In silico Annotation and Antigenic Peptide Identification of *Staphylococcus* Enterotoxins

[1]D.Gayathri, [2]M.Shiva Prakash
*Ideal Degree College For Women- Hyderabad, 2Scientist-E, Department of Microbiology & Immunology, National Institute Of Nutrition (ICMR) Hyderabad
*Email: chilly79in@yahoo.co.in, [2] drmspnin@gmail.com

Received: 11th April 2016, Accepted: 22nd April 2016, Published: May 2016

ABSTRACT
Staphylococcus has long back gained its importance in causing several diseases especially skin diseases. It is known to have the capability of enterotoxin production which constitutes for its pathogenicity. Being resistant to several antibiotics, causes hurdles in the treatment. The current work aimed to annotate in detail the selected Enterotoxins SEA, SEB, SEC, SED and SEE. The annotation includes the collection of gene sequences from NCBI Data base, Translating and obtaining the protein product of the sequences. Identification of functional domains in the protein sequence, structure prediction, Antigenic site identification and propensity calculations were performed. The 3D Structure of the protein was observed in Rasmol Software. The Foreignness of the sample to the human proteome was identified using TFAST Y of SDSC Biology workbench. The overall work emphasis on the analysis of enterotoxin pathogenicity of Staphylococcus aureus.

KEYWORDS
Enterotoxin, Antigenicity, Foreignness, TFASTY

INTRODUCTION

Staphylococcus aureus is gram positive cocci found in clusters. It is long being known for its pathogenicity against humans by causing several skin related aberrations. This group of bacteria are known to produce enterotoxins which are the actual cause for their pathogenicity. These are facultative anaerobic species, showing a positive test for Catalase and nitrate reduction. The enterotoxins produced by staphylococcus include SEA, SEB, SEC, SED and SEE. All the toxins are known to cause human skin infections. The use of antibiotics to treat staphylococcal infections would face a severe challenge due to the antibiotic resistance shown by these strains. Methicillin resistant staphylococcus is well known for its toughness towards several antibiotics used for its treatment. The enterotoxins produced by these bacteria are proteinacious in nature and have potential pathogenicity to provoke the immune components of the host thereby causing infections. Staphylococcal human skin infections especially by Community associated Methicillin resistant *staphylococcus aureus* (CA-MRSA) have been a global health problem within a short span of time.

Use of Bioinformatics in the analysis of biological data especially the medical data has been the current era research. Several insilico tools and softwares enable the complete annotation of the gene and its translated protein product. In the current work the physiological, Biochemical, functional and structural parameters of the 5 Enterotoxins A, B, C, D and E have been considered. Several tools like Protparam, TFASTY, EMBOSS, phyre etc have been used to perform the annotation. The study focusses on identifying the exact antigenic regions on these enterotoxin units so as to develop a structural analogue or a vaccine against the same to prevent the infections.

MATERIALS AND METHODS:

Sequence Retrieval from Databases

NCBI data base has been used to collect the gene sequences of all 5 enterotoxins A, B, C, D and E of Staphylococcus. The database not only provides the sequence but also the annotations like the length, regions, sites, domains functions etc. The sequences and basic annotation of all the five enterotoxins has been collected from the database. NCBI is a public database a product of US Govt. It has data from all round the globe obtained from research organizations and individual's submissions. It has a free access for all its contents.

Translation of the gene sequence to its protein product using TRANSEQ:

In order to study the protein product of these enterotoxins the translation has been performed insilico using TRANSEQ tool. The translated product can further be used for the protein annotation. Transeq tool is a product of EBI. The tool is designed to translate the given DNA sequences into their protein products using 6 frames. The product with maximum size can be used as the final protein product. The insilico translation can be of immense use to analyse the protein products of several important genes whose translated sequences are unavailable in the databases.

Identification of Antigenic Sites on the protein using EMBOSS

Emboss is an online tool used for the identification of antigenic sites on a given protein. The prediction is using the method of Kolaskar and Tongaonkar.

Bering based on a single parameter the method is ready to use and simple to understand. The major application of this technology is in the field of vaccine development. As per the available data and proved algorithms it can be understood that hydrophobic residues Cys, Leu and Valine when occur on the surface of the protein are more likely to be the part of antigenic sites. In the current study a comparative analysis has been made to determine the total number of antigenic sites present in the protein and further the comparison is also made on the propensity of antigenicity.

Antigenic Propensity Calculation using PVS Server:

Although antigenic site identification has been identified using Emboss antigenic, there is an importance in the identification of degree of antigenicity among the peptides which is termed as antigenic propensity. PVS Server is designed to identify the antigenic sites in the given protein along with their propensity calculation and graph providing the summery in the form of peaks. The highest peak is an indication of maximum antigenic propensity which makes the peptide more pathogenic or antigenic. All the 5 proteins were checked for their antigenic propensity and were compared for their pathogenic nature.

Measurement of foreignness using TFASTY

TFASTY is an online tool available at SDSC Biology workbench. The tool would compare the given user entered sequence with any one of the selected database in TFASTY. The expected result is the degree of identity among the user entered sequence and the database sequence. Higher the identity among the samples closer are the organisms. The technique is employed in the current study to detect weather the enterotoxin protein of bacteria shares any degree of similarity to the human proteome so as to reduce its antigenicity. The 5 toxin proteins are compared with the translated human CDS so as to identify the degree of similarity. In case the identity is less than 30 % the sequences are said to be foreign to the humans. Higher is the foreignness grater is the antigenicity.

Structure prediction using PHYRE:

Phyre is a structure prediction tool used for the designing or obtaining the 3D structure of the protein. The tool works by comparing the user entered sequenced with all the data base sequences of RCSB PDB whose structures are available. The result is the Pair wise alignment and identity prediction of all the database sequences with the entered query sequence. The sequence that share the similarity greater than a certain threshold are displayed in the output along with their pdb id's whose structure can directly be downloaded from the PDB Bank. The user should note that, a separate visualization software is required to view the 3D structure downloaded here as it contains the coordinates as a script not the 3D structure. The output contains Id's list along with the E value Score, Identity and other parameters to select the best as per the similarity.

RASMOL visualization:

Rasmol is a free online software used for the visualization of 3D structures of proteins. The software is basically a command line programme requiring the proper commands to apply on the structure. The software can only be used for structure visualization but not manipulation or alteration. The software can also recognise the polar and non-polar residues, loops, coils and sheets etc. It is a user friendly standalone software specifically used for protein structure visualizations.

Results and Discussion:

All the five translated sequences of enterotoxins were found to be with the lengths of SEA: 257 a.a, SEB: 266 a.a, SEC: 201 a.a, SED: 258 a.a, and SEE: 233 a.a

TRANSEQ for translation of genes to proteins:

The protein sequences of all the five proteins have been displayed below in their FASTA Format. The same sequences have been used for further analysis.

>SEA

M K K T A F T L LL F I A L T L T T S P L V N G
S E K S E E I N E K D L R K K S E L Q G A L
G N L K Q I Y YY N E K A K T E N K E S H D
Q F L Q H T I L F K G F F T D H S W Y N D L
L V D F D S K D I V D K Y K G K K V D L Y G
A Y Y G Y Q C A G G T P N K T A C M Y G G
V T L H D N N R L T E E K K V P I N L W L D
G K Q N T V P L E T V K T N K K N V T V Q E
L D L Q A R R Y L Q E K Y N L Y N S D V F D
G K V Q R G L I V F H T S T E P S V N Y D L
F G A Q G Q Y S N T LL R I Y R D N K T I N S
E N M H I D I Y L Y

>SEB

M Y K R L F I S H V I L I F A L I L V I S T P N V
L A E S Q P D P K P D E L H K S S K F T G
L M E N M K V L Y D D N H V S A I N V K S I
D Q F L Y F D L I Y S I K D T K L G N Y D N V
R V E F K N K D L A D K Y K D K Y V D V F G
A N Y YY Q C Y F S K K T N D I N S H Q T **D**
K R K T C M Y G G V T E H N G N Q L D K Y
R S I T V R V F E D G K N L L S F D V Q T N
K KK V T A Q E L D Y L T R H Y L V K N K K L
Y E F N N S P Y E T G Y I K F I E N E N S F W
Y D M M P A P G D K F D Q S K Y L M M Y N
D N K M V D S K D V K I E V Y L T T K KK

>SEC

MTPFITYITRAHVSLHAFSFTEMI
QIYVMIIFFIAFCFISPVMFYQLW
AFIAPGLHNNERQFIYKYSFFSV
LLFCAGVAFAFYVGFPIIIQFALK
LSLTLNISPVIGFKAYLVELIRWL
FTFGILFQLPILFIGLAKFGLIDIT
SLKHYRKYIYFACFVLASIIAPPD
LTLNILLTLPLILLFEFSMFIVKF
TCRGKPPTH

>SED

MAQHFKSKNVDVYPIRYSINCY
GGEIDRTACTYGGVTPHEGNKL
KERKKIPINLWINGVQKEVSLDK
VQTDKKNVTVQELDAQARRYLQ
KDLKLYNNDTLGGKIQRGKIEFD
SSDGSKVSYDLFDVKGDFPEKQ
LRIYSDNKTLSTEHLHIDIYLYEK

>SEE

MGNVMNLYTSPPVEGRGVINSR
QFLSHDLIFPIEYNEVKTELENTE
LANNYKDKKVDIFGVPYFYTCII
PKSEPDINQNFGGCCMYGGLTF
NSSENERDKLITVQVTIDNRQSL
GFTITTNKNMVTIQELDYKARH
WLTKEKKLYEFDGSAFESGYIK
FTEKNNTSFWFDLFPKKELVPF
VPYKFLNIYGDNKVVDSKSIKM

Antigenic Peptide prediction using EMBOSS:

Table 1 showing the total antigenic sites on the proteins under study

S.No	Name of Protein	Total Antigenic Sites
1	SEA	13
2	SEB	7
3	SEC	4
4	SED	8
5	SEE	6

The above table shows the total number of antigenic sites on each of the protein as predicted by EMBOSS antigenic. As per the above results the SEA has the highest no of antigenic sites on its surface, which may constitute for a high antigenicity compared to the other proteins.

Table 2: Showing the Antigenic Sites of Each Protein

SNO	Protein	Antigenic Peptides
1	SEA	AFTLLL**F**IALTLTTSPLVNG, QR**G**LIVFHT, GKKVDLYGAYYG**Y**QCAG, GNLKQ**I**YYYN, DQFLQ**H**TILFKG, TVPL**E**TV, KVPI**N**LW, YNDLLV**D**FD, NVTV**Q**ELDLQARRY, ACM**Y**GGVTLH, EPSVN**Y**DLFGAQG, NTLL**RI, K**DIVDKY
2	SEB	RLFISHVIL**I**FALILVISTPNVLAES,DKYVDVFGANYYY**Q**CYFS,SKDVKI**E**VYLTT, VTAQELDYLTR**H**YLVKNK,KVLYDDNHVSAINVKSIDQFLYF**D**LIYSI, LLS**F**DV, SITVR**V**F
3	SEC	FIYKYSFFSV**L**LFCAGVAFAFYVGFPIIIQFALKLSLTLNISPVIGFKAYLVELIRWL, TFGILFQLPILFIGLAKFGLIDITSLKHYRKYIYFA**C**FVLASIIAPPDLTLNILLTLPLILLFE FSMFIVKFTCRGK, IQIYVMIIFFIAFCF**IS**PVMFYQLWAFIAPG, FITYITRAHVS**L**HAFS
4	SED	TEHLHI**D**IYL,NVDV**Y**PIRYSINCYGG,QKEVSL**D**KV,GSKVS**Y**DLFDVKGD, NVTV**Q**ELD,QK**D**LKLY, EKQL**R**IY, RTAC**T**YG
5	SEE	KKELVPVPYKFLNI, KKVDIFGVPY**F**YTCIIPK, LI**T**VQVT, GGCCM**Y**G, DNKVVD**S**K, NSRQFLS**H**DLIFPIEYK

As per the above table 2 only the SEA has maximum number of antigenic sites.

PVS Server for the prediction of Antigenic Propensity of all the above peptides:

After identification of the total number of antigenic sites each of the protein PVS server has been used to identify the antigenic propensity of these epitopic regions. The PVS graph displays the total number of antigenic sites along with their degree of antigenicity / antigenic propensity in the form of a graph with peaks indicating highest antigenicity. The plots of all 5 proteins has been displayed below.

Fig 1: Antigenic propensity curve of Peptides SEA, SEB, SEC, SED and SEE.

Fig 1a: SEA

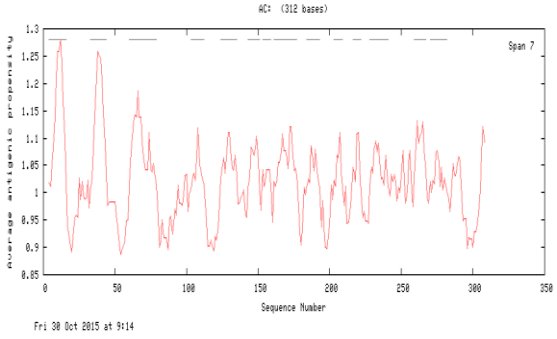


Table 1a: Corresponding Table for Antigenic propensity of SEA

There are 14 antigenic determinants in your sequence:

n	Start Position	Sequence	End Position
1	4	ASTAPHYLCCCAL	16
2	33	STAPHYLCCCSA	44
3	60	TAFTLLLFIALTLTTSPLVN	79
4	103	LGNLKQIYYY	112
5	124	HDQFLQHTILFK	135
6	143	WYNDLLVDF	151
7	153	SKDIVDK	159
8	161	KGKKVDLYGAYYGYQCA	177
9	184	TACMYGGVTL	193
10	203	KKVPINL	209
11	216	NTVPLET	222
12	228	KNVTVQELDLQARR	241
13	259	VQRGLIVFH	267
14	270	TEPSVNYDLFGAQ	282

From the above table and graph it can be inferred that the maximum antigenic propensity has been shown by the peptide 4-16 and 33-44, showing highest peaks.

Fig 1b: Plot of SEB

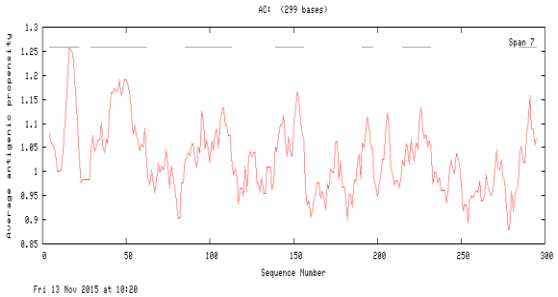


Table 1b for the antigenic propensities

n	Start Position	Sequence	End Position
1	4	AALAFSESTAPHYLCCCSA	22
2	29	ARESCLMYKRLFISHVILIFALILVISTPNVLAE	62
3	85	MKVLYDDNHVSAINVKSIDQFLYFDLIYS	113
4	139	KDKYVDVFGANYYYQCYF	156
5	191	RSITVRV	197
6	215	KVTAQELDYLTRHYLVKN	232
7	285	DSKDVKIEVYL	295

From the above table and plot it can be inferred that the maximum antigenic propensity was found in the peptide 29-62.

Fig 1c: Plot of SEC

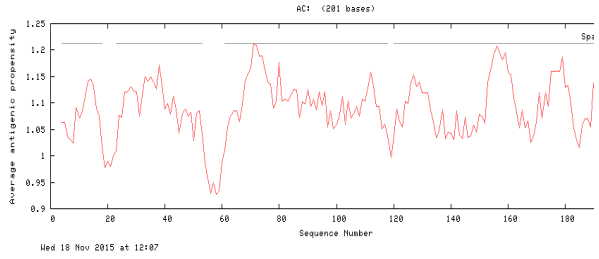


Table 1c for antigenic propensities

n	Start Position	Sequence	End Position
1	4	FITYITRAHVSLHAF	18
2	23	MIQIYVMIIFFIAFCFISPVMFYQLWAFIAP	53
3	61	QFIYKYSFFSVLLFCAGVAFAFYVGFPIIIQFALKLSLTLNISPVIGFKAYLVELIRW	118
4	120	FTFGILFQLPILFIGLAKFGLIDITSLKHYRKYIYFACFVLASIIAPPDLTLNILLTLPLILLFEFSMFIVKFTCRG	196

From the results it can be inferred that the peptide 61 to 118 has maximum antigenic propensity.

Fig 1d plot of SED

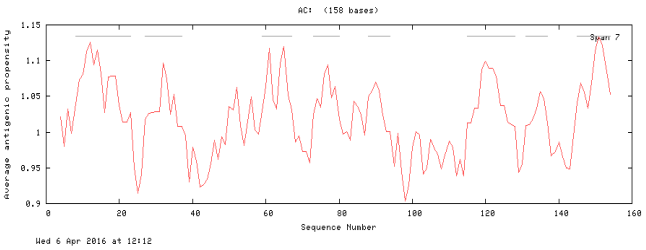


Table 1d for antigenic peptide

There are 8 antigenic determinants in your sequence:

n	Start Position	Sequence	End Position
1	8	KNVDVYPIRYSINCYG	23
2	27	DRTACTYGGVT	37
3	59	VQKEVSLDK	67
4	73	KNVTVQEL	80
5	88	LQKDLKL	94
6	115	DGSKVSYDLFDVKG	128
7	131	PEKQLRI	137
8	145	STEHLHIDIY	154

Contact Pedro Reche
Last Update: 6 April 2016

It can be inferred that the peptide 145 to 154 has the maximum antigenic propensity when compared to all the other peptides of the protein.

Fig 1e Plot of SEE

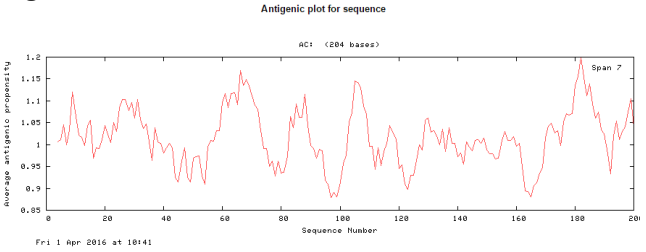


Table 1e for antigenic peptides

There are 6 antigenic determinants in your sequence:

n	Start Position	Sequence	End Position
1	19	INSRQFLSHDLIFPIEY	35
2	56	DKKVDIFGVPYFYTCIIP	73
3	83	FGGCCMY	89
4	103	KLITVQV	109
5	176	PKKELVPFVPYKFLN	190
6	193	GDNKVVDS	200

Contact Pedro Reche
Last Update: 1 April 2016

From the above results of SEE it can be inferred that the peptide 176 to 190 has the maximum antigenic propensity.

3D Structure Prediction of all the 5 peptides using Phyre server:

Phyre is an online tool used for the prediction of 3D structures of the given proteins using the BLAST Algorithm and RCSB PDB databank.The structure codes called pdb id's can be obtained in the result. However a visualization tool must be used for the final visualization of the structure.

As per the results of phyre the pdb id's obtained for the 5 peptides are 1SXT, 1D6E, 4B4A, not available and 1XXG (alternative). The structures of SED and SEE were unavailable in the databases. However as per the sequence similarity the structure of SEA and SEG can be used for SED and SEE respectively.

Visualization of 3D Structures in Rasmol:

Fig 2: 3D Structure Visualization of SEA or SED: 1SXT, RASMOL

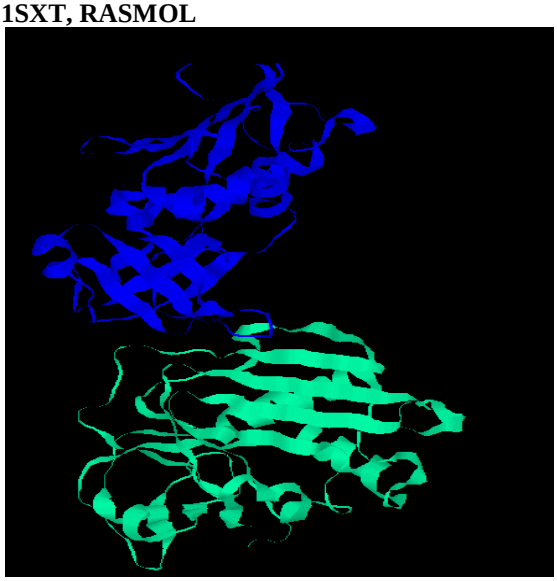


Fig 3: 3D Structure of SEB: 1D6E in Rasmol:

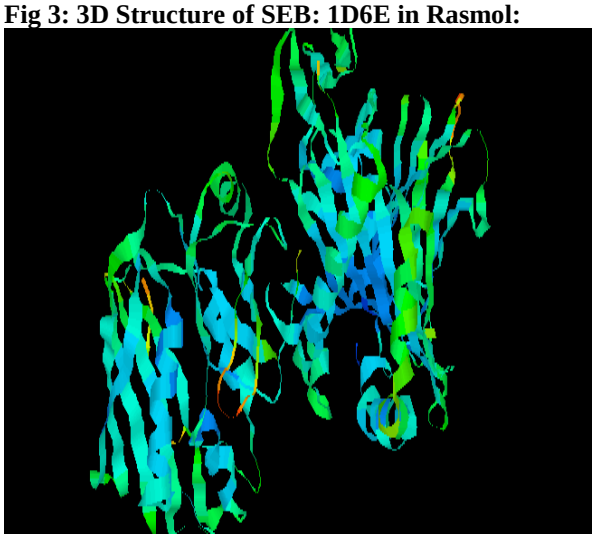


Fig 4: 3D Structure of SEC: 4B4A in Rasmol

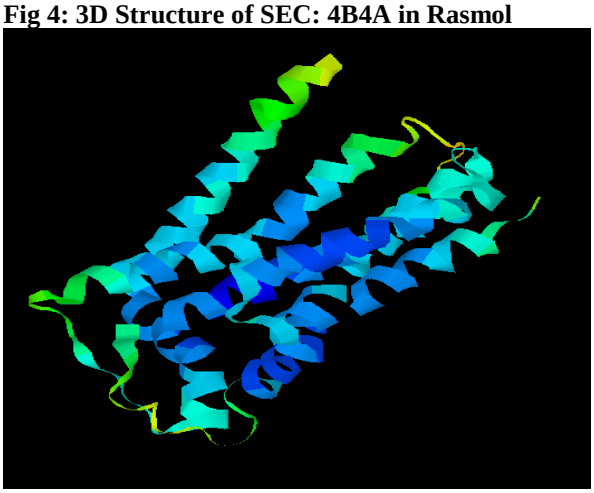
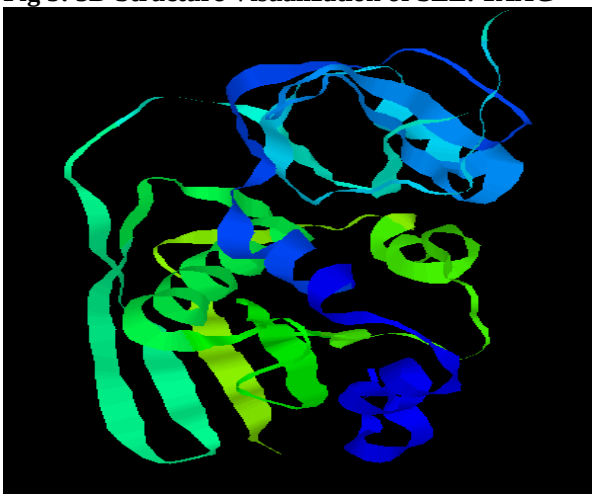


Fig 5: 3D Structure Visualization of SEE: 1XXG



From the above 3D structures it can be inferred that the structures of SEA, SEB and SED are similar and the Structures SEC and SEE differ widely.

TFASTY Results for the Confirmation of the Foreignness:

The tool has been used to identify the similarity of these five sequences with the human proteome so as to analyse their foreignness. The results are shown below:

Fig 6: TFASTY results for the protein SEA with Human CDS

The above figure shows that there is no sequence in human CDS that matches with the sequence of SEA stating No library with E_0<5.00

Fig 7: TFASTY results for the protein SEB with Human CDS

The above figure shows that there is no sequence in human CDS that matches with the sequence of SEB stating No library with E_0<5.00

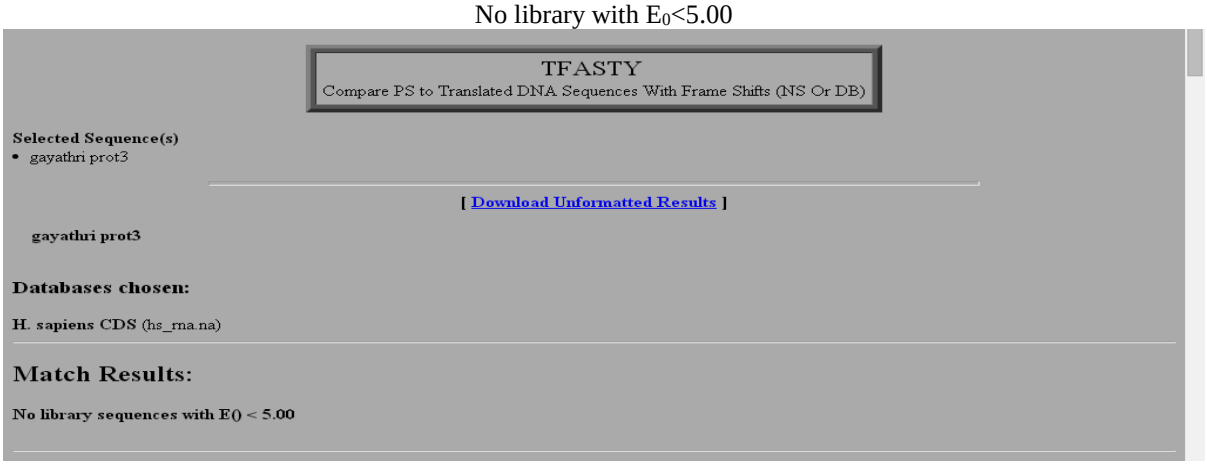
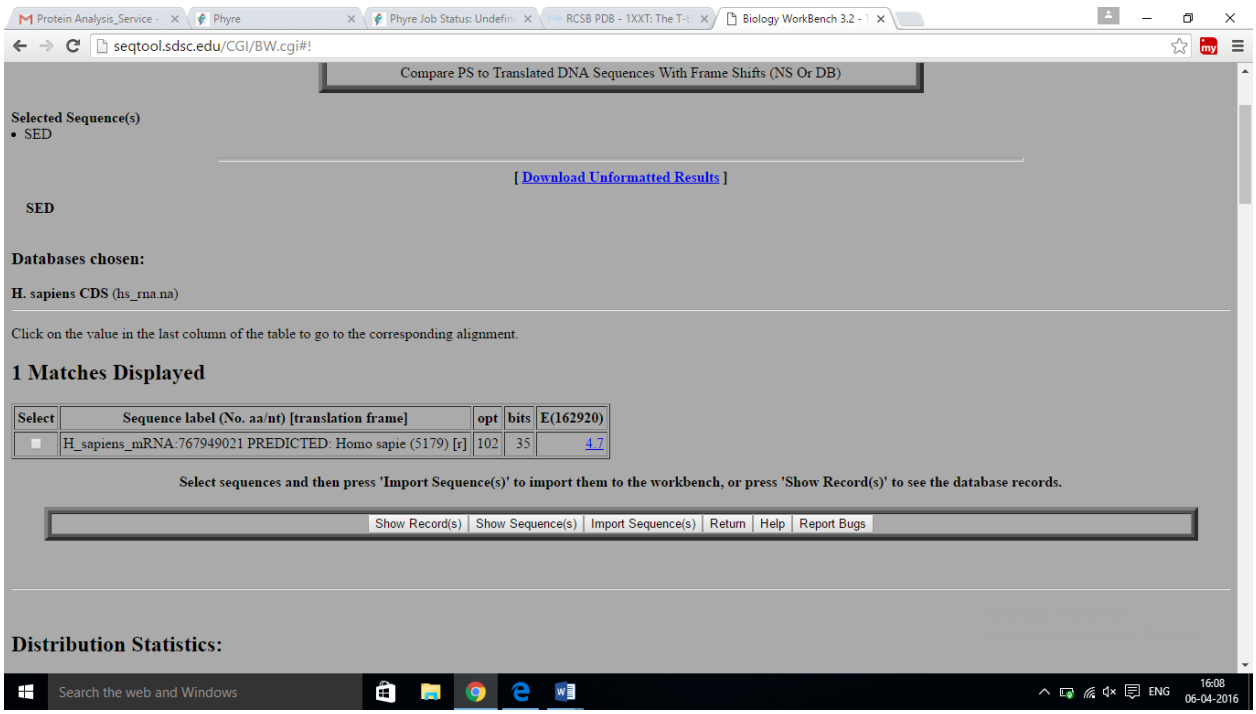


Fig 8: TFASTY results for the protein SEC with Human CDS

The above figure shows that there is no sequence in human CDS that matches with the sequence of SEC stating No library with E_0<5.00



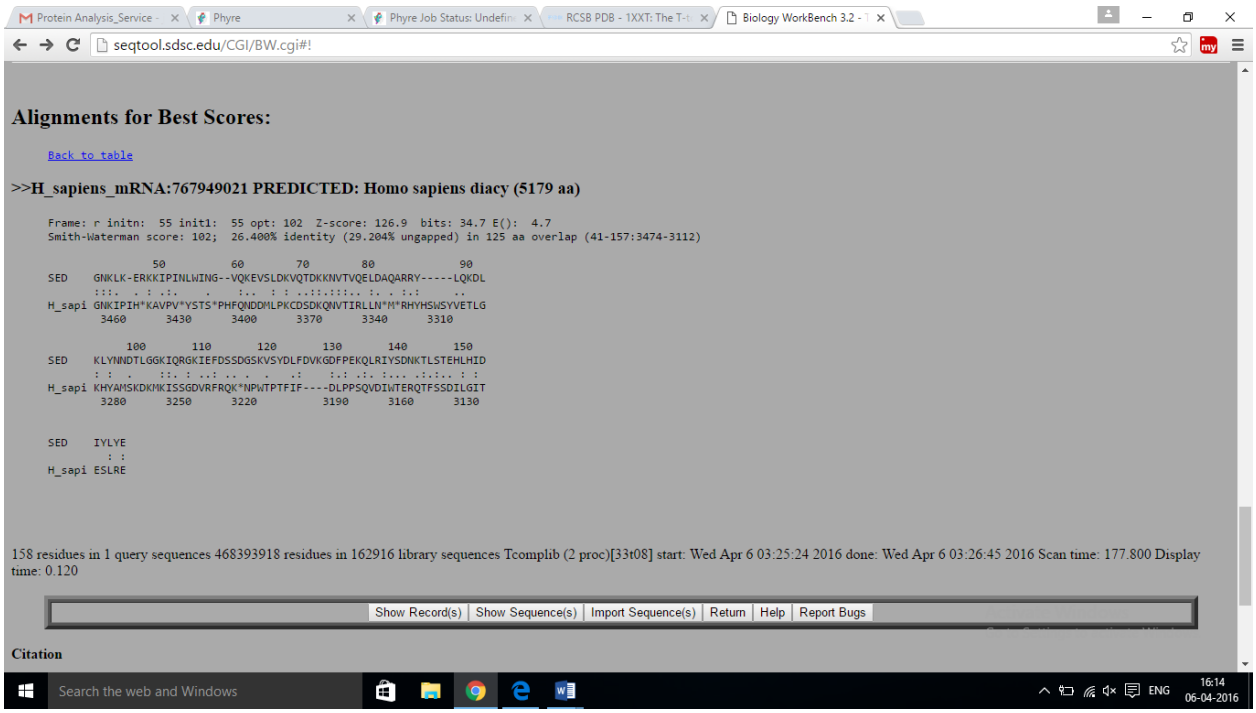


Fig 9: TFASTY results for the protein SED with Human CDS

The above result shows that one of the protein of Homo sapiens diacy shares similarity with the bacterial protein to an identity of 26%

Fig 10: TFASTY results for the protein SED with Human CDS

As per the above results there is no protein in the homo sapiens that shares similarity with the bacterial species with E_0<5.00

Similar results are obtained with TFASTY performed for SEB, C, D, and E which indicates that all the proteins are highly foreign to human proteome. The foreignness pertains to their pathogenicity. Fig: 6,7,8,9,10.

Conclusion

As per the above analysis one important conclusion though known theoretically is that the pathogenicity and foreignness of all the five peptides to the human proteome. Higher the degree of foreignness greater is the chances of causing diseases. It can also be concluded that out of all the 5 Peptides SEA has maximum number of pathogenic sites which imparts for its high pathogenicity. The work also emphasized on the identification of pathogenic sites called paratopic regions on the proteins that can bind to and recognized by the epitomic sites of antibodies. Structural analysis of the peptides proved that SEA, SEB and SED show greater structural similarity when compared to the SEC and SEE.

Bibliography:

1. LIU Xu-wei,GE Wen-xia(Xinjiang Coll.of Agric.Vocat'n.Technol.,Changji831100) " Enterotoxins of Staphylococcus aureus", Journal Of Microbiology 2008-05

2. Marie AlixPeyrat†et,al ., "egc, A Highly Prevalent Operon of Enterotoxin Gene, Forms a Putative Nursery of Superantigens in Staphylococcus aureus"

3. R. MoninaKlevens, DDS, MPHet, al., Invasive Methicillin Resistant Staphylococcus aureus Infections in the United States, JAMA, October 17- 2007

4. National Center for Biotechnology Information, U.S. National Library of Medicine8600 Rockville Pike, BethesdaMD, 20894USA

5. EMBOSS: The European Molecular Biology Open Software SuiteRice P., Longden I. and BleasbyA.Trends in Genetics. 2000 16(6):276-277 doi:10.1016/S0168-9525(00)02024-2

6. A new bioinformatics analysis tools framework at EMBL-EBI (2010)Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez RNucleic acids research 2010 Jul, 38 Suppl: W695-9

7. Kolaskar,AS and Tongaonkar,PC (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Letters 276: 172-174.

