ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

An Automated Proteome based Pipeline for identifying Potential Drug Targets in Disease Causing Organisms.

M Udayakumar*, S Guru Vignesh, S Soorya Prakash, R Senthilkumar, and U Bharathram.

School of Chemical and Biotechnology, SASTRA University, Tanjore – 613401, Tamil Nadu, India.

ABSTRACT

In the present world of extensively expanding data, there is an existing need for the continual identification of the drug targets of these pathogens which has been endowed with an easier solution by the sequencing of the proteomes of various micro-organisms. Hence, an *in silico* approach that provides an automated prediction of potential drug targets for a wide range of pathogens, employing proteome based analysis, would be the most imperative requirement with respect to the drug discovery process. This *in silico* approach has been implemented in the server by a three step pipeline i)identifying the non-homologous proteins of the pathogens in comparison to the host,ii) verifying the essential proteins and iii)asserting that the proteins are involved with the pathogen specific pathways exclusive of the host pathways. The validation step was performed by subjecting the predicted drug targets to CELLO server. Three case studies (*Mycobacterium tuberculosis* – reviewed, *Vibrio cholerae* – unreviewed and reviewed) were used to execute the three step procedure and the predicted drug targets were found to be prevalent in cytoplasmic and membrane regions of the cell thereby confirming their effectiveness.This automated computational framework will play an vital role in Bioinformatics applications.

Keywords: Potential drug targets, Pathogen specific pathway proteins, Sub-cellular localization, DEG, CELLO,

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Novel medicines for the continually increasing diseases are obtained by the process of drug discovery and with respect to the drug discovery process, the identification of drug targets[1] becomes the all-important step. The process of drug target identification involves acquiring a molecular level understanding of a specific disease state and includes analysis of gene sequences, protein structures, protein interactions and metabolic pathways and its main objective lies in identifying suitable targets whose biological activity could be linked to a pathological pathway[2]. A major technique which is operating at a very high level in this area of identifying drug targets is the *in silico* approach[3]. A survey conducted in 1996 showed that there were 483 drug targets which accounted for the current therapies (45% receptors, 28% enzymes, 5% ion channels and 2% nuclear receptors)[4] and with the completion of the human genome, the identification of drug targets for various common diseases has grown much further. By the beginning of the millennium, it was predicted on the basis of bioinformatics analysis that successful target classes alone, such as receptors, enzymes and ion channels, could be predicted to amount to ~6500, which indicated the huge potential for target discovery and a decade later, we sit over close to thousands of potential drug targets identified by the *in silico* approach and a large volume of data that can be examined further in order to predict novel drug targets. The availability of a large amount of such data can be attributed to the completion of the genome sequencing projects of numerous pathogenic bacteria along with the successful completion of the human genome project[5] and comparative genomics[6], which employs a subtractive approach to the information available between the pathogen and the host, can be utilized to the aforementioned available data to provide information about the genes that have considerable probability to be essential to the pathogen and not to the host.

METHODOLOGY

Computational Pipeline for Drug Target identification

When the drug is administered, the general concern is that the protein targeted must be a protein which is possessed by the pathogen and not by the host. Therefore, a method of confirmation that the proteins of the pathogen are non-homologous in comparison to the host is mandatory[7]. From Figure 1, the workflow depicts both the manual and automated approach. The manual approach has been carried out in many of the research articles for the identification of potential drug targets and it is a time consuming process. To overcome these problems, we developed an automated computational pipeline that finds and evaluates the drug targets of the given organism. The initial task for the pipeline is setting up the *Homo sapiens* protein database and it was facilitated with the help of the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/). The data for the pathogens that we principally tested was fetched from the Uniprot[8] database (http://www.uniprot.org/). A protein BLAST[9] operation was performed with a user determined e-value between the two sets of data to separate out the non-homologous sequences. The e-value is an important criterion with respect to the protein BLAST operation since it describes the number of hits that occur by chance and is inversely proportional with relation to the score of the match.The second step in the prediction of potential drug targets would be to ensure that the identified non-homologous proteins are mandatory for the survival of the pathogen. Hence the secondary step involves keying out the essential proteins. A BLASTp operation is carried out with a user determined e-value and bit-score carried out between the non-homologous sequences and the essential protein sequences of the corresponding organism. The essential proteins encoding for the pathogen were obtained from DEG 10.8 (http://tubic.tju.edu.cn/deg/). The Database of Essential Genes[10] contains all the essential genes that are currently available which can be used for performing the BLAST operation against the query sequences and the presence of homologous genes indicates the possibility of the queried genes also being essential.The final step involves ascertaining the proteins that are exclusively involved with the pathogen specific pathways. Hence, dispensing with the proteins that correspond to the pathways of the host which may lead to adverse side effects is important. This step is performed by gathering the pathogen specific pathways of the corresponding organism and executing a protein BLAST operation with a user determined e-value and bit-score between the essential proteins obtained and the proteins accumulated as functioning in the pathogen specific pathways. The proteins that correspond to the pathogen specific pathways were collected from KOBAS 2.0[11] server (http://kobas.cbi.pku.edu.cn) which annotates an input set of genes with putative disease pathways by the process of mapping to the genes with known annotations.After the completion of these three steps, the resulting proteins were settled on as to be the potential drug targets for the pathogen in question. Drug targets are primarily fixated to be prevalent in either the cytoplasm or the membrane[12]. And hence, the determination of the sub-cellular localization of

ISSN: 0975-8585

the resulting proteins was considered as the validation step. CELLO server (http://cello.life.nctu.edu.tw/) was utilized for this purpose[13]. CELLO server is a multi-class SVM classification system that uses 4 different types of sequence coding schemes to finally assign the sub cellular localizations for the inputted proteins.

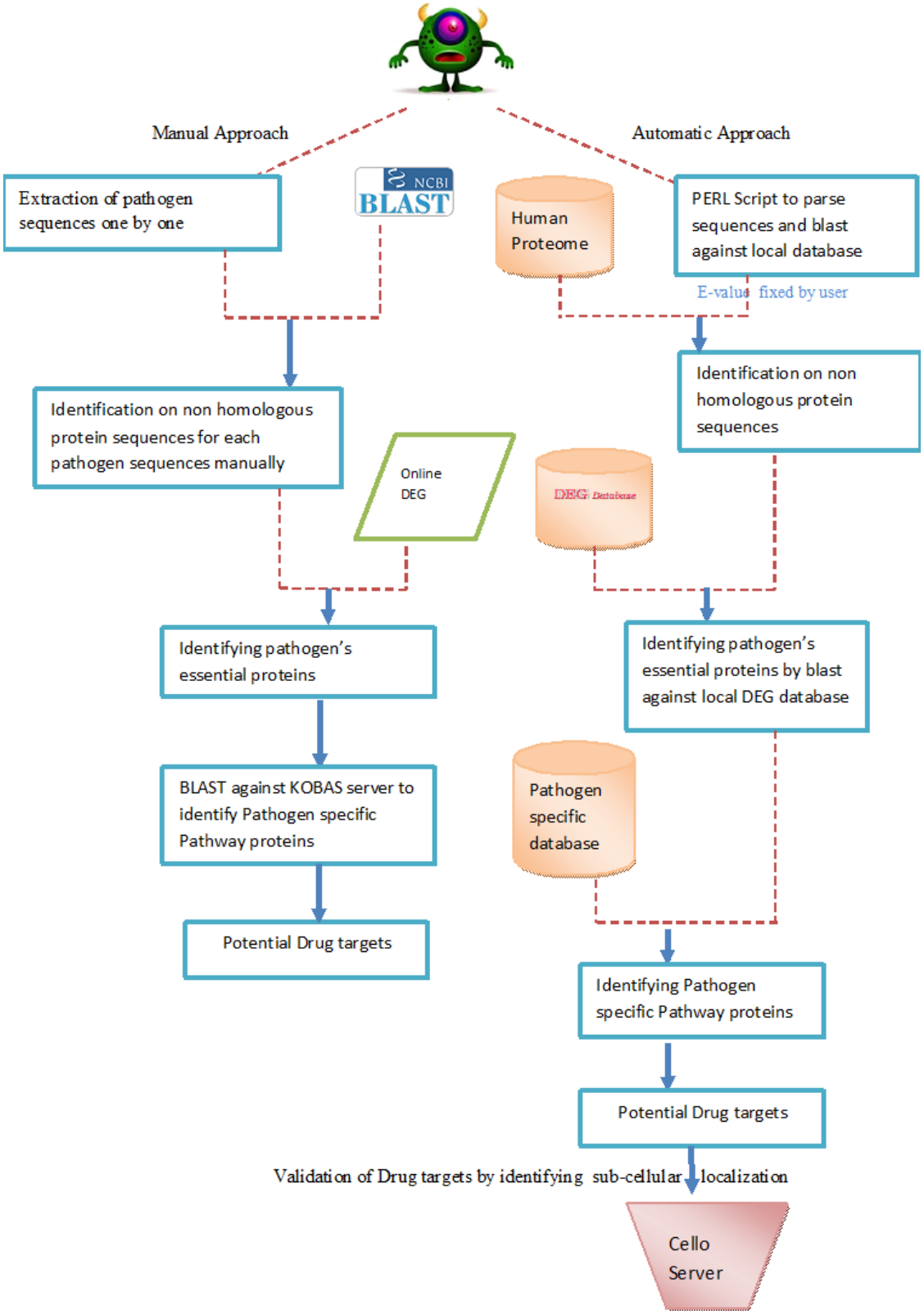

Figure 1: Computational Pipeline for Drug Target identification

ISSN: 0975-8585

Implementation of Computational pipeline

Figure 2 depicts the various navigation steps carried out in the design and development of the drug target identification server .Any *in silico* approach includes the integration of two or more databases. This integration becomes more valuable when a pipeline is designed for a wide range of organisms and ensures that targeting the drug targets would impair the pathogen.The graphical user interface of the server was designed with the aim of facilitating easy accessibility for the user. The front end of the application was designed projecting the main objective of predicting drug targets as an option along with breaking down the three-step procedure of the server into three options in order to help the user directly proceed to a step of his necessity in accordance with his data. "Identifying drug targets" hyperlink provides an option to upload the file containing the protein sequences of the organism in study or paste the list of sequences in the text box. The various user options are provided to choose the organism name for which the protein sequences are submitted and also the matrix type of BLOSUM and PAM variants .The default matrix-type is fixed as BLOSUM 62 and finally The E-value can also be entered by the user in the failure of which, the default value of 5 is considered for the execution of the process.The next step in the pipeline is 'Identification of essential proteins' page has very similar options in comparison to the input page for the identification of non-homologous sequences process. The additional feature available is the option that allows the user to set the value for the bit-score. The default bit-score value has been fixed as 100.Followed by 'Identification of proteins corresponding to pathogen specific pathways' page also has options that allow the user to provide the matrix type, the bit-score and the e-value.

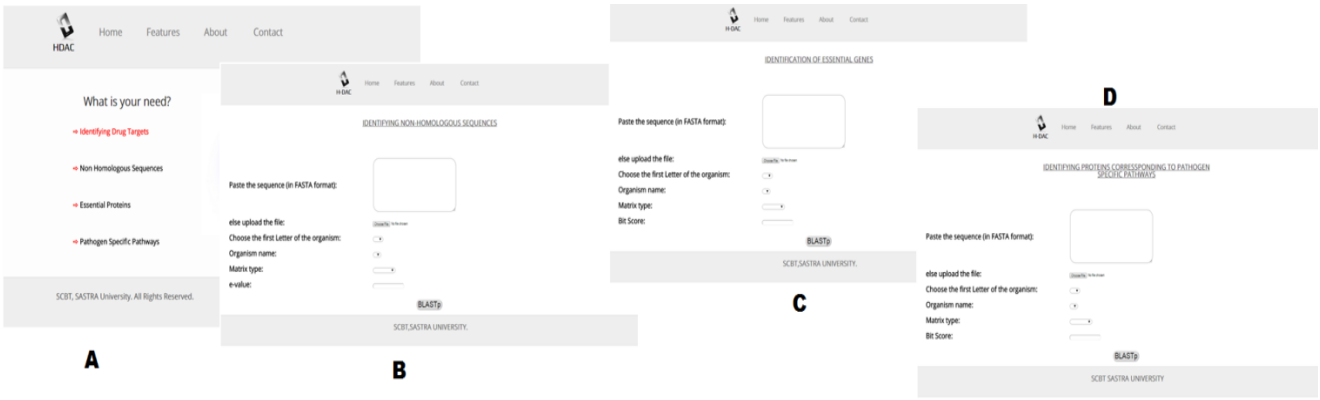

Figure 2: (A) Homepage of the webserver (B) 'Identification of non-homologous sequences in the server (C) Identification of essential proteins' (D) Identification of pathogen pathway specific proteins' in the server

RESULTS AND DISCUSSION

The protocol that had been locked in on as the best way to predict drug targets was tested with three case studies – *Mycobacterium tuberculosis* (reviewed), *Vibrio cholerae* (reviewed) and *Vibrio cholerae* (unreviewed).

Case Study 1: *Mycobacterium tuberculosis* (reviewed)

The protein sequences of the strain – *Mycobacterium tuberculosis* H37Rv – were fetched from the Uniprot database. There were 2037 protein sequences available for the considered case. These 2037 proteins were subjected to BLASTp against the database comprising of the entire protein sequences of *Homo sapiens* and the protein sequences that satisfied the user determined criteria (e-value > 5 in our case) were separated. 221 non-homologous sequences were obtained and these proteins were subjected to BLASTp operation against the 614 essential proteins of *Mycobacteirum tuberculosis* and the proteins that fit the user determined criteria (e-value < 0 and bit-score > 100 in our case) were gathered resulting in 47 proteins to be determined as essential. These 47 proteins were then submitted to a BLASTp operation against the proteins that corresponded exclusively to the pathways of *Mycobacterium tuberculosis* and the proteins that met the criteria (e-value <0 and bit-score >100 in our case) were settled on as to be proteins involved with the pathogen specific pathways and were labelled as potential drug targets. The 9 potential drug targets obtained by this procedure were verified with the CELLO server with respect to their sub-cellular localization and were validated. All 9 drug targets were found to be in the extracellular region which confirmed their potentiality as shown in **Figure 3**.

ISSN: 0975-8585

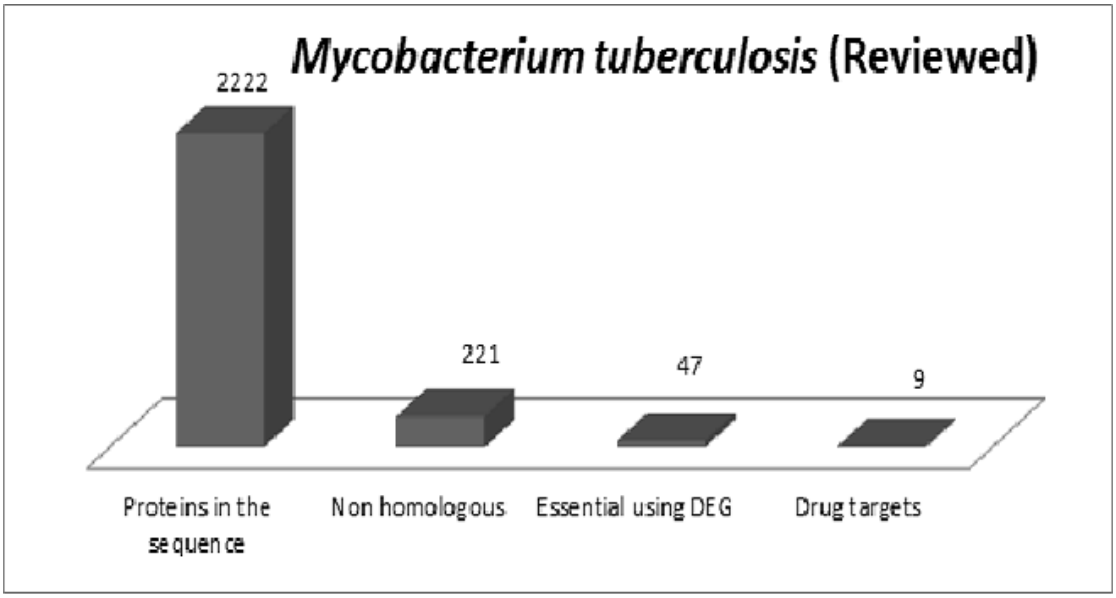

Figure 3: Graph for *Mycobacterium tuberculosis* (reviewed) case study

Case Study 2: *Vibrio cholerae* (reviewed)

The protein sequences of the strain – *Vibrio cholerae* serotype O1 (strain ATCC 39315 / El Tor Inaba N16961) – were fetched from the Uniprot database. There were 956 protein sequences available for the considered case. These 956 proteins were subjected to BLASTp against the database comprising of the entire protein sequences of *Homo sapiens* and the protein sequences that satisfied the user determined criteria (e-value > 5 in our case) were separated. 179 non-homologous sequences were obtained and these proteins were subjected to BLASTp operation against the 779 essential proteins of *Vibrio cholerae* and the proteins that fit the user determined criteria (e-value < 0 and bit-score > 100 in our case) were gathered resulting in 49 proteins to be determined as essential. These 49 proteins were then submitted to a BLASTp operation against the proteins that corresponded exclusively to the pathways of *Vibrio cholerae* and the proteins that met the criteria (e-value <0 and bit-score >100 in our case) were settled on as to be proteins involved with the pathogen specific pathways and were labelled as potential drug targets. The 1 potential drug target obtained by this procedure was verified with the CELLO server with respect to their sub-cellular localization and was validated. The drug target was found to be in the cytoplasmic region which confirmed its potentiality as shown in Figure 4.

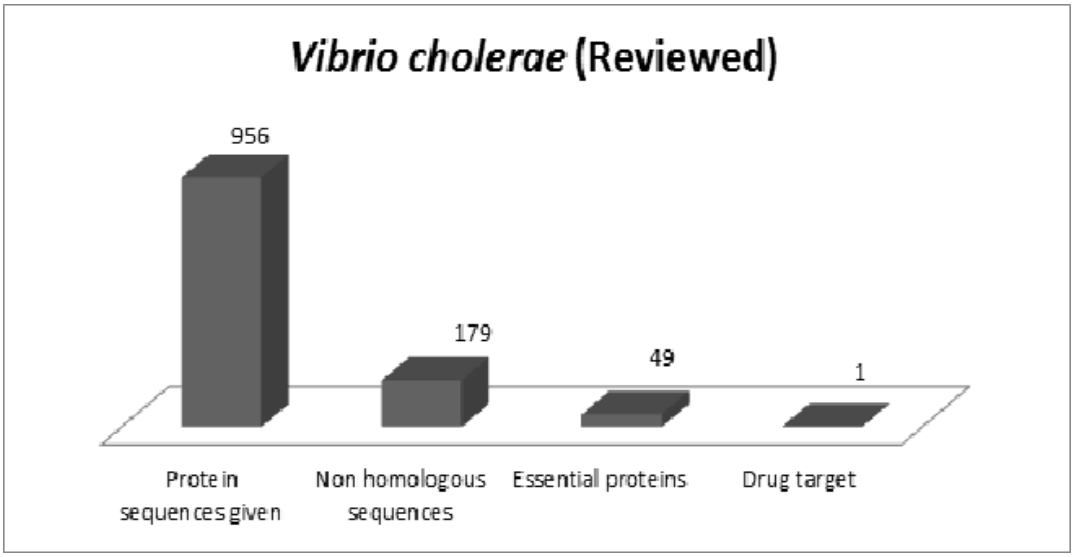

Figure 4: Graph for *Vibrio cholerae* (reviewed) case study

ISSN: 0975-8585

Case Study 3: *Vibrio cholerae* (unreviewed)

The unreviewed protein sequences of the strain – *Vibrio cholerae* serotype O1 (strain ATCC 39315 / El Tor Inaba N16961) – were fetched from the Uniprot database. There were 2833 protein sequences available for the considered case. These proteins were subjected to BLASTp against the database comprising of the entire protein sequences of *Homo sapiens* and the protein sequences that satisfied the user determined criteria (e-value > 5 in our case) were separated. 940 non-homologous sequences were obtained and these proteins were subjected to BLASTp operation against the 779 essential proteins of *Vibrio cholerae* and the proteins that fit the user determined criteria (e-value < 0 and bit-score > 100 in our case) were gathered resulting in 128 proteins to be determined as essential. These 128 proteins were then submitted to a BLASTp operation against the proteins that corresponded exclusively to the pathways of *Vibrio cholerae* and the proteins that met the criteria (e-value <0 and bit-score >100 in our case) were settled on as to be proteins involved with the pathogen specific pathways and were labelled as potential drug targets. The 5 potential drug targets obtained by this procedure were verified with the CELLO server with respect to their sub-cellular localization and were validated. 4 drug targets were found to be in the cytoplasmic region and 1 drug target in the inner membrane region which confirmed their potentiality as shown in Figure 5.

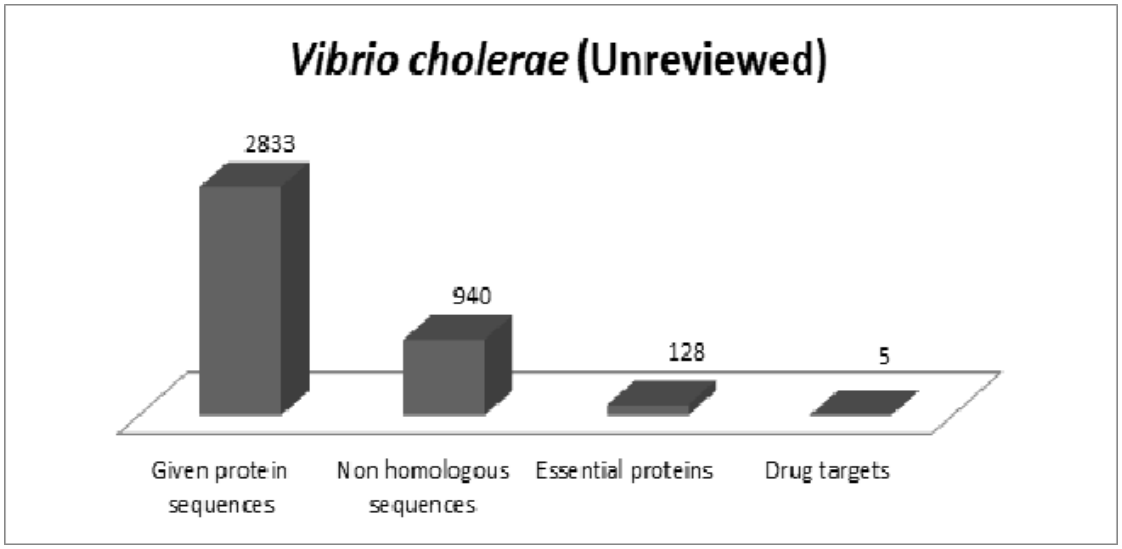

Figure 5: Graph for *Vibrio cholerae* (unreviewed) case study

Protein ID	Sub-Cellular Localization
P9WP21	Extracellular membrane
P9WI31	Extracellular membrane
P9WI33	Extracellular membrane
P9WI07	Extracellular membrane
P9WI05	Extracellular membrane
P9WI03	Extracellular membrane
P9WI09	Extracellular membrane
P9WHZ9	Extracellular membrane
P9WHZ3	Extracellular membrane

Table 1 : Listing the Protein ID's and Sub-cellular Localizations of Potential Drug targets identified for *Mycobacterium tuberculosis* (Reviewed)

Protein ID	Sub-Cellular Localization
Q9KPY2	Cytoplasm

Table 2 :Listing the Protein ID's and Sub-cellular Localizations of Potential Drug targets identified for *Vibrio cholerae* (Reviewed)

ISSN: 0975-8585

Protein ID	Sub-Cellular Localization
Q9KUE2	Cytoplasm
Q9KRE8	Cytoplasm
Q9KQY9	Cytoplasm
H9L4T6	Cytoplasm
Q9KTK5	Inner Membrane

Table 3 : Listing the Protein ID's and Sub-cellular Localizations of Potential Drug targets identified for *Vibrio cholerae* (Unreviewed)

CONCLUSION

Prediction of novel drug targets is of utmost importance in the drug discovery process. Equally important is an automated procedure that is not time-consuming and is not restricted to a particular disease or pathogen. Hence, we believe that our server that performs the process of drug target prediction in an efficient manner and is compatible for a wide range of organisms, is a very important contribution to the process of drug discovery. Additional features like '3D Model Generation' and 'Inhibitor Docking' will be added to the server in future and it will to be an effective *in silico* technique in Drug designing process for design and development in next generation drugs.

ACKNOWLEDGEMENTS

We thank SASTRA University for providing us with a wonderful infrastructure to successfully complete our research. The authors sincerely acknowledge Dr. S. Swaminathan, Dean, Sponsored Research, SASTRA University for providing the scope of the work and for encouraging our research

REFERENCES

[1] Peter Imming, Christian Sinning ,Achim Meyer Nature Reviews Drug Discovery 2006 ;5:821-834.
[2] Georg CT, Angelo Reggiani Trends in Pharmacological Sciences 2001; 22: 23-26.
[3] Jeffrey Augen Drug Discovery Today 2002 ;7: 315-323.
[4] Zhenran JIANG , Yanhong ZHOU Journal of Integrative Bioinformatics 2005;14:2(1)
[5] Chain PS,Grafham DV,Fulton RS,Fitzgerald MG,Hostetler J,Muzny D,Ali J,Birren B,Bruce DC,Buhay C,Cole JR,Ding Y,Dugan S,Field D,Garrity GM,Gibbs R,Graves T,Han CS,Harrison SH,Highlander S,Hugenholtz P,Khouri HM,Kodira CD,Kolker E,Kyrpides NC,Lang D,Lapidus A,Malfatti SA,Markowitz V,Metha T,Nelson KE,Parkhill J,Pitluck S,Qin X,Read TD,Schmutz J,Sozhamannan S,Sterk P,Strausberg RL,Sutton G,Thomson NR,Tiedje JM,Weinstock G,Wollam A,Detter JC Science 2009;326:236-237.
[6] Anirban Dutta, Shashi kr.Singh, Payel Gosh, Runni Mukherjee, Sayak Mitter, Debashis Bandyopadhyay In silico Biology 2006 ; 6: 43-47.
[7] Debmalya Barh, Anil Kumar In silico bilogy 2009; 9:225-231.
[8] Bairoch A,Apweiler R,Wu CH,Barker WC,Boeckmann B,Ferro S,Gasteiger E,Huang H,Lopez R,Magrane M,Martin MJ,Natale DA,O'Donovan C,Redaschi N,Yeh LSL Nucleic Acids Research 2005 33: D154-D159.
[9] Altschul SF, Gish W, Miller W, Myers EW , Lipman DJ Journal of Molecular Biology 1990 ;215:403-410.
[10] Zhang T, Lin y Nucleic Acids Research 2009 ;37: D455-D458.
[11] Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L Nucleic Acids Research 2011 ;39:W316-W322.
[12] Butt AM, Nasrullah I, Tahir S, Tong Y Plos One. 2012 ;7: e43080.
[13] Yu CS, Lin CJ, Hwang JK Protein science 2009 ;13: 1402-1406.