IJPSR (2012), Vol. 3, Issue 07 (Research Article)

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 27 February, 2012; received in revised form 27 June, 2012; accepted 29 June, 2012

IN-SILICO ANALYSIS AND HOMOLOGY MODELING OF TARGET PROTEINS FOR *CLOSTRIDIUM BOTULINUM*

Chirag Prajapati*[1] and Chintan Bhagat [2]

Department of Computer Science (Bioinformatics) [1], Department of Biotechnology [2], Veer Narmad South Gujarat University, Surat, Gujarat, India

Keywords:
Computational tools,
Isoelectric point,
Disulphide bridge,
Homology model,
Homology modeling,
Ramachandran plot

Correspondence to Author:

Chirag Prajapati

Assistant Professor, Department of Computer Science, Veer Narmad South Gujarat University, Udhna-Magdalla Road, Surat-395007, Gujarat, 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. Recently adopting strategies for drug designing is based on comparative genomics approach, it gives a set of genes that are likely to be essential to the pathogen but absent in the host. By performing homology searches and structural modeling we can determine which of these proteins can function as the most effective surface epitope. This provides novel targets for functional inhibitors will result in discovery of novel therapeutic compounds active against bacteria. In this study, we used proteins that are potential target for *Clostridium botulinum*, this include, MATE efflux family protein, ComEC/Rec2 family protein, formate/nitrite transporter family protein. Physico-chemical characterization interprets properties such as pI, EC, AI, GRAVY and instability index and provides data about these proteins and their properties. Prediction of patterns, disulfide bridges and secondary structure were performed for functional characterization. Three dimensional structures for these proteins were not available as yet at PDB. Therefore, homology models for these proteins were developed. The modeling of the three dimensional structure of these proteins were performed by swiss model & modeller. The models were validated using protein structure checking tools PROCHECK and WHAT IF. These structures will provide a good foundation for functional analysis of experimentally derived crystal structures and also for drug designing.

INTRODUCTION: Completion of human and pathogenic bacteria genome provided lot of raw material for in silico analysis [1]. Essential genes are those important for the survival of an organism and therefore considered foundation of life. 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 sequence 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 yield a drug which is highly selective against pathogen with respect to human host. A subtractive genome approach and bioinformatics provide opportunities for finding the optimal drug targets [3]. Proteins that cooperate towards a common biological function are located in sub cellular compartment.

Eukaryotic cell has evolved highly elaborated subcellular compartment but prokaryotes (gram negative bacteria) have five major subcellular localizations (outer membrane, inner membrane, periplasm, cytoplasm and extracellular), specialized in different biochemical process [4, 5].

Thus, identification of various subcellular proteins helps in development of drug candidate. By subjecting four non-structural membrane proteins (MATE efflux family protein, Com/Rec2 family protein, Formate/ Nitrite transport family protein, Hypothetical protein CLI_0953) to homology searches and structural modeling we can determine which of these proteins can function as most effective surface epitope. Screening against such novel targets results in designing of functional inhibitors [6, 7]. This will result in discovery of novel therapeutic compounds active against bacteria including the increased number of antibiotic resistant clinical strains.

Computational tools provide researches to understand physicochemical and structural properties of proteins. A large number of computation tools are available from different sources for making prediction regarding the identification and structure prediction of proteins. The major drawbacks of experimental methods that have been used to characterize the proteins of various organisms are time consuming, costly and fact that this methods not amendable to high throughput techniques.

In-silico approaches provide a viable solution to these problems. The amino acids sequence provides most of the information required for determining and characterizing molecule's function, physical and chemical properties.

Computationally based characterization of the features of proteins found or predicted in completely sequenced proteomes is an important task in search for knowledge of protein function. In this paper, the in silico analysis and homology modeling studies of target proteins were reported.

Three dimensional structures for these proteins were yet not available. Hence to describe it structural features and to understand molecular function, the model structures for these proteins were constructed.

MATERIALS AND METHODS: Sequences of target proteins of *clostridium botulinum* were retrieved from the NCBI, a public domain database. **Table 1** shows the protein sequences considered in this study. The protein sequences were retrieved in FASTA format and used for further analysis.

Physicochemical characterization: For physico-chemical characterization, theoretical isoelectric point (pI), molecular weight (M. wt), total number of positive and negative residues ($R^{+/-}$), extinction coefficient (EC) [8], instability index (II) [9], aliphatic index (AI) [10] and grand average hydropathy (GRAVY) [11] were computed using Expasy's Protparam server [12]. The results were shown in **Table 2**.

Functional characterization: The SOSUI server [13] performed the identification of transmembrane regions. **Table 3** represents the transmembrane region identified for these proteins. Disulphide bonds are important in determining functional linkages. **Table 4** shows prediction of "SS" bonds using primary structure (protein sequence data) by tool CYS_REC http://linux1.softberry.com/berry.phtml?topic=cys_rec &group=programs&subgroup=propt. CYS_REC identifies the position of cysteins, total number of cysteins present and pattern, if present, of pairs in the protein sequence [14].

Secondary Structure Prediction: SOPMA was employed for calculating the secondary structural features of the proteins sequences considered for this study. The results were represented in **Table 5**.

Model building and evaluation: The modeling of three dimensional structures of the protein was performed by two homology modeling programs, Swissmodel and Modeller [15, 16]. The constructed 3D models were energy minimized in CHIRON by short discrete molecular dynamics (DMD) simulation [17]. The overall stereochemical property of the protein was assessed by Ramchandran plot analysis [18].

The validation for structure models obtained from the two software tools was performed by using PROCHECK [19]. The models were further checked with WHAT IF [20]. The results of PROCHECK and WHAT IF analysis was shown in **Table 6 and 7** respectively. Structural analysis was performed and figures representations were generated with YASARA.

RESULT AND DISCUSSION: Table 1 shows target proteins considered in this study. These proteins sequence were retrieved from NCBI, a public domain database, in FASTA format and used for further analysis. Parameters computed using Expasy's ProtParam tool was represented in Table 2. The calculated isoelectric point (pI) will be useful because at pI, solubility is least and mobility in an electrofoucsing system is zero. Isoelectric point (pI) is the pH at which surface of protein is covered with charge but net charge of protein is zero. At pI proteins are compact and stable.

The computated pI values of all proteins were greater than 7 indicate that these proteins were considered as basic. The computated isoelectric point will be useful for developing buffer system for purification by isoelectric focusing method. Although Expasy's ProtParam computes the extinction coefficient (EC) for 276, 278, 279, 280 and 282 nm wavelengths, 280 nm is favored because proteins absorb light strongly there while other substances commonly in protein solutions do not. Extinction coefficient of proteins at 280nm is ranging from 25565 to 61715 M^{-1} cm^{-1} with respect to the concentration of Cys, Trp and Tyr. The high EC of RFP indicates presence of high concentration of Cys, Trp and Tyr.

The computed ECs help in the quantitative study of protein-protein and protein-ligand interactions in solution. The instability index provides an estimate of stability of protein in test tube. There are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. This method assigns a weight value of instability. Using these weight values, it is possible to compute an instability index (II). A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable. The instability index values for proteins were found to be ranging from 23.51 to 34.86. The result classified NTFP & MEFP were more stable than RFP and NTFP is more stable among all (Table 2).

The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic site chains (A, V, I and L) is regarded as positive factor for the increase of thermal stability of globular proteins. AI for proteins sequences ranged from 120.86 to 136.45. The very high AI of all protein sequences indicates that these proteins may be stable for wide temperature range. The Grand Average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all amino acids, divided by the number of residues in the sequence. GRAVY indices of proteins ranging from 0.356 to 0.871. The lower range of value indicates the possibility of better interaction with water.

Functional analysis of these proteins includes prediction of transmembrane region and disulfide bond. SOSUI distinguishes between membrane and soluble proteins from amino acids sequences. The transmembrane regions and their length were tabulated in Table 3. The server SOSUI classifies all proteins as membrane proteins. The transmembrane region is rich in hydrophobic amino acids. As disulphide bridges play important role in determining thermostability of these proteins. CYS_REC was used to determine the cysteine residues and disulphide bonds. Possible pairing and pattern with probability were shown in Table 4. Results show that RFP contains disulphide linkages.

The secondary structures of proteins were predicted by SOPMA (Self Optimized Prediction Method with Alignment) which correctly predicts 69.5% of amino acids for a state description of the secondary structure prediction (Geourjon and Deléage, 1995). The secondary structure indicates whether a given amino acid lies in a helix, strand or coil. Secondary structure features as predicted using SOPMA were represented in Table 5. The results revealed that alpha helix dominated among secondary structure elements followed by extended strand, random coil and beta turns for all sequences. The secondary structure were predicted by using default parameters (Window width: 17, similarity threshold: 8 and number of states: 4).

Three dimensional structures are predicted for proteins where such data is unavailable. There is lack of experimental structures for these proteins considered. Out of four protein sequences, three dimensional structure was not modeled for only HPC. The other three proteins for which the three dimensional structures were modeled includes MEFP, RFP, NTFP. The modeling of the three dimensional structure of the protein was performed by two

homology modeling programs, Swiss Model and Modeller. The constructed three dimensional models were energy minimized using CHIRON by short discrete molecular dynamics (DMD) simulations. The phi and psi distribution of the Ramachandran Map generated by of non glycine, non proline residues were summarized in Table 6. A comparison of the results obtained from Swiss Model and Modeller, shows that the models generated by two different software tools for NTFP were almost same. Thus NTFP would be used in designing drug. The final modeled structures were visualized by YASARA that was shown in **Figure 1**. The stereo chemical quality of the predicted models and accuracy of the protein model was evaluated after the refinement process using Ramachandran Map calculations computed with the PROCHECK program. The results of Ramachandran plot of all proteins were shown in **Figure 2**. In the Ramachandran plot analysis, the modeled structure for MEFP, RFP and NTFP has 65.0%, 60.0% and 83.4% residue respectively in most favored region. This shows that model for NTFP has good quality. The modeled structures of these proteins were also validated by other structure verification servers WHAT IF. The results were shown in Table 7.The analysis revealed RMS Z-score of the modeled protein NTFP is 0.840 which is near to 1.0.

TABLE 1: PROTEIN SEQUENCES CONSIDERED FOR THE STUDY

Accession number [NCBI]	Length [Amino Acid]	Description
YP_001392059	473	MATE efflux family protein [MEFP]
YP_001392246	595	ComEC/Rec2 family protein [RFP]
YP_001390312	254	formate/nitrite transporter family protein [NTFP]
YP_001390221	156	Hypothetical protein CLI_0953 [HPC]

TABLE 2: PARAMETERS COMPUTED USING EXPASY'S PROTPARAM TOOL

Proteins	Accession number	II	M.wt	pI	EC	-R	+R	AI	GRAVY
MEFP	YP_001392059	23.70	51809	9.54	43780	28	46	136.45	0.871
RFP	YP_001392246	34.86	68744	9.49	61715	43	77	120.86	0.356
NTFP	YP_001390312	23.51	27388	9.35	32680	12	20	121.34	0.731
HPC	YP_001390221	25.42	17345	9.73	25565	5	14	124.94	0.841

TABLE 3: TRANSMEMBRANE REGIONS IDENTIFIED BY SOSUI SERVER

Proteins	Accession number	No of transmembrane region	Length	Type of protein
MEFP	YP_001392059	12	23	Transmembrane
RFP	YP_001392246	9	23	Transmembrane
NTFP	YP_001390312	7	23	Transmembrane
HPC	YP_001390221	4	23	Transmembrane

TABLE 4: DISULPHIDE (SS) BOND PATTERN OF PAIRS PREDICTED, BY CYS_REC

Proteins	Accession number	CYS_REC
MEFP	YP_001392059	Cys present but no linkage
RFP	YP_001392246	Cys87-cys473, cys315-cys331
NTFP	YP_001390312	Cys present but no linkage
HPC	YP_001390221	-

TABLE 5: CALCULATED SECONDARY STRUCTURE ELEMENTS BY SOPMA.

Proteins	MEFP	RFP	NTFP	HPC
Alpha helix	57.08 %	32.77 %	58.27 %	53.85 %
310 helix	0.0 %	0.0 %	0.0 %	0.0 %
Pi helix	0.0 %	0.0 %	0.0 %	0.0 %
Beta bridge	0.0 %	0.0 %	0.0 %	0.0 %
Extended strand	19.24 %	33.28 %	19.69 %	19.87 %
Beta turn	4.44 %	7.06 %	3.54 %	7.05 %
Bend region	0.0 %	0.0 %	0.0 %	0.0 %
Random coil	19.24 %	26.89 %	18.50 %	19.23 %
Ambiguous state	0.0 %	0.0 %	0.0 %	0.0 %
Other state	0.0 %	0.0 %	0.0 %	0.0 %

TABLE 6: RAMACHANDRAN PLOT CALCULATION AND COMPARATIVE ANALYSIS OF THE MODELS FROM SWISS-MODEL AND MODELLER COMPUTED WITH THE PROCHECK PROGRAM

Server	Proteins	MEFP	RFP	NTFP
Swiss model	Residues in the most Favored Region	84.0 %	69.0 %	89.9 %
	Residues in additionally allowed region	13.4 %	20.7 %	7.2 %
	Residues in generously allowed region	1.1 %	3.4 %	1.4 %
	Residues in disallowed region	1.6 %	6.9 %	1.4 %
Modeller	Residues in the most Favored Region	65.0 %	60.0 %	83.4 %
	Residues in additionally allowed region	30.1 %	32.9 %	11.2 %
	Residues in generously allowed region	3.5 %	4.5 %	2.7 %
	Residues in disallowed region	1.4 %	2.5 %	2.7 %

TABLE 7: RMS Z-SCORE FOR BOND ANGLES OF MODELED PROTEIN STRUCTURE USING WHAT IF

Software used for model generation	Proteins	RMS Z score for bond angle
Swiss-model	MEFP	0.745
	RFP	0.732
	NTFP	0.773
Modeller	MEFP	0.746
	RFP	0.736
	NTFP	0.840

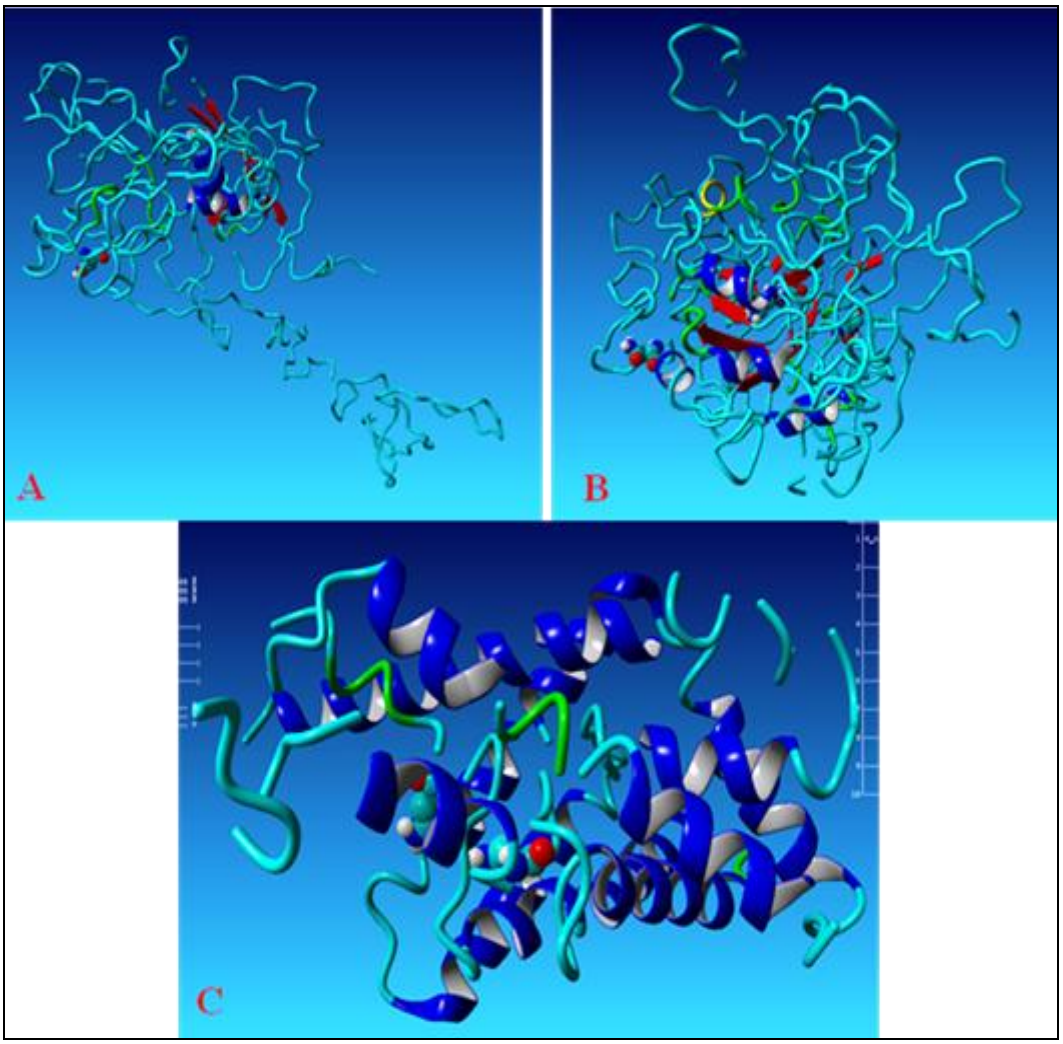

FIGURE 1: STRUCTURE OF PROTEINS MODELED BY MODELLER AND THE RIBBON STRUCTURE WAS VISUALIZED BY YASARA. FIG. A, B & C REPRESENTS PROTEINS MEFP, RFP & NTFP, RESPECTIVELY

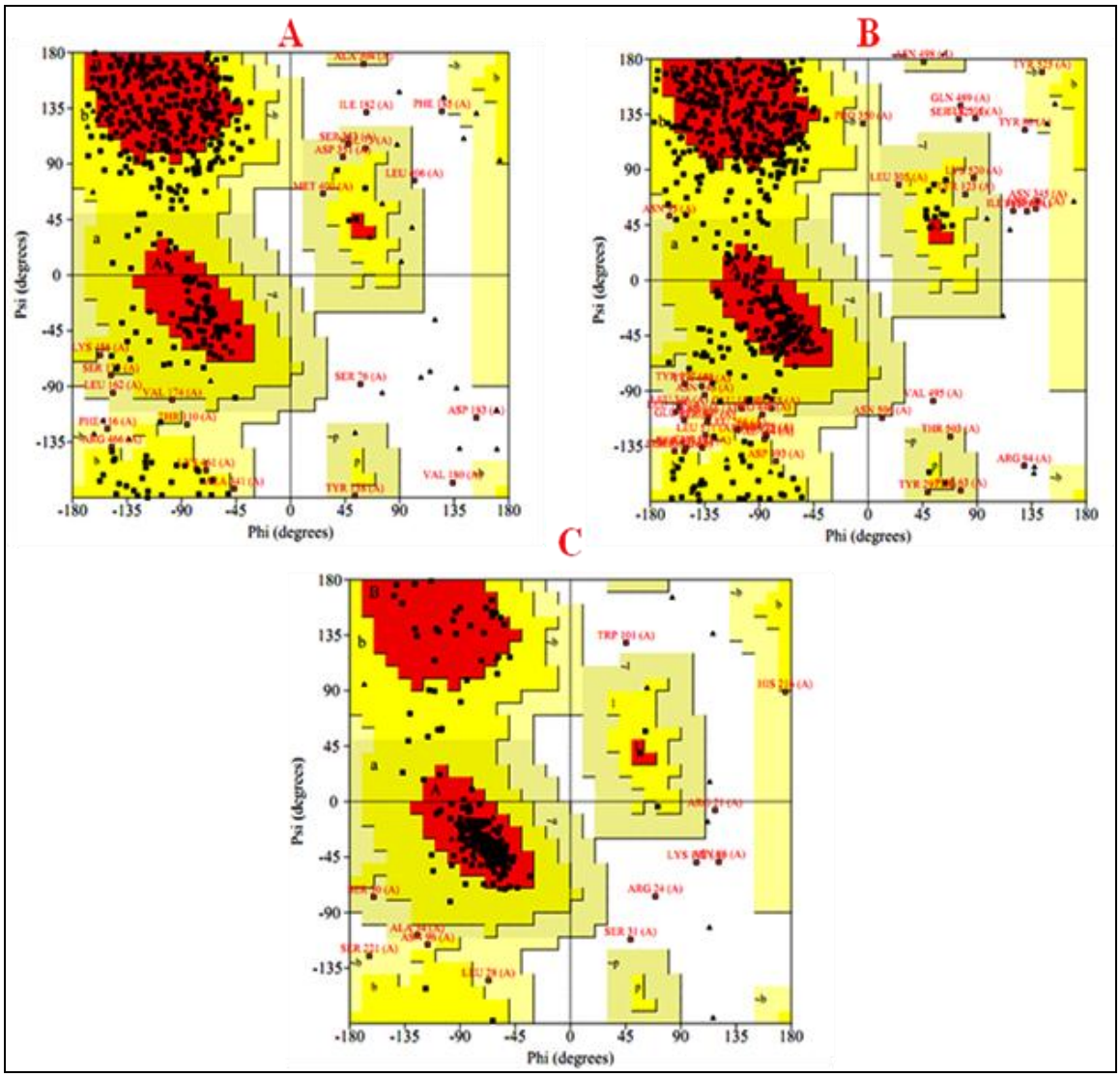

FIGURE 2: RAMACHANDRAN PLOT VALIDATION FOR PROTEINS MEFP, RFP & NTFP SHOWN IN FIGURE A, B & C RESPECTIVELY

CONCLUSION: In this study, four proteins MEFP, RFP, NTFP and HPC were selected but for HPC model were not generated. For remaining three proteins physicochemical analysis, theoretically isoelectric point, molecular weight, total number of positive and negative residues, extinction coefficient, instability index, aliphatic index and grand average hydropathy were performed. Results shows that NTFP is most stable protein and has considerably high aliphatic index.

Functional analysis was performed by SOSUI server. Secondary structure analysis shows that alpha helix dominated among secondary structure. Among three proteins, NTFP shows up to 83.4% of residues are in most favored region which indicated good model. The structures were checked using PROCHECK and WHAT IF.

This shows that the modeled structure of NTFP will provide a good foundation for functional analysis of experimentally derived crystal structures and also for drug designing against *Clostridium botulinum.*

ACKNOWLEDGEMENT: The authors are thankful to the staff members of Departments and authorities of Veer Narmad South Gujarat University, for providing support to the study.

REFERENCES:

1. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S and Bandyopadhyay D: *In silico* identification of potential therapeutic targets in the human pathogen Helicobacter pylori.in silico biology 2006, 6:0005.
2. Sarangi A.N, Aggarwal R, Rahman Q and Trivedi N: subtractive Genomics Approach for in Silico Identification and Characterization of Novel Drug Targets in Neisseria Meningitides Serogroup B B. J Comput Sci Syst Biol 2009; 2: 255-258.

3. Reddy E.H. and Satpathy G.R: Identification of potential targets and lead molecules for designing inhibitory drugs against Chlamydophila pneumoniae online journal of bioinformatics 2009; 1: 14-28.
4. Lu Z, Szafron D, Greiner R, Lu P and Wishart D.S : Predicting subcellular localization of proteins using machine-learned classifiers Bioinformatics 2004; 20: 547-556
5. Garg A. and Raghava G. P. S. BMC Bioinformatics 2008; 9:503.
6. Koteswara Reddy G and Nagamalleswara Rao K: International Journal of Bioinformatics Research 2010; 2:12-16.
7. Ren Zhang and Yan Lin: DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Research 2009; 37:455-458
8. Gill SC, Von Hippel PH Extinction coefficient. Anal Biochem 1989; 182: 319- 328.
9. Guruprasad K, Reddy BVP and Pandit MW: Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Prot Eng 1990; 4: 155-164.
10. Ikai AJ Thermo stability and aliphatic index of globular proteins. J Biochem 1980; 88: 1895 1898.
11. Kyte J and Doolottle RF: A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157: 105-132.
12. Gasteiger E Protein Identification and Analysis Tools on the ExPASy Server. In: John M Walker ed, The Proteomics Protocols Handbook, Humana Press 2005; 571-607.
13. Hirokawa T, Boon-Chieng S, Mitaku SB SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatic 1998; 14: 378-379.
14. Ferre F and Clote P: Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 2005; 21: 2336-2346.
15. Arnold K, Bordoli L, Kopp J, and Schwede T: The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics 2006; 22:195-201.
16. Sali A and Blundell TL: Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 1993; 234: 779-815
17. Chiron: Ramachandran, S., Kota, P., Ding, F. and Dokholyan, N. V., PROTEINS: Structure, Function and Bioinformatics 2011; 79: 261-270.
18. Ramachandran GN, Ramakrishnan C, Sasisekhran V : Stereochemistry of polypeptide chain confi guarations. J Mol Biol 1963; 7: 95-99.
19. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996; 8: 477-486.
20. Vriend G WHAT IF: A molecular modeling and drug design program. J Mol Graph 1990; 8: 52-56.

How to cite this article:
Prajapati C and Bhagat C: *In-Silico* Analysis and Homology Modeling of Targets Proteins for *Clostridium botulinum*. *Int J Pharm Sci Res*, 2012; Vol. 3(7): 2050-2056.