In silico identification of potential therapeutic targets in *Clostridium botulinum* by the approach subtractive genomics

Koteswara Reddy G.[1]*, Nagamalleswara Rao K.[2], Phani Rama Krishna B.[3] and Aravind S.[4]
[*1]Department of Biotechnology, Bapatla Engineering College, Bapatla-522101, AP, India
kotireddy.nit@gmail.com, Ph: +91 9908594414.
[2]Department of Chemical Engineering, Bapatla Engineering College, Bapatla-522101, AP, India
[3&4] Department of Biotechnology, Bapatla Engineering College, Bapatla-522101, AP, India

Abstract- The completion of genome sequences of pathogenic bacteria and the completion of human genome project has provided lot amount of data that can be utilized to design vaccines and drug targets. One of the recently adopting strategies for drug designing is based on comparative genomics approach, in which the subtraction dataset between the host and the pathogen genome provides information for a set of genes that are likely to be essential to the pathogen but absent in the host. This approach has been used to identify vaccine and drug targets of *Pseudomonas aeruginosa* and *Helicobacter pylori*. We have used the same approach to identify the vaccine and drug targets of *Clostridium botulinum F strain*. Our analysis has revealed that out of 3631 coding sequences of the pathogen, 446 represent essential genes that have no human homolog. We have further analyzed these 446 genes by the protein sequence database to list some 96 genes whose products are possibly exposed on the pathogen surface. This preliminary work reported here identifies a small subset of the *Clostridium botulinum F strain* proteome that might be investigated further for identifying potential drug and vaccine targets in this pathogen.

Key Words - *Clostridium botulinum F strain*, Subtractive genomics approach, Comparative genomics approach, Drug and Vaccine targets.

Background

The completion of human genome project and the completion of genome sequences of pathogenic bacteria have increased the momentum of field of drug discovery against threatening human pathogens. The sequencing of pathogenic bacteria has provided a lot amount of raw material for *in silico* analysis [1]. Identification of bacterial genes that are non-homologous to human genes and important for the survival of bacteria is one of the promising means to identify novel drug targets. Availability of genome sequences of pathogens has provided a tremendous amount of information that can be useful in drug target and vaccine target identification [2]. The target should be essential for growth and viability of the organism, should provide selectivity, and should yield a drug which is highly selective against pathogen with respect to human host. Essential genes are those important for the survival of an organism, and therefore considered a foundation of life. A subtractive genomics approach and bioinformatics provide opportunities for finding the optimal drug targets [3]. A subtractive genomics has been successfully used by authors to locate novel drug targets in *Pseudomonas aeruginosa* [4]. The work has been effectively complemented with the compilation of the Database of Essential Genes (DEG) for a number of pathogenic micro-organism [1-5].The whole approach is built on the assumption that the target should play an important role in the survival of the pathogen and it should not have any conserved homolog in the human host. Non-human homologous can eradicate possibilities of cross contamination that might be harmful to the human host. The subtractive genomics approach is subtractive because we focus on the complement of the genome of the pathogen that is essential for the viability of the pathogen but is not present in the human. The present work makes use of the subtractive genomics approach, and the database of essential genes (DEG). The database of essential genes records currently available essential genes. For prokaryotes, the DEG database contains essential genes in more than 10 bacteria, such as *E. coli, B. subtilis, H. pylori, S. pneumoniae, M. genitalium* and *H. influenza* [1, 6-7] whereas for eukaryotes, the DEG database contains those in yeast, humans, mice, worms, fruit flies, zebra fish and the plant *A. thaliana*, by blasting query sequence with the prokaryotes sequence in the DEG database we can find out whether the query sequence is essential or not. By using these two it is possible to identify the potential therapeutic targets of *Clostridium botulinum* F strain. Sub cellular localization plays a key role to elucidate the functions of a protein. Therefore, proteins that cooperate towards a common biological function are located in the same sub cellular compartment. Eukaryotic cell has evolved highly elaborated subcellular compartments but prokaryotes (Gram-negative bacteria) too have 5 major subcellular localizations (outer membrane, inner membrane, periplasm, cytoplasm, and extracellular), specialized in distinct biochemical process [8, 9]. The prokaryotes are the causative agent of most of the deadly disease and widespread of epidemics, hence, biologists are paying much attention for the functional annotation of prokaryotic proteins. This may further guide the determination of virulence factors as well as new pattern of resistance for antibiotic agents in pathogenic bacteria. Hence, prediction of protein subcellular localization of gram-negative bacteria would be very useful in the field of

molecular biology, cell biology, pharmacology, and medical science. In a present study, systematic attempt has been made to develop the SVM (Support Vector Machines) based method for the prediction of subcellular localization of prokaryotic proteins Botulism (Latin, *botulus*, "sausage") also known as botulinus intoxication is a rare but serious paralytic illness caused by botulinum toxin, which is produced by the bacterium *Clostridium botulinum*. The toxin enters to the body in one of four ways: by colonization of the digestive tract by the bacterium in children (infant botulism) or adults (adult intestinal toxemia), by ingestion of toxin from foodstuffs (foodborne botulism) or by contamination of a wound by the bacterium (wound botulism).All forms lead to paralysis that typically starts with the muscles of the face and then spreads towards the limbs. In severe forms, it leads to paralysis of the breathing muscles and causes respiratory failure. In view of this life-threatening complication, all suspected cases of botulism are treated as medical emergencies, and public health officials are usually involved to prevent further cases from the same source [10]. The muscle weakness of botulism characteristically starts in the muscles supplied by the cranial nerves. This group of twelve nerves controls eye movements, the facial muscles and the muscles controlling chewing and swallowing. Double vision, drooping of both eyelids, loss of facial expression and swallowing problems may occur, as well as difficulty with talking. The weakness then spreads to the arms (starting in the shoulders and proceeding to the forearms) and legs (again from the thighs down to the feet).Severe botulism leads to reduced movement of the muscles of respiration, and hence problems with gas exchange. This may be experienced as dyspnea (difficulty breathing), but when severe can lead to respiratory failure: to the buildup of unexhaled carbon dioxide and its resultant depressant effect on the brain. This may lead to coma and eventually death if untreated. In addition to affecting the voluntary muscles, it can also cause disruptions in the autonomic nervous system. This is experienced as a dry mouth and throat (due to decreased production of saliva), postural hypotension (decreased blood pressure on standing, with resultant lightheadedness and risk of blackouts), and eventually constipation (due to decreased peristalsis). Some of the toxins (B and E) also precipitate nausea and vomiting [6]. *Clostridium botulinum* would normally be harmless to humans, but it can infect by a virus. The viral DNA, integrated into the bacterial genome, causes the host to produce toxins [11]. Neurotoxin production is the unifying feature of the species *C. botulinum*. Seven types of toxins have been identified and allocated a letter (A-G). *Clostridium botulinum* producing B and F toxin types have been isolated from human botulism cases in New Mexico and California [12]. The toxin type has been designated Bf as the type B toxin was found in excess to the type F Similarly, strains producing Ab and Af toxins have been reported[13].Organisms genetically identified as other *Clostridium* species have caused human botulism, *Clostridium butyricum* producing type E toxin and *Clostridium baratii* producing type F toxin [14,15]. The ability of *C. botulinum* to naturally transfer neurotoxin genes to other clostridia is concerning, especially in the food industry where preservation systems are designed to destroy or inhibit only *C. botulinum* but not other *Clostridium* species.

Materials and methods

[1] The complete genome and protein sequences of *Clostridium botulinum* F strain were downloaded from the National Center for the Biotechnology Information (NCBI) server *(ftp://ftp.ncbi.nlm.nih.gov/genomes/).*

[2] Coding sequences having less than 100 amino acids were screened out because coding sequences having less than 100 amino acids were less likely to represent essential genes from protein table *(www.ncbi.nlm.nih.gov/sites/entrez).*

[3] These coding sequences were subjected to BLASTX against the DEG database *(http://tubic.tju.edu.cn/deg/).* Expectation value (E-value) cut-off of 0.00001 was used to screen out coding sequences that are likely to be essential.

[4] Remaining coding sequences were subjected to BLASTX against human genome provided by the NCBI server *(http://www.ncbi.nlm.nih.gov/)* with default parameters to find out essential and non human homologs.

[5] The homologs to human genome were excluded; the essential, non-human homologs were listed out.

[6] Among the essential, non-human homolog coding sequences, their functional elements i.e. enzymes were listed out because they are potential drug targets.

[7] The protein products corresponding to the final selected genes were further analyzed with the database of protein sub cellular localization in bacteria *(http://db.psort.org/)* to compile the final list of proteins which were presumably located on the surface to design vaccine and drug targets.

[8] The hybrid the SVM (Support Vector Machines) module encapsulates the complete information of a protein such as amino acid composition, composition of physico-chemical properties, dipeptide composition, and PSI-BLAST output. The reliability index (RI) assignment is used to measure the level of certainty in the

prediction for a particular sequence. Hence, it is helpful to gain the confidence of the users about the prediction. The strategy used for assigning the RI is similar as used previously. The RI was assigned according to the difference between the highest and second highest SVM (Support Vector Machines) output scores. The reliability index for the hybrid approach based methods was calculated using following equation [16].

$$RI = \begin{cases} INT(\Delta * 5/3 + 1 & \text{if } 0 \le \Delta < 4, \\ 5 & \text{if } \Delta \ge 4. \end{cases}$$

Where, Δ is the difference between the highest and second highest SVM output scores

$$\text{Accuracy} = \frac{p(x)}{Exp(x)}$$

Where, *x* can be any subcellular location (cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular) *exp*(x) is the number of sequences observed in location *x*, *p*(x) is the number of correctly predicted sequences of location *x*.

[9] By using the graphpad prism software Version5.02, graph was plotted between Reliability Index (RI) value and Expected Accuracy (EA) on x and y axis respectively [17].

Results and discussions

The results obtained by this approach were summarized in Table-1. The objective of the work was to find and locate those essential genes of *C.botulinum F strain* that play important roles in the normal functioning of the bacterium within the host and to shortlist them in the view of drug targeting. Identification of non-human homologs in the essential genes of *C.botulinum F strain* with subsequent screening of the proteome to find the corresponding protein product are likely to lead to development of drugs that specifically interact with the pathogen. The non-human homologs of the surface proteins would represent ideal vaccine targets.

Our analysis has identified 1508 essential genes and by subjecting this essential genes to BLASTX against human genome provided by the NCBI server resulted in 446 essential, non human homolog genes. By further analyzing these essential and non-human homolog genes, we found 96 proteins that are possibly located on the membrane of the pathogen. They were found to represent either integral membrane proteins or outer membrane proteins that were linked to the membrane through some other molecule. Out of these 96 membrane proteins, 46 have no structures described in protein table *(www.ncbi.nlm.nih.gov/sites/entrez)*.In order to confirm the prediction reliability index (RI) assignment was carried out for the hybrid module. As depicted from the RI curve, good Expected Accuracies (EA) that is 90% and 98.1% was obtained with RI=4 and 5 respectively. It has also been observed that ~74% of the sequences have RI=5. Hence, the present method can annotate subcellular localization of prokaryotic proteins more reliably. We further analyzed those 46 non-structural membrane proteins and identified 10 proteins with RI value and EA as 5 and 98.1% respectively (Summarized in Table-2) from the graph (fig.1). Number of approaches for new vaccine and drug development exist, including sub-unit protein and DNA vaccines; recombinant vaccines; auxotrophic organisms to deliver genes and so on. Testing such candidates is tiresome and expensive. Bioinformatics enables us to reduce substantially the number of such candidates to test.

The computational genomics approach stated here is likely to speed up drug discovery process by removing hindrances like dead ends or toxicity that are encountered in classical approaches. The ninety six membrane associated proteins of *C.botulinum F strain* are invariably linked with essential metabolic and signal transduction pathways.

Conclusion

By subjecting the above 10 non-structural membrane proteins (table-2) to fold-level homology searches and structural modeling we can determine which of these proteins can function as the most effective surface epitope. Screening against such novel targets for functional inhibitors will result in discovery of novel therapeutic compounds active against bacteria, including the increased number of antibiotic resistant clinical strains.

Acknowledgements

We are thankful to Department of Biotechnology (DBT), Bapatla Engineering College and Bapatla Educational Society for their financial assistance. Authors are thankful to E.Hari Kishan Reddy, IIT-Hyderabad.

References

[1] Dutta A., Singh S.K., Ghosh P., Mukherjee R., Mitter S. and Bandyopadhyay D. (2006) *In Silico Biology* 6, 0005.

[2] Sarangi A.N., Aggarwal R., Rahman Q., Trivedi N. (2009) *B. J Comput Sci Syst Biol* 2: 255-258.

[3] Reddy E.H. and Satpathy G.R. (2009) *online journal of bioinformatics* 10(1): 14-28.

[4] Sakharkar K.R., Sakharkar M. K. and Chow V.T.K. (2004) *In Silico Biol*. 4, 0028.
[5] Zhang R., Ou H. Y. and Zhang C. T. (2004) *Nucleic Acids Res.* 32, D271-D272.
[6] Manoj Bhasin, Aarti Garg and G.P.S. (2005) *Bioinformatics Advance Access publication online* 21(10):2522-252.
[7] Galperin M. Y. and Koonin E. V. (1999) *Curr. Opin. Biotechnol.* 10, 571-578.
[8] Lu Z., Szafron D., Greiner R., Lu P. and Wishart D.S. (2004) *Bioinformatics,* 20: 547-556.
[9] Garg A. and Raghava G. P. S. (2008*) BMC Bioinformatics* 9:503.
[10] Tomb J.F., White O., Kerlavage A.R., Clayton R.A., Sutton G.G. and Fleischmann R.D. (1997) *Nature*; 388:539-47.
[11] Doyle Michael P. (2007) *Food Microbiology: Fundamentals and Frontiers*-ASM Press
[12] Hathaway C. L. and Mc Croskey L.M. (1987) *J. Clin. Microbiolozy* 25:2334–2338.
[13] Aureli P., Fenicia L., Pasolini B., Gianfrancesche M., Mccroskey J.M. and Hathaway C. L. (1986) *J. Infect. Dis.* 154:207–211.
[14] Hall J. D., McCroskey L.M., Pincomb B.J. and Hatheway C. L. (1985*) J. Clin. Microbiol.* 21:654–655.
[15] Notermans S. and Havellar A.H. (1980) *Antonie van Leeuwenhoek* 46:511–514.
[16] Saha S. and Raghava G. P. S. (2006) *Genomics Proteomics & Bioinformatics* 4:42-7.
[17] Graph pad prism software Version-5.02 a*vailable at http://www.graphpad.com/demos/*

Table 1- Classification of the Genes in Clostridium botulinum F strain

Total Number of coding Genes	**3631**
Genes greater than 100 amino acids (aa's)	3126
Essential genes greater than 100 aa's	1508
Essential and non human homologs genes greater than 100 aa's	446
Non membranous functional elements i.e. essential, Non-human homologs	350
Membrane associated non-human homologs of essential genes	96
Non structural ,Membrane associated non-human homologs of essential genes	46
Non-structural, Membrane associated non-human homologs of essential genes with RI value (5) and EA (98.1%)	10

Table 2- List of the 10 non- structural membrane proteins from the graph (fig.1) with RI value (5) and EA (98.1%) of Clostridium botulinum F strain.

S.NO	**Name of the protein**	*** RI**	*** EA**	**Gene ID**
1	CDP-diacylglycerol-serine O-phosphatidyltransferase	5	98.1	5405625
2	Undecaprenyl pyrophosphate phosphatase	5	98.1	5403669
3	Hypothetical protein CLI_0953	5	98.1	5403661
4	Formate/nitrite transporter family protein	5	98.1	5402527
5	Sodium:alanine symporter family protein	5	98.1	5404223
6	Sodium:alaninesymporter family protein	5	98.1	5405335
7	Na+/H+ antiporter family protein	5	98.1	5403165
8	MATE efflux family protein	5	98.1	5404439
9	ComEC/Rec2 family protein	5	98.1	5403983
10	Putative polysaccharide transporter	5	98.1	5403180

*Where RI indicates Reliability Index**,** EA indicates Expected Accuracy in percentage

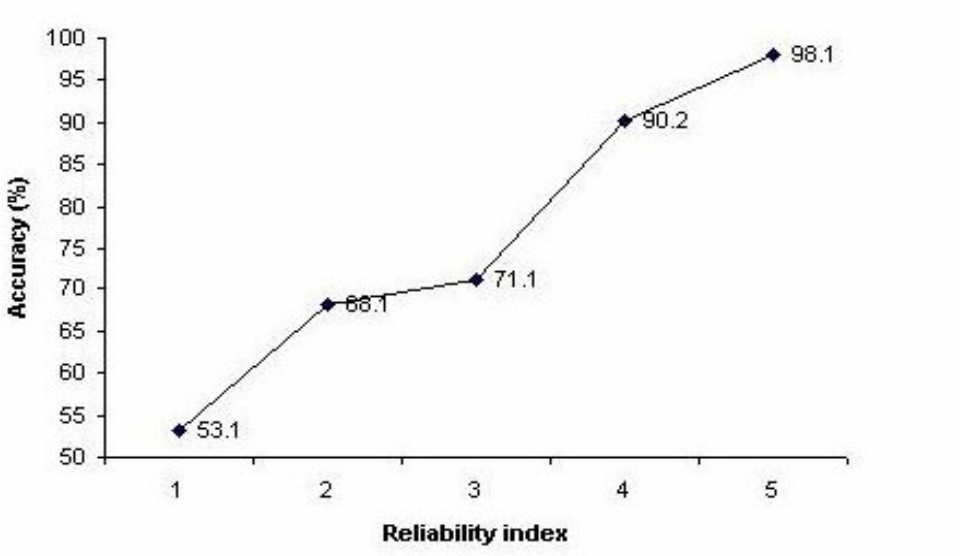

Fig. 1- The graph was plotted between Reliability index (RI) vs. Expected accuracy (EA) for 46 non-structural membrane proteins by Graph pad prism software Version5.02.