Research Article

IDENTIFICATION OF DRUG AND VACCINE TARGETS IN *CLOSTRIDIUM BOTULINUM-A* BY THE APPROACH *IN-SILICO* SUBTRACTIVE GENOMICS

G. Koteswara Reddy[1*], K. Nagamalleswara Rao[2] and Prof. P. Rajendra Prasad[3]

Address for Correspondence

[1*] Department of Biotechnology, Bapatla Engineering College, Bapatla-522 101, A.P, India, [2] Department of Chemical Engineering, Bapatla Engineering College, Bapatla-522 101,A.P,[3]Department of Chemical Engineering, AU College of Engineering, Andhra University,Visakhapatnam-530003,AP,India

E Mail: kotireddy.nit@gmail.com, kanidarapunag2001@gmail.com, prprasad_au@yahoo.com

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 A strain*. Our analysis has revealed that out of 3404 coding sequences of the pathogen, 180 represent essential genes that have no human homolog. We have further analyzed these 180 genes by the protein sequence database to list some 22 genes whose products are possibly exposed on the pathogen surface. This preliminary work reported here identifies a small subset of the *Clostridium botulinum* A strain proteome that might be investigated further for identifying potential drug and vaccine targets in this pathogen. The computational genomics approach stated here is likely to speed up the drug and vaccine discovery process.

KEYWORDS *Clostridium botulinum* A strain, Subtractive genomics approach, Comparative genomics approach, Drug and Vaccine targets.

INTRODUCTION

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 [1]. The sequencing of pathogenic bacteria has provided a lot amount of raw material for *in silico* analysis. 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. 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 [2]. A subtractive genomics has been successfully used by authors to locate novel drug targets in *Pseudomonas aeruginosa* [3]. The work has been effectively complemented with the compilation of the Database of Essential Genes (DEG) for a number of pathogenic micro-organism [4, 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 [6]. 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* [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* A 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. 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 [8].

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 [7].

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 [7]. *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 [9]. 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 [10]. 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[11].Organisms genetically identified as other *Clostridium* species have caused human botulism, *Clostridium butyricum* producing type E toxin and *Clostridium baratii* producing type F toxin [11,12,13]. 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 no other *Clostridium* species [14].

MATERIALS AND METHODS

1. The complete genome and protein sequences of *Clostridium botulinum* A 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. http://www.imtech.res.in/raghava/pslpred/submit.html.
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.

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

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* [8].

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 A 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 A 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 1458 essential genes from the database of essential genes and by subjecting these essential genes to BLASTX against human genome provided by the NCBI server resulted in 180 essential, non-human homolog genes are summarized in table-1. The protein products corresponding to the final selected genes were further analyzed with the database of protein sub cellular localization in bacteria to compile the final list of proteins which were presumably located on the surface to design vaccine and drug targets. By further analyzing these essential and non-human homolog genes, we found 22 proteins that are possibly located on the membrane of the pathogen and 158 proteins that are non-membranous functional elements (Among essential, non-human homologs) are summarized in table-2. The protein sub cellular localization studies described that the proteins represent either integral membrane proteins or outer membrane proteins that were linked to the membrane through some other molecule. The list of the possible inner and outer membrane proteins among 22 Membrane associated non-human homologs of essential genes are identified as 16 and 1, which are summarized in table-3 and table-4. Five cytoplasmic proteins are identified from protein sub cellular localization database, which are summarized in table-5. 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 used for 22 inner membrane proteins. In order to confirm the prediction reliability index (RI) assignment was carried out for the hybrid module with good Expected Accuracies (EA) parameters used for this study. It has also has been observed that 10 proteins with EA greater than 90% and RI greater than 4 respectively, which are summarized in table-6. Four possible inner membrane proteins with EA ≥ 90% & RI ≥4 has been observed that they don not have crystallographic structures from the protein table and corresponding their molecular function has also described in table-7. Hence, the present method can annotate subcellular localization of prokaryotic proteins more reliably. 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 the drug and vaccine discovery process by removing hindrances like dead ends or toxicity that are encountered in classical approaches. The 22 membrane associated proteins of *C.botulinum A strain* are invariably linked with essential metabolic and signal transduction pathways.

Table 1: Classification of the Genes in *Clostridium botulinum Strain A hall*

Total no of coding sequences	**3404**
Genes whose products are greater than 100 amino acids	2938
Essential genes (E-value cut-off 0.00001)	1458
Essential genes having non-human homolog	180
Non-membranous functional elements (Among essential, non-human homologs)	158
Membrane associated non-human homologs of essential genes	22
Inner membrane proteins among 22	16
Cytoplasmic proteins among 22	5
Outer membrane proteins among 22	1
Accuracy ≥ 90% & Reliability Index ≥4 proteins among 22	10
Inner membrane proteins having no crystallographic structures with Accuracy ≥ 90% & Reliability Index ≥4 proteins among 22	4

Table 2: List of the possible membrane proteins of *Clostridium botulinum* identified by subtraction

S.No	Gene ID	Name of the protein	*EA	*RI
1	5400235	F0F1 ATP synthase subunit epsilon	53.1	1
2	5399974	Iron chelate ABC transporter permease protein-	90.2	4
3	5398542	Methyl-accepting chemotaxis protein	53.1	1
4	5400096	Proton/sodium-glutamate symporter	90.2	4
5	5399642	PTS system, beta-glucoside-specific, IIABC component	71.1	3
6	5400238	Putative anion ABC transporter, permease protein	71.1	3
7	5398961	ABC transporter, permease protein	71.1	3
8	5400353	Molybdate ABC transporter, permease protein	90.2	4
9	5398484	ATP phosphoribosyltransferase catalytic subunit	71.1	3
10	5400600	Type II secretion system protein F	53.1	1
11	5401608	Tungstate ABC transporter, permease protein	90.2	4
12	5400228	Penicillin-binding protein	53.1	1
13	5400018	Acetyl-CoA carboxylase, carboxyl transferase, alpha subunit	90.2	4
14	5400548	Sodium:alanine symporter family protein	98.1	5
15	5400416	Putative oligopeptide ABC transporter, permease protein	71.1	3
16	5399034	Peptide/opine/nickel uptake ABC transporter permease	90.2	4
17	5398738	His/Glu/Gln/Arg/opine ABC transporter permease	90.2	4
18	5399723	Na+/H+ antiporter family protein	98.1	5
19	5400107	Major facilitator family transporter	90.2	4
20	5400449	Transglycosylase SLT domain-containing protein	53.1	1
21	5400414	Preprotein translocase subunit YajC	53.1	1
22	5400119	Sporulation integral membrane protein YtvI	71.1	3

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

Table 3: List of the possible inner membrane proteins among 22 Membrane associated non-human homologs of essential genes are 16.

S.No	Gene ID	Name of the protein
1	5399974	Iron chelate ABC transporter permease protein
2	5398542	Methyl-accepting chemotaxis protein
3	5400096	Proton/sodium-glutamate symporter
4	5399642	PTS system, beta-glucoside-specific, IIABC component
5	5400238	Putative anion ABC transporter, permease protein
6	5398961	ABC transporter, permease protein
7	5400353	Molybdate ABC transporter, permease protein
8	5401608	Tungstate ABC transporter, permease protein
9	5400228	Penicillin-binding protein
10	5400548	Sodium:alanine symporter family protein
11	5400416	Putative oligopeptide ABC transporter, permease protein
12	5399034	Peptide/opine/nickel uptake ABC transporter permease
13	5398738	His/Glu/Gln/Arg/opine ABC transporter permease
14	5399723	Na+/H+ antiporter family protein
15	5400107	Major facilitator family transporter
16	5400119	Sporulation integral membrane protein YtvI

Table 4: List of the possible Outer membrane proteins among 22 Membrane associated non-human homologs of essential genes are identified as One.

S.No	Gene ID	Name of the protein
1	5400449	Transglycosylase SLT domain-containing protein

Table 5: List of the possible Cytoplasmic proteins among 22 Membrane associated non-human homologs of essential genes are identified as Five.

SNo	Gene ID	Name of the protein
1	5400235	F0F1 ATP synthase subunit epsilon
2	5398484	ATP phosphoribosyltransferase catalytic subunit
3	5400600	Type II secretion system protein F
4	*5400018*	*Acetyl-CoA carboxylase, carboxyl transferase, alpha subunit*
5	5400414	Preprotein translocase subunit YajC

Table 6: List of the possible Inner membrane proteins with Accuracy ≥ 90% & Reliability Index ≥4, among 22 proteins are identified as Ten

S.No	Gene ID	Name of the protein	ACC	RI
1	5399974	Iron chelate ABC transporter permease protein-	90.2	4
2	5400096	Proton/sodium-glutamate symporter-	90.2	4
3	5400353	Molybdate ABC transporter, permease protein	90.2	4
4	5401608	Tungstate ABC transporter, permease protein	90.2	4
5	5400018	Acetyl-CoA carboxylase, carboxyl transferase, alpha subunit	90.2	4
6	5400548	Sodium:alanine symporter family protein	98.1	5
7	5399034	Peptide/opine/nickel uptake ABC transporter permease	90.2	4
8	5398738	His/Glu/Gln/Arg/opine ABC transporter permease	90.2	4
9	5399723	Na+/H+ antiporter family protein	98.1	5
10	5400107	Major facilitator family transporter	90.2	4

Table 7: List of the possible Inner membrane proteins having no crystallographic structures with Accuracy ≥ 90% & Reliability Index ≥4, among 22 proteins are identified four with molecular function.

S.No	Gene ID	Name of the protein	Molecular function
1	5400548	Sodium: alanine symporter family protein	Sodium-amino acid symporter activity
2	5399034	Peptide/opine/nickel uptake ABC transporter permease	Substrate transporter activity
3	5399723	Na+/H+ antiporter family protein	Sodium-Hydrogen antiporter activity
4	5400107	Major facilitator family transporter	Trans membrane transporter

CONCLUSION

By subjecting the above four non-structural membrane proteins (table-7) 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. The computational genomics approach stated here is likely to speed up the drug and vaccine discovery process.

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. Anirban Dutta, Shashi Kr. Singh, Payel Ghosh, Runni Mukherjee, Sayak Mitter and Debashis Bandyopadhyay “In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori” *In Silico Biology* (2006) 6,0005.
2. E.H. Harikishan Reddy and G.R. Satpathy “Identification of potential targets and lead molecules for designing inhibitory

drugs against *Chlamydophila pneumonia" Online journal of bioinformatics (*2009)10(1), 14-28.

3. K.R. Sakharkar, M. K. Sakharkar and V.T.K. Chow "A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa" *In Silico Biology* (2004) 4, 0028**.**
4. Ren Zhang, O.U. Hong-Yu and ChunTing Zhang "DEG, a Database of Essential Genes" *Nucleic Acids Research (*2004) 32, D271-D272.
5. Ren Zhang and Yan Lin "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes" *Nucleic Acids Research (*2009) 37, D455-D458.
6. G. Koteswara Reddy, K. Nagamalleswara Rao, B. Phani Rama Krishna and S. Aravind "In silico identification of potential therapeutic targets in Clostridium botulinum by the approach subtractive genomics" *International Journal of Bioinformatics Research* (2010) 2, 2:12-16.
7. J.F. Tomb, O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton and R.D. Fleischmann "Comparative Genomics and Understanding of Microbial Biology" *Nature (*1997) 388,539-47.
8. Manoj Bhasin, Aarti Garg and GPS "PSLpred: prediction of subcellular localization of bacterial proteins" *Bioinformatics Advance Access publication online* (2005) 21(10), 2522-2528.
9. Doyle and P. Michael "Food Microbiology: Fundamentals and Frontiers" *ASM Press (*2007).
10. C.L. Hathaway and L.M. McCroskey "Examination of feces and serum for diagnosis of infant botulism in 336 patients" *Journal of Clinical Microbiology* (1987) 25,2334–2338.
11. P. Aureli, L. Fenicia, B. Pasolini, M. Gianfrancesche, J.M. Mccroskey and C.L. Hathaway "Two cases of type E infant botulism caused by neurotoxigenic Clostridium botulinum in Italy" *J. Infect. Dis. (1986*), 154,207–211.
12. J. D. Hall, L.M. McCroskey, B.J. Pincomb and C.L. Hatheway "Isolation of an organism resembling Clostridium baratii which produces a type F botulinal toxin from an infant with botulism" *Journal of Clinical Microbiology (*1985) 21,654–655.
13. S. Notermans and A.H. Havellar "Removal and inactivation of botulinum toxin during production of drinking water from surface water" *Antonie van Leeuwenhoek* (1980) 46,511–514.
14. B.A. Satterfield, A.F. Stewart, C.S. Lew, D.O. Pickett, M.N. Cohen, E.A. Moore, P.F. Luedtke, K.L. Neill and R. Robison "A quadruplex real-time PCR assay for rapid detection and differentiation of the Clostridium botulinum toxin genes A, B, E and F" *J. Med. Microbiology(* 2010) 59, 55-64.