Chapter V

Characterization of pathogenic genes through condensed matrix method, case study through bacterial zeta toxin

5.1: Introduction

The Zeta protein causes Gram-negative bacteria (Escherichia coli) to cease growing and form long cells with many chromosomes, clearly unable to divide. Similarly, when the toxin is introduced into yeast cells it changes their morphology and halts budding, and in large quantities it causes death. Human cancer cells also die from the toxic effect of the Zeta poison. The poison protein Zeta is extraordinarily large compared other toxins (287 vs. around 100 amino acids) and with the exception of its frequently encountered nucleotide binding motif, it does not show any similarities to any known proteins. The molecular mechanism by which the Zeta toxin operates has not yet been discovered (Zielenkiewcz & Dmowski 2009). Where alignment based and structural based phylogeny fails nucleotide triplet based method give light towards molecular phylogeny.

Availability of nucleotide sequences of zeta toxin motivated us to construct phylogram using their nucleotide sequence, which will complement the phylogram obtained by sequence similarity. In this work we have done the molecular phylogeny of zeta toxin using their nucleotide sequence and without making any sequence alignment. It is a based on a method developed by basak et al associating DNA sequences with a set of sequence invariant. In this work we have quantified the string, which favors the direct comparison of the sequences. A sequence invariant, as considered as a number independent of the labels A, C, G, T standing for adenine (A), cytosine (C), guanine (G), and thiamine (T). We have form the matrix associated with each sequence and calculated the leading Eigen value of the matrices to see the variation of leading Eigen values associated with the string and the relationship between the enzymes. We have also build a phylogram using the Eigen values of the characteristics matrices of zeta toxin.

Our results complement the observation with the earlier studies based on multiple sequence alignment and structural alignment. The uniqueness of this method is that it does not employ sequence alignment of complete nucleotide sequence of the corresponding gene.

5.2: Material and Methods

The nucleotide sequences of zeta toxin of some pathogenic and non pathogenic bacteria were obtained from www.img.jgi.doe.gov. In a DNA sequence of four letters, there are 64 possible triplets (subsequence of length3) that can occur, starting from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC, etc. A 4×4×4 cubic matrix with 64 entries that denote the frequencies of occurrence of all the 64 triplets in a DNA sequence are introduced. For the cubic matrix, three groups of 4×4×4 matrices, {M1, M2, M3, M4}, {M5, M6, M7, M8}, {M9, M10, M11, M12}, can be obtained, each group of which contain all entries of the cubic (see Table I). Usually the group of 4×4 matrices {M1, M2, M3, M4} as the representative of the cubic matrix. The four matrices contain not only the information about frequencies of occurrence of all triplets of a DNA sequence but also the information about the frequencies of occurrence of pairs and every letter in a DNA sequence. For example, the number of all TG-pair in a DNA sequence is equal to the row sum of the third row in M2 plus ∂, where $\partial = 0$ if the last two letters of the DNA sequence are not TG and $\partial = 1$ otherwise. The frequency of occurrence of any pair in a DNA sequence can obtain by the above method. In addition, the frequencies of occurrence of four letters A, T, G, C are, respectively, equal to the sum of all entries of M1, M2, M3, M4 plus ∂, where ∂ are, respectively, equal to the number of A, T, G, C in the last two letters of the DNA sequence. The column sums of M1, M2, M3, M4 just denote the number of pairs of distance two in a DNA sequence (Randic, et al 2001; Randic & Basak 2001).

Table I: Three Groups of Four 4 × 4 matrices, {M_1, M_2, M_3, M_4},{M_5, M_6, M_7, M_8}, and {M_9, M_{10}, M_{11}, M_{12}}Listing All 64 possible XYZ Entries, Where X, Y, Z = A, C, G, T.

M_1				M_2				M_3				M_4			
AAA	AAT	AAG	AAC	TAA	TAT	TAG	TAC	GAA	GAT	GAG	GAC	CAA	CAT	CAG	CAC
ATA	ATT	ATG	ATC	TTA	TTT	TTG	TTC	GTA	GTT	GTG	GTC	CTA	CTT	CTG	CTC
AGA	AGT	AGG	AGC	TGA	TGT	TGG	TGC	GGA	GGT	GGG	GGC	CGA	CGT	CGG	CGC
AGA	ACT	ACG	ACC	TCA	TCT	TCG	TCC	GCA	GCT	GCG	GCC	CCA	CCT	CCG	CCC
M_5				M_6				M_7				M_8			
AAA	AAT	AAG	AAC	ATA	ATT	ATG	ATC	AGA	AGT	AGG	AGC	ACA	ACT	ACG	ACC
TAA	TAT	TAG	TAC	TTA	TTT	TTG	TTC	TGA	TGT	TGG	TGC	TCA	TCT	TCG	TCC
GAA	GAT	GAG	GAC	GTA	GTT	GTG	GTC	GGA	GGT	GGG	GGC	GCA	GCT	GCG	GCC
CAA	CAT	CAG	CAC	CTA	CTT	CTG	CTC	CGA	CGT	CGG	CGC	CCA	CCT	CCG	CCC

M_9				M_{10}				M_{11}				M_{12}			
AAA	TAA	AAG	CAA	AAT	TAT	GAT	CAG	AAG	TAG	GAG	CAG	AAC	TAC	GAC	CAC
ATA	TTA	GTA	TAC	ATT	TTT	GTT	CTG	TAG	TTG	GTG	CTG	ATC	CTT	TCG	CTC
AGA	TGA	GGA	CGA	AGT	TGT	GGT	CGG	AGG	TGG	GGG	CGG	AGC	GGC	GCG	CGC
ACA	TCA	GCA	ACC	ACT	TCT	GCT	CCG	ACG	TCG	GCG	CCG	ACC	CCT	GCC	CCC

We developed our own program in C++ to count all the possible triplets of t-RNA synthetase and formed the matrices by using all the possible triplets. Also we have calculated the leading Eigen values of the matrices by using MATHLAB (Version 4) (Toh *et al.,* 1999) software. We have constructed a distance matrix of the synthetases by summing the square of the difference of eigen values. A phylogram of the synthetases are constructed by the cluster analysis of the similarity matrix using phylip (Felsenstein 1989).

5.3: Results

The lengths of the zeta toxin of some pathogenic and non-pathogenic bacteria are given in Table 2. It is clear that the enzymes differ considerably in length. Firstly, we took the nucleotide sequence of zeta toxin of pathogenic and non pathogenic bacteria listed in Table 3 and counted the frequencies of occurrence of all the 64 triplets then the group of 4 × 4 matrices {M1, M2, M3, M4} as the representative of the cubic matrix are constructed.

Table 2: The lengths of the zeta toxin of some pathogenic and non-pathogenic bacteria

Bacteria Name	Short name	Length	Nature
Pseudomonas fluorescens	PSE	759	Non pathogenic
Frankia sp. CcI3	FR1	1053	Non pathogenic
Frankia sp. CcI3	FR2	1353	Non pathogenic
Mesorhizobium sp. BNC1 plasmid 1	MES	1761	Non pathogenic
Alteromonas macleodii 'Deep ecotype	ALT	720	Non pathogenic
Streptococcus pneumoniae ATCC 700669	STR	759	Pathogenic
Neisseria cinerea ATCC 14685	NEI	720	Pathogenic
Enterococcus faecalis TX0104	ENT	441	Pathogenic
Oribacterium sinus F0268	ORI	783	Pathogenic
O.algarvensis Gamma1	OLA	306	Non pathogenic
Crenothrix polyspora	CRE	2154	Non pathogenic

The leading Eigen values of each matrix are evaluated. The leading Eigen values of each matrix of those bacteria are represented in Table 4. The distance matrices of the synthetases are constructed by summing up the square of the difference of eigen values. The distance matrix for bacteria mentioned in Table 1 is given in Table 4.

Table 3: The leading Eigen values of each matrix of bacteria are represented in Table II.

Name of Bacteria	Short Name	M1	M2	M3	M4
Pseudomonas fluorescens	PSE	63.9742	59.4986	66.1942	61.1813
Frankia sp. CcI3	FR1	40.9835	45.5497	83.7531	1.0101
Frankia sp. CcI3	FR2	51.1907	38.5775	82.8222	1.001
Mesorhizobium sp. BNC1 plasmid 1	MES	39.4057	47.476	87.3356	86.5084
Alteromonas macleodii 'Deep ecotype	ALT	80.975	64.4988	59.6351	51.9077
Streptococcus pneumoniae ATCC 700669	STR	101.2985	68.0832	59.42	49.6625
Neisseria cinerea ATCC 14685	NEI	92.048	69.8199	70.3941	41.2137
Enterococcus faecalis TX0104	ENT	93.5377	88.5706	54.7326	37.3998
Oribacterium sinus F0268	ORI	108.5584	71.2367	65.8962	40.403
Olavius algarvensis Gamma1	OLA	71.8841	53.1558	57.2465	69.9674
Crenothrix polyspora	CRE	72.5779	69.704	60.1925	54.6687

Table 4: The distance matrix of zeta toxins of some pathogenic and non-pathogenic bacteria mentioned in Table 1

⇩	PSE	FR1	FR2	MES	ALT	STR	NEI	ENT	ORI	OLA	CRE
PSE	0	4652.03	4499.27	1836.57	443.05	1645.37	1311.01	2416.11	2557.36	260.05	256.61
FR1	4652.03	0	153.67	7328.99	5130.63	7104.81	4991.42	6779.14	7096.86	6470.41	5015.98
FR2	4499.27	153.67	0	7549.96	4688.15	6296.98	4416.93	6406.48	6196.68	6051.22	4818.6
MES	1836.57	7328.99	7549.96	0	3982.31	6392.28	5609.09	8093.65	7932.02	2266.07	3344.99
ALT	443.05	5130.63	4688.15	3982.31	0	430.98	381.04	971.79	977.8	543.17	105.54
STR	1645.37	7104.81	6296.98	6392.28	430.98	0	280.4	652.31	190.33	1505.05	853.16
NEI	1311.01	4991.42	4416.93	5609.09	381.04	280.4	0	613.64	295.49	1683.91	664.21
ENT	2416.11	6779.14	6406.48	8093.65	971.79	652.31	613.64	0	659.73	2790.05	1123.29
ORI	2557.36	7096.86	6196.68	7932.02	977.8	190.33	295.49	659.73	0	2620.79	1532.99
OLA	260.05	6470.41	6051.22	2266.07	543.17	1505.05	1683.91	2790.05	2620.79	0	517.05
CRE	256.61	5015.98	4818.6	3344.99	105.54	853.16	664.21	1123.29	1532.99	517.05	0

Using the distance matrices phylograms are constructed, which are represented in Figures 1. Figure 2 represented the Phylogram of zeta toxins based on ClustalW (Thompson *et al.*, 1994).

Figure 1: Phylogram of zeta toxins based on condensed matrix method.

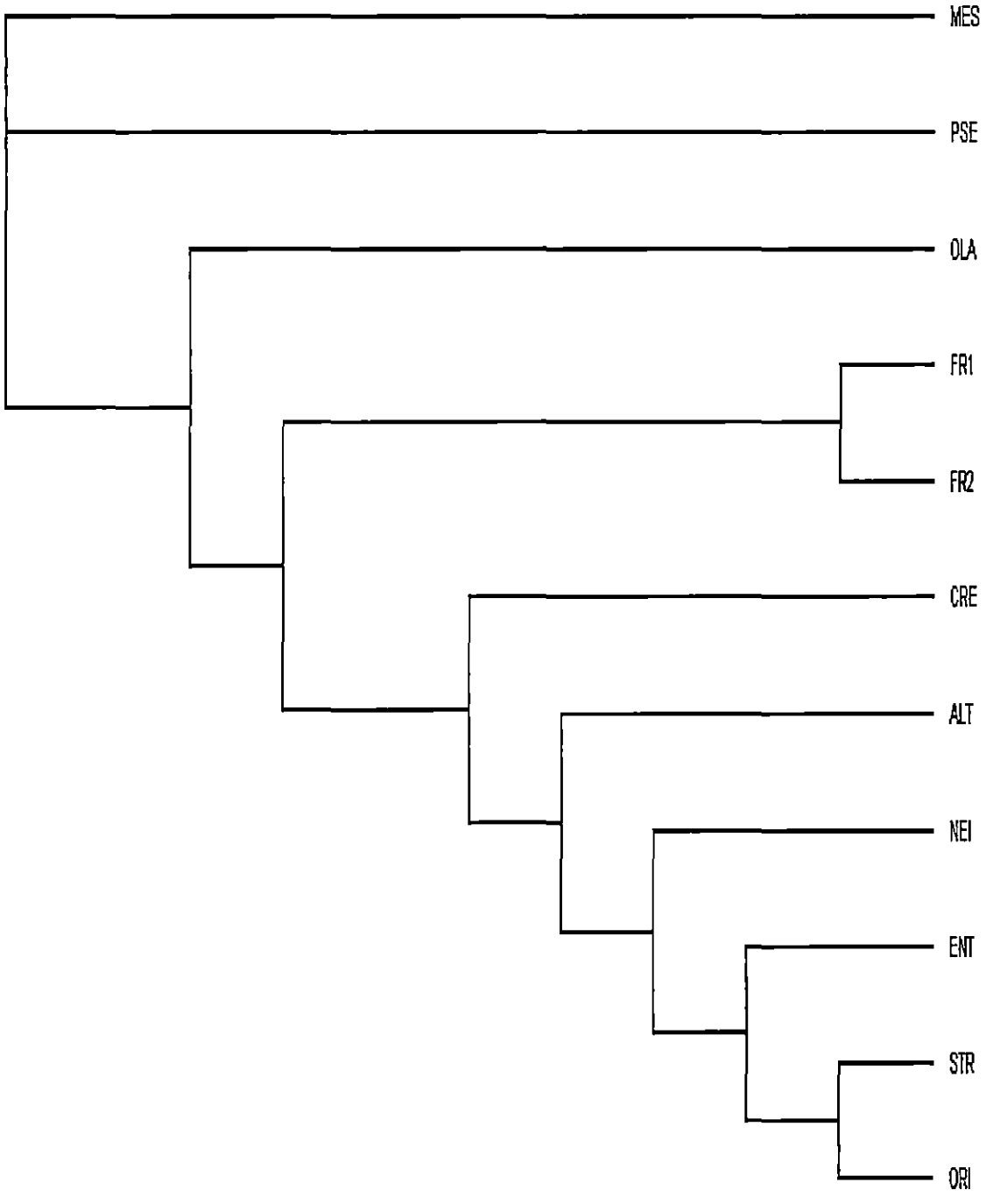
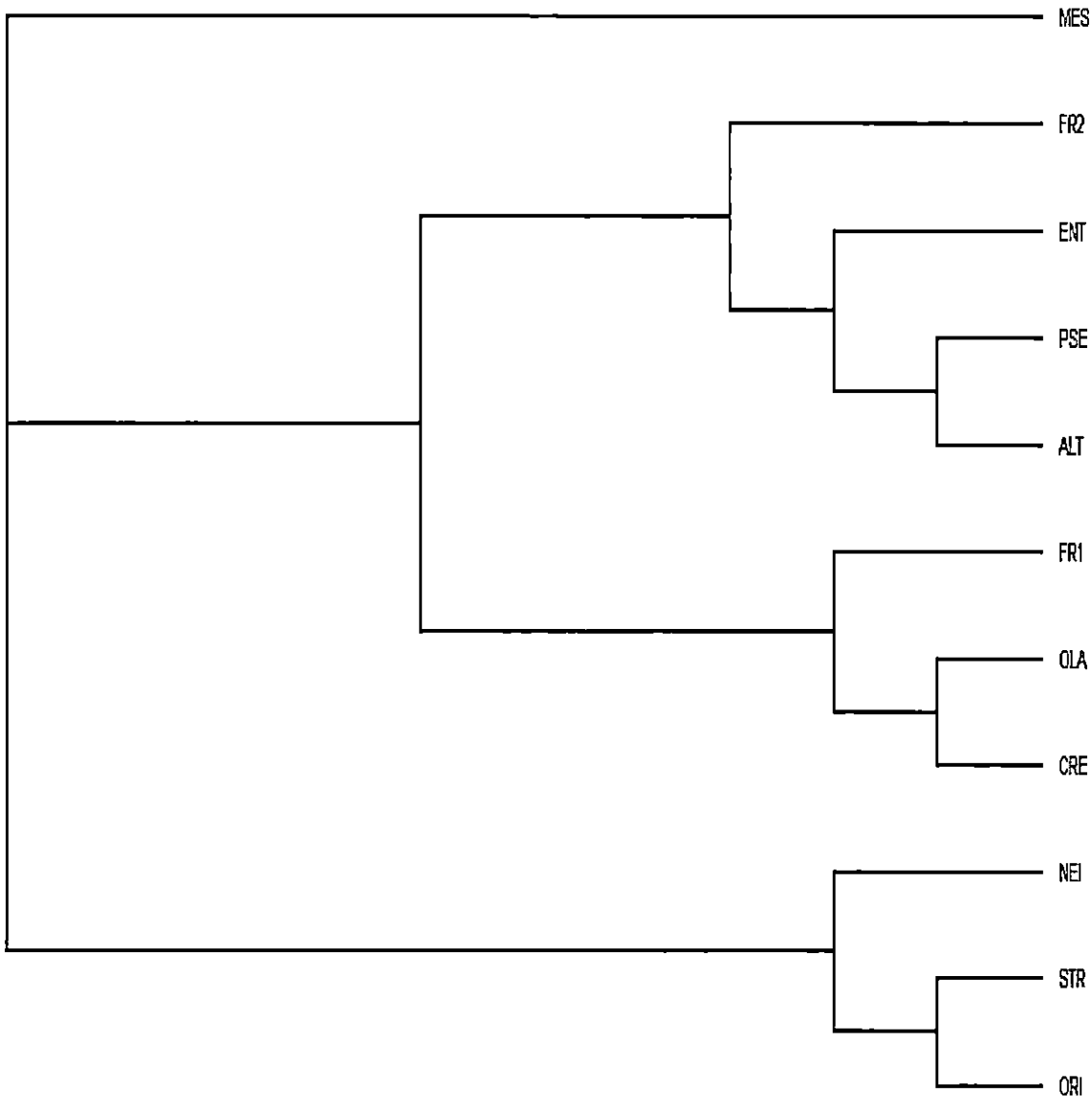


Figure 2: Phylogram of zeta toxins based on ClustalW.



5.4: Discussion

From Figure 1 is seen that, zeta toxin pathogenic and non-pathogenic bacteria forms different cluster. It is also seen that two zeta toxins of Frankia sp. Cc13 are in the same clade. From Figure 2 it is seen that separation are not clear between pathogenic and non-pathogenic bacteria which indicates superiority of condensed matrix method. It is also seen that two zeta toxins of Frankia sp. Cc13 are in the different clade.

Sequence comparison quickly becomes unreliable at this and lower levels of sequence identity. In this regime of similarity, it becomes difficult to distinguish between correctly aligned homologous sequences and unrelated sequences or random alignments. Structure based phylogeny has limited scope because adequate number of structures are not yet solved to draw any general conclusion. The nucleotide triplet based phylogeny is free from above mentioned limitations and it considers the full length of the genes for construction of phylogram. From the separation of zeta toxin by condensed matrix method may help to identify pathogenic or non-pathogenic strain/species.

5.3: References

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2), *Cladistics*, **5**:164-166.

Randic M & Basak SC (2001) Characterization of DNA Primary Sequences Based on the Average Distances between Bases, *J Chem Inf Comput Sci.*, **41**:561-568.

Randic M, Guo X & Basak SC (2001), On the characterization of DNA primary sequences by triplet of nucleic Acid Based, *J Chem Inf Comput Sci.*, **41**: 619-626.

Thompson JD, Higgins DG & Gibson TJ (1994), CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680

Toh KC, Todd MJ & Tutuncu RH (1999) SDPT3 - a Matlab software package for semidefinite programming, Optimization Methods and Software, **11**:545--581.

Zielenkiewcz U & Dmowski M (2009) Systems ensuring plasmid inheritance Epsilon vs. Zeta. *Research in Progress Microbiology*, **1**: 38-39.