms and viruses clustered together, leading to a clear distinction between different phylogenetic branches, agreeing to the evolutionary taxonomy.

B. *(Hypothesis 2) Prediction of Virus targets through Proteomic Sequences*

1) *Plots:*

The plots obtained in this case can be reffered in Figure 3 and Figure 4.

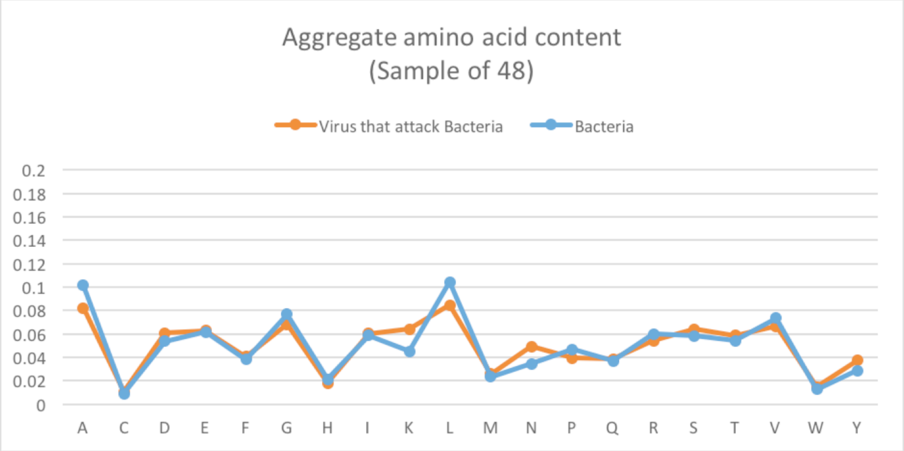

FIG. 3- AGGREGATED VECTORS OF THE BACTERIA AND THEIR TARGET VIRUS.

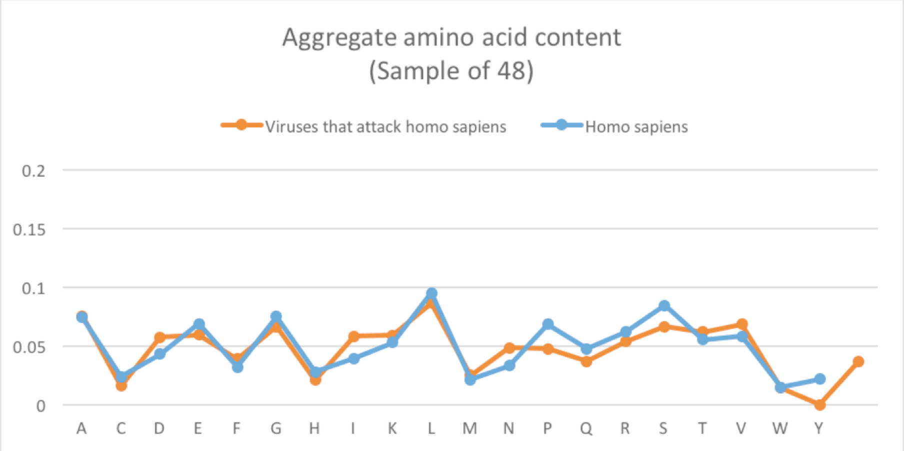

FIG. 4 AGGREGATED VECTORS OF THE TARGET VIRUS AND A HUMAN PROTEOME SAMPLE.

2) k-Mean Clustering:

The partitions obtained from the k-Mean clustering resulted in following for the data:

TABLE V
STATISTICS FROM K-MEAN CLUSTERING (96 SAMPLES)

Sample	Clustered	Total	Ratio	Interpretations
Bacterial Proteome	33	48	68.7%	All bacteria do not get clustered in the same partition.
Viral Proteome	32	48	66.6%	No proper clustering of viruses.

The above proves that the data is too similar to be divided into two proper classes and hence our hypothesis can be assumed to be true to an accuracy of approximately 32.3%. Hence, there is a probability that we can map a virus to its host by knowing only its proteome sequence up to an accuracy of 32.3%. Increasing the sample length of bacteria to 250 does establish the hypothesis and increases the accuracy to 42%. Further increasing it to 692, an accuracy of 43% can be achieved.

The result proves our hypothesis wrong. This means that one cannot correctly map a virus to its host based only upon the proteome sequence of the two.

TABLE VI
STATISTICS FROM K-MEAN CLUSTERING (50 SAMPLES)

Sample	Clustered	Total	Ratio	Interpretations
Homo sapiens	1	1	100%	Expected result for a single point data.
Virus that attack Human	30	49	61.2%	28 out of 48 viruses get clustered and tagged into a partition.

Similar result can be observed here with only 38.78% accuracy in predicting the host of a virus.

TABLE VII
STATISTICS FROM K-MEAN CLUSTERING (96 SAMPLES)

Sample	Clustered	Total	Ratio	Interpretations
Virus that attack Bacteria	24	48	50%	No proper clustering.
Virus that attack Human	28	48	58.3%	28 out of 48 viruses get clustered and tagged into a partition.

The above result shows that no clear difference can be established between the viruses that have different hosts. Hence, proteome sequence cannot be the only measure to predict and tag viruses for their hosts.

C. (Hypothesis 3) Significance of small Proteomic Sequences

The average of the number of occurrences of the sequences in the proteome was found out to be 1.91 and its variance to be 12.85. This suggested that there could be some sequences that occurred multiple times. Following were the sequences with highest frequency of occurrences in descending order:

TABLE VIII
STATISTICS FROM K-MEAN CLUSTERING (96 SAMPLES)

1614	679	671	654	494
GGAGG	GGNGG	AGGAG	GAGGA	GGTGG

381	293	284	282	275
GGDGG	NGGAG	GNGGA	GNGGN	AAAA

Trying to find the presence of these sequences in known motifs of Mycobacterium Tuberculosis could not produce any concrete result. Though, looking at the above table we can conclude that the sequences are Glycine (G) rich. Referring to the literature [7], [8], [9] we can see that these Glycine rich proteins are in fact important as they regulate the cell outer structure (antigen) of this particular bacteria.

Moreover, by Table VII, relative frequency of Glycine in secondary structures infer that it is predominantly found in turns, suggesting, Glycine must be an important amino acid which modulates the structure of the proteins in *Mycobacterium Tuberculosis*.

V. REFERENCES

[1] Mayr, Ernst & Bock, W.J. (2002), "Classifications and other ordering systems" (PDF), J. Zool. Syst. Evol. Research 40 (4): 169–94, doi:10.1046/j.1439-0469.2002.00211.x

[2] Fiona Cunningham, *et.al.*
Ensembl 2015
Nucleic Acids Research 2015 43 Database issue:D662-D669
doi: 10.1093/nar/gku1010

[3] Inselberg, Alfred (1985). "The Plane with Parallel Coordinates". Visual Computer 1 (4): 69–91. doi:10.1007/BF01898350.

[4] Karl Pearson (June 20, 1895) "Notes on regression and inheritance in the case of two parents," Proceedings of the Royal Society of London, 58 : 240–242.

[5] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.

[6] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" (PDF). Philosophical Magazine 2 (11): 559–572. doi:10.1080/14786440109462720.

[7] M. C. Marcelino and L. M. Gierasch, “Roles of ß-turns in protein folding: From peptide models to protein engineering,” Biopolymers, vol. 89, no. 5, pp. 380–391, 2008.

[8] S. Banu, N. Honoré, B. Saint-Joanis, D. Philpott, M. C. Prévost, and S. T. Cole, “Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens?,” Mol. Microbiol., vol. 44, no. 1, pp. 9–19, 2002.

[9] Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. Section 3.6, The Amino Acid Sequence of a Protein Determines Its Three-Dimensional Structure. Available from: http://www.ncbi.nlm.nih.gov/books/NBK22342/.