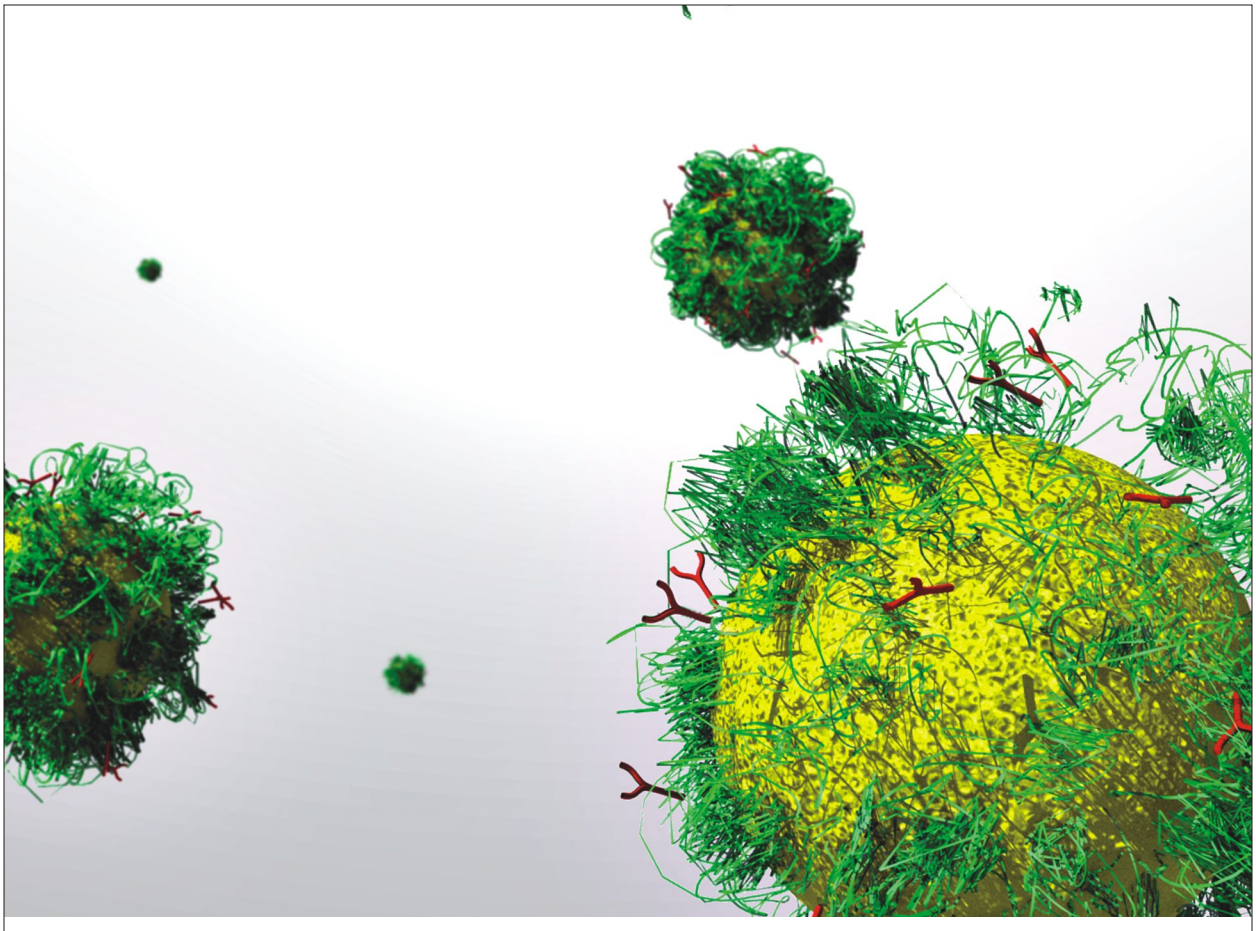

American Journal of
Drug Discovery and Development

ISSN 2150-427X

Academic
Journals Inc.

www.academicjournals.com

American Journal of Drug Discovery and Development 1 (1): 70-74, 2011
ISSN 2150-427x / DOI: 10.3923/ajdd.2011.70.74

Database of *in silico* Predicted Potential Drug Target Proteins in Common Bacterial Human Pathogens

Sushil Kumar Shakyawar, Arun Goyal and Vikash Kumar Dubey
Department of Biotechnology, Indian Institute of Technology, Guwahati-781 039, Assam, India

Corresponding Author: Dr. Vikash Kumar Dubey, Department of Biotechnology, Indian Institute of Technology, Guwahati-781 039, Assam, India Tel: +91- 361-2582203 Fax: +91-361-2582249

ABSTRACT

A Drug Target Protein (DTP) Database has been developed having mainly *in silico* predicted potential drug target proteins and non human homologous genes in bacterial human pathogens. The drug targets of pathogens available in literature are selected to store in the database. Currently 10 bacterial pathogens are considered as initials of the database with their general information (disease caused, common symptoms, available drugs etc). Secondly, the two statistical applications mainly (A) Amino Acid Dynamics and (B) amino acid composition were added in the database for analysis of sequence of drug target proteins. The webpage and data connectivity was maintained by PHP codes. The utility of database is for drug designing and vaccine development for selected pathogen.

Key words: Drug discovery, statistical analysis, computational biology

INTRODUCTION

To strengthen the work bench of finding, analysis of drug targets in human pathogen is current challenge for pharmaceutical and medical research. Huge availability of protein sequence and computational approaches for identifying target proteins/genes has been a great support for further experimental work. Many drug targets have been identified in such pathogens by various computational tools (Suthar *et al*., 2009). Another extensive study related to identification of drug targets of selected foodborne pathogens by our group is currently under way (unpublished data). Up to date there is no database to structure these data; therefore a Drug Target Protein (DTP) Database has been developed which contains mainly proteins in pathogens identified as potential drug targets. List of non human homologous genes and general information (disease, symptoms, preventive measures etc.) of the pathogens are also included. Simultaneously, to analyse the annoted drug target protein sequence a PHP based web tool Statistical analysis has been interpreted in the database. Currently, there are few drug targets database are available but they list human proteins as drug targets for various human diseases (http://www.sciclips.com/sciclips/drug-targets-main.do). These databases do not include drug targets of pathogens. Kumar *et al*. (2007) have predicted drug targets for *Brugia malayi*, causing lymphatic filariasis. Moreover, there are few other drug target prediction reports which are complied as The TDR Targets Database (Aguero *et al*., 2008). However, TDR Targets Database includes drug targets for pathogen causing tropical diseases only.

The amino acids sequence and frequency of various amino acids dictates properties of the protein (Kong *et al*., 1992; Yoon *et al*., 2000). The database server facilitates two kind of sequence analysis (a) amino acid dynamics and (b) amino acid composition in the target protein. Amino acid dynamics measures frequency of occurrence of a particular amino acid along the protein sequence whereas second application is simply finding composition of various amino acids in whole protein sequence. The plot of frequency along the selected sequence is useful to analyze segment wise characteristic profile of the protein as well as nature of its core. The relative trend of occurrence and composition of amino acids gives idea about nature of the protein. In case of enzyme protein, dynamics of specific amino acid plays major role in maintaining 3D structure of protein and further enzyme specificity and its activity.

MATERIALS AND METHODS

The experiments were done at Indian Institute of Technology Guwahati by Mr. Sushil Kumar Shakyawar as part of his B.Tech project during May 2009 to July 2009.

Data collection: Available general information (disease caused, common symptoms, available drugs etc.) about pathogens in the database were collected from Wikipedia or literature. Some of the drug targets included in the database are already published / communicated by our group. The protein sequences of selected pathogens were downloaded from NCBI and analyzed as described earlier (Suthar *et al*., 2009). In brief, the gene sequence of *Homo sapiens* and various pathogens were downloaded from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nlm.nih.gov/genomes/). The essential gene sequences were downloaded from the Database of Essential Genes (Zhang and Lin, 2009; Zhang *et al*., 2004) using the open GENparser (https://launchpad.net/opengenparser). Pathogens paralog sequences at 60% identity were purged using CD-HIT and excluded from further analysis (Li *et al*., 2001). Subsequently, the results were subjected to BLASTP analysis against human protein sequences to identify non-homologous genes in *Leishmania infantum* at an expectation E-value cutoff of 10-3. The non-homologs were aligned with sequences acquired from DEG using BLASTP at an E-value cutoff of 10-10 and 30% identity.

RESULTS AND DISCUSSION

Pathogenic infections still remain a leading cause of global disease burden. The impact of these diseases in India and across the globe has been tremendous at socioeconomic and public health levels. As most of the pathogens are developing resistance against the available drugs, If the new drug development efforts are stopped, current trends suggest that some diseases will have no effective therapies within the next ten years (Levin, 2001; Johnsen *et al*., 2009). Thus the development of new drug against bacterial disease is most urgent and continued requirement. As in addition to strong public health structure, the strategy to combat these diseases needs a, effective communication among the general public, we will include some other useful details like common symptoms, available treatment etc in the database which may be a valuable resource. Few studies have been reported about identification of drug targets by experimental as well as *in-silico* methods from pathogens (Smith, 2004; Chong *et al*., 2006; Suthar *et al*., 2009) but currently there are not database available listing these drug targets proteins. A Drug Target Protein (DTP) Database was prepared using data available in literature or published/submitted by our group. Additionally, sequence analysis tool is included in the web server.

A simple text based web model a (Fig. 1) of the database was developed to store and avail all the information and data generated which is stored in .csv files with connectivity of files by PHP

Fig. 1: Screenshot of the database

programming. Further in utilization of the database the two tools were interpreted within the database for analysis of the protein sequence especially those which were annoted as potential drug targets. A completely user defined input based tool was developed in the database to study variation of amino acid along the sequence of protein to study its residual nature. The algorithm of the tool is based on frequency calculation in residues of 10 amino acids along the sequence. The whole protein sequence was first chopped into group of 10 residues along the sequence. In each residues the frequency of user selected amino acid was calculated as following-

$$\text{Frequency of amino acid in a residue of 10 amino acid} = \frac{\text{(No. of amino acid in selected residue)}}{\text{(Total No. of amino acid in whole sequence)}}$$

The frequency along length of sequence has been used to study dynamics of amino acid. Secondly the compositional analysis of all the 20 amino acids in selected protein sequence were performed in single step and histogram was plotted as output file. Further search application developed in database facilitate user to search a gene ID mainly in drug target protein and nonhuman homologous database.

The developed database is useful resources for designing more effective vaccines and drugs for pathogens which are more resilient or for which drugs are yet to be discovered. The final proteins annoted as potential drug targets can be retrieved from Using the Database, Statistical analysis for two tools developed for residual and sequential analysis of the selected protein and Search for querying the gene ID whether it belongs to drug target of non human homologous.

For validation of the tools developed, DNA polymerase III subunit beta protein of *Mycobacterium tuberculosis* was considered for dynamic analysis of amino acid along its sequence as well as compositional analysis. Dynamic analysis all four amino acids viz. Phe, Lle, Ile and Leu along the sequence were studied by the tool in our database (Fig. 2a, b). Commonly, protein

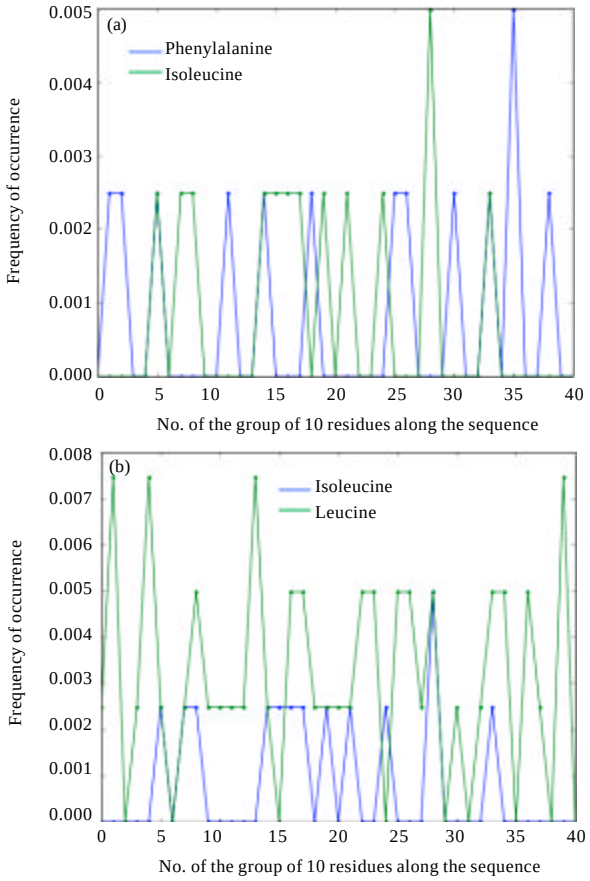

Fig. 2: (a) Dynamics of phenylalanine and isoleucine in DNA polymerase III subunit beta protein of *Mycobacterium tuberculosis*. (B) Dynamics of leucine and Isoleucine in DNA polymerase III subunit beta protein of *Mycobacterium tuberculosis*

function is predicted by aligning the sequence with other protein sequence of known function. However, this method fails when the query sequence does not show significant sequence similarity with other protein. We are further trying to understand if function of a protein can be predicted by analysis of various amino acids frequencies along the protein sequence. The database is a valuable resource of researcher for planning experimental studies.

ACKNOWLEDGMENT

Assistance of Mr. Nanu Alan Kachari (Scientific Officer Gr.II, Department of CSE, IIT Guwahati) in preparation of the database is highly acknowledged. Infrastructural facility provided by IIT Guwahati is acknowledged.

REFERENCES

Aguero, F., B. Al-Lazikani, M. Aslett, M. Berriman and F.S. Buckner *et al.*, 2008. Genomic-scale prioritization of drug targets: The TDR targets database. Nat. Rev. Drug Discovery, 7: 900-907.

Chong, C.E., B.S. Lim, S. Nathan and R. Mohamed, 2006. *In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In silico* Biol., 6: 0031-0031.

Johnsen, P.J., J.P. Townsend, T. Bohn, G.S. Simonsen, A. Sundsfjord and K.M. Nielsen, 2009. Factors affecting the reversal of antimicrobial-drug resistance. Lancet Infect. Dis., 9: 357-364.

Kong, XP., R. Onrust, M. O'Donnell and J. Kuriyan, 1992. Three-dimensiona structure of the beta subunit of E. *coli* DNA polymerase III holoenzyme: A sliding DNA Clamp. Cell, 69: 425-437.

Kumar, S., K. Chaudhary, J.M. Foster, J.F. Novelli and Y. Zhang *et al.*, 2007. Mining predicted essential genes of Brugia malayi for nematode drug targets. Plos One, 2: 1189-1189.

Levin, B.R., 2001. Minimizing potential resistance: A population dynamics view. Clin. Infect. Dis., 33: S161-S169.

Li, W., L. Jaroszewski and A. Godzik, 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17: 282-283.

Smith, C., 2004. Drug target identification: A question of biology. Nature, 428: 225-231.

Suthar, N., A. Goyal and V.K. Dubey, 2009. Identification of potential drug targets of *Leishmania Infantum* by in-silico genome analysis. Lett. Drug Des. Discovery, 6: 620-622.

Yoon, H.G., H.Y. Kim, Y.H. Lim, H.K. Kim, D.H. Shin, B.S. Hong and H.Y. Cho, 2000. Identification of essential amino acid residues for catalytic activity and thermostability of novel chitosanase by site-directed mutagenesis. Applied Microbiol. Biotechnol., 56: 173-180.

Zhang, R., H.Y. Ou and C.T. Zhang, 2004. DEG, a database of essential genes. Nucleic Acids Res., 32: 271-272.

Zhang, R. and Y. Lin, 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res., 37: 455-458.