IMPORTANT DATABASES AND TOOLS TO IDENTIFY PROMISING DRUG TARGETS BY SUBTRACTIVE GENOMICS APPROACH – A REVIEW

Preeti Dora[1], V.G. Shanmuga Priya[2], Preeti.S.H[3], U.M. Muddapur[4], Rakesh.N.R[5]

[1]Department of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum-8, Karnataka, India

[2,3,4,5]R & D Centre, Department of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum-8, Karnataka, India

Abstract

Most of the pathogenic organisms have antibiotic resistant strains hence there is a need to identify new drug targets to design new drugs to combat the disease.Subtractive genomics approach is one of the recently adopted methodologies to identify novel drug targets specific to the pathogen to avoid any cross reactivity and side effects.This methodology uses various tools and databases to find essential proteins that are indispensible for the survival or virulence of the organism andare absent in the host. A survey was done on the databases and tools generally usedby the researchers to carry out this work on different pathogenic organism. Databases like NCBI (National Centre for Biotechnology Information) genome database, DEG BlastP (Database of Essential Genes and Genomes), BlastP (Basic Local Alignment Search Tool), KEGG (Kyoto Encyclopaedia for Genes and Genomes) database, Drug bank database are used in most of the studies. Tools and web servers such as CD HIT (Cluster database at high identity with tolerance) and CELLO are mostly used.

Keywords: - *NCBI (National Centre for Biotechnology Information), DEG BlastP (Database of Essential Genes and Genomes), BlastP (Basic Local Alignment Search Tool), KEGG (Kyoto Encyclopaedia for Genes and Genomes), CD HIT(Cluster database at high identity with tolerance), DEG-BlastP (Database of essential Genes Protein Blast)*

----------***----------

1. INTRODUCTION

The modern approach called Subtractive Genomics approach is one of the most useful and helpful methodology in identification of potential targets to design new drugs. This procedure has been successfully used to identify potential drug targets in various pathogens like *Salmonella typhi*, and *Helicobacter pylori, Neisseria Meningitides serogroup B* and *Betasubunit of DNA polymeraseIII in Streptococcus* species, *etc*.

1.1 Subtractive Genomics

Subtractive genomics approach eliminates the sequences that are homologous to both host and the pathogens and finds those proteins that are essential for the pathogen but absent in the host. The protein sequences that are exclusively present in the pathogen are to be used as a drug targets. Studies based on subtractive genomics approach could facilitate the selection, processing and development of strain-specific drugs against various pathogens. The availability of the complete genome sequence of many microbes has made the in-silico identification of drug targets an easy task. The critical genes or proteins which are indispensible for the survival of the pathogen and which are absent in the host can be screened out using the subtractive genomics approach. Sequences exclusively present in the pathogen and are essential to it are considered for further analysis. The essential proteins that are identified should be involved in replication, survival of the pathogen and various metabolic pathways and should have no human homology so that the drug designed should be effective to the pathogen and should not cause any cross reactivity to the host. It is one of the useful methodologies which help in potential therapeutic drug development.

This procedure is successfully carried out by various authors to identify novel drug targets specific to the pathogen. The procedure is carried out using various databases, tools, web servers etc.The methodology reduces the time required in identification of drug targets. The identification of drug targets through this modern approach helps in designing novel drugs and vaccines specific to the pathogen.

1.2 Studies on Pathogenic Organisms

To identify novel drug targets, subtractive genomics approach has been carried out for various disease causing pathogens.

Subtractive genomics approach has been carried out on *Streptococcus* species (John J. V.V.Umrania, 2009) which consisted of 1859 proteins, the protein sequences having less than 100 amino acids were removed from the list. Among

1859 sequences few were eliminate and remaining protein set were subjected to CD HIT at 60% threshold identity which compares and clusters biological sequences. It resulted in 802 protein sequences. These set of protein sequences were then subjected to BlastP to identify non homologous proteins which are specific to the pathogen to the pathogen. The resulted protein sequences that were Non homologous were subjected to DEG BlastP to identify essential proteins that are indispensible for the survival of the pathogen. There were 406 Non homologous sequences among them 283 were found to be essential proteins. Pathway analysis of essential proteins was performed using KEGG database, the proteins specific to the pathogen were considered as putative drug targets.

The same process was carried out for *Samonella typhi*(Bhawna Rathi, 2012) which causes typhoid. The complete proteome of the pathogen and host were retrieved from Swiss-prot and NCBI genome database which consisted of 4178 protein sequences among them 300 protein sequences were found to be essential. CD HIT analysis was carried out to exclude duplicate proteins. The pathway analysis of 300 proteins was carried out using KEGG pathway among these 149 protein sequences were exclusively involved in several pathways of the pathogen.8 pathways were exclusively present in the pathogen which consists of 27 enzymes which can be used as drug targets. Subcellular localization of essential proteins was carried out using PA-SUB which revealed 11 proteins lie on the outer membrane of the cell which could be probable vaccine targets.

Subtractive genomics approach was also carried out for *Neisseria meningitides* serogroup B (Aditya Sarangi, 2009). The complete proteome of the pathogen was retrieved from Swiss-prot and NCBI. The protein sequences were then subjected to BlastP against human proteome to exclude non-human homologous. Essential proteins were identified using DEG BlastP against human. Among 1461 protein sequences 362 proteins were found to be essential to the bacterium. Metabolic pathways analysis was done using KAAS (Kegg automatic annotation server) which revealed 35 enzymes specific to the pathogen. Subcellular localization prediction was carried using v2.5 (PA-SUB).

In silico identification of putative drug targets was carried for *Klebsiella pneumonia* MGH78578 (John Georrge and Valentina Umrania, 2011). The complete proteome of the pathogen was retrieved from Uniprot which consisted of 5126 protein sequences. Removal of duplicate proteins was carried out using CD HIT web server which resulted in 4462 protein sequences, these protein sequences were than subjected to BlastP against human proteome which revealed in 2321 protein sequences that are non-homologous to human. To identify essential proteins DEG BlastP was carried out which resulted in 453 protein sequences among them 108 pathways were identified using KEGG. Drug bank analysis was also carried and 3proteins were used as drug targets.

Subtractive genomics approach was carried for the identification of novel drug targets in *C.trachomatis* strain *(*Khalida Shoukat, Nadia Raheed, Mohammed Sajid, 2012) .The complete proteome was retrieved from NCBI genome database. The complete proteome was retrieved from NCBI which consisted of 895 protein sequences, the duplicate protein sequences were excluded using CD HIT web server at 80% threshold. BlastP was performed to identify non-homologous protein sequences which revealed 623 protein sequences were non homologous to the host genome. BlastP against DEG database resulted in 203 protein -sequences that are essential to C.trachomatis D/UW/-3/Cx. Among 203 protein sequences 39 were found to be the part of membrane of pathogen using PSORT tool. Metabolic pathway analysis was carried out using KAAS server which revealed that 7 proteins were unique to the pathogen.

2. GENERAL WORKFLOW CARRIED OUT:

The above said studies all seem to follow a common work flow. It includes:

1. Retrieval of complete proteome from NCBI or Swiss-prot
2. Removal of sequences having less than 100 amino acids.
3. Removal of paralogs using CD HIT web server.
4. Identification of non-homologous protein sequences using BlastP.
5. Identification of essential proteins using DEG BlastP.
6. Subcellular localization of essential proteins using PA-SUB tool.
7. Metabolic pathway analysis using KEGG database.
8. Drug bank database analysis.

2.1 Common Tools and Databases used for the Study

For the identification of the essential proteins using subtractive genomics approach various tools and databases are used for the study.

NCBI (National Centre for Biotechnology Information): The complete proteome of the pathogen can be retrieved using NCBI genome database.(www.ncbi.nlm.nih.gov/). It is a database that contains information about chemicals and Bioassays, Data and software, DNA and RNA, Domains and Structure, Genes and Expression, Genetics and Medicine etc. It also contains databases, downloads, submissions, and tools.

CD HIT web server: Cluster database at high identity with tolerance is used to cluster and compare biological sequences. The complete proteome might contain redundant sequences or duplicate sequences that can be eliminated at a threshold of 60% using CD HIT.

BlastP (Basic local alignment search tool): The proteins are subjected to BlastP (protein-protein blast) against human proteome with expectation value (E value) e-4 to identify.

DEG BlastP (Database of essential genes and genomes): DEG BlastP against human proteome is performed to identify essential proteins that are indispensible for the survival of organism. For protein sequence to be essential there are certain parameters thatshould be considered. The parameters are: Sequence identity, E-Value, Bit-score.Sequence Identity refers to percentage of matches of the same amino acid residues between two aligned sequences. For a protein sequence to be significant identity should be greater than 35%. E-Value is the Expect value and provides information about the likelihood that a given sequence matches is purely by chance. Lower the E-value more significant the sequence match. E-Value should be zero or negative in order to be significant. If the E-value is less than 10; the sequence under consideration is either unrelated or distantly related. **Bit score** is another prominent statistical indicator used in addition to the E-Value in Blast output. The bit score measures sequence similarity independent of query sequence length and database size. Higher the bit score, more highly significant the match is. All the protein sequences with bit score more than 100 bits are considered for further analysis.

CELLO TOOL: CELLO tool was used to predict the protein location. It is an approach based on the two-level SVM system. The first level comprises a number of SVM classifiers each based on a specific type of feature vectors derived from sequences. The second level SVM classifiers function as the jury machine to generate the probability distribution of decisions for possible localization. Sub cellular localization involves the computational prediction of where a protein resides in the cell. Prediction of protein subcellular localization is an important component as it predicts the protein function and genome annotation, and it can aid the identification of targets. Subcellular localization analysis of essential protein sequences can be done by any of the prediction tool.

KEGG (Kyoto encyclopaedia for genes and genomes):KEGG stands for Kyoto Encyclopaedia for Genes and Genomes, has become major resource for pathway analysis and contains a wealth of data associated with pathways, genes, genomes, chemical compounds and reaction information, in addition to links to outside resources such as PubMed (Kanehisa el al., 2006) (www.genome.jp/kegg/). The metabolic pathway analysis can be studied using KEGG database, the pathway analysis of essential non-homologous proteins is important to know the function of the proteins.

DRUG BANK DATABASE: Drug bank database is a unique bioinformatics and chemo informatics resource that combines both drug data and comprehensive with drug target information. It gives us information whether the identified targets are already targeted or need to be targeted. (www.drugbank.ca/)

3. CONCLUSION

The step by step screening of protein sequences using subtractive genomics approach helps in identification of pathogen specific drug targets which helps in designing therapeutic drugs in treating various infections and diseases.The procedure carried out by researchers to identify strain specific drug targetsincludes retrievalof complete proteome from NCBI or Swiss-prot, removal of duplicate sequences using CD HIT web server, identification of essential proteins using DEG-BlastP against human, identification of non-homologous protein sequences using Blastp against human proteome. PA-SUB tool can be used for sub-cellular localization prediction and pathway analysis can be carried out using KEGG pathway database. Drug bank analysis gives us the information about the selection of putative drug targets. Hence this indicates that these basic databases and tools are highly informative and useful and can be engaged to find species specific drug targets to combat various pathogenic organisms.

REFERENCES

[1] Sarangi AN, Aggarwal R, Rahman Q, Trivedi N(2009) Subtractive Genomics approach for in Silico Identification and Characterization of Novel Drug Targets in Neisseria Meningitides Serotype B. J Comput Sci Syst Biol 2:255-258.doi: 10.4172/jcb.1000038

[2] Lew, J.F., D.L. Swerdlow, M.E. Dance, P.M. Griffin, C.A. Bopp, M.J. Gillenwater, T. Mercatante, R.I. Glass. 1991. An outbreak of shigellosis aboard a cruise ship caused by a multiple-antibiotic-resistant strain of *Shigella flexneri*. *American Journal of Epidemiology*. 134(4): 413-420

[3] Parija SC (Jan 1, 2009). *Textbook of Microbiology & Immunology*. India: Elsevier. ISBN 8131221636.

[4] Sack RB, Rahman M, Yunus M, Khan EH. Antimicrobial resistance in organismscausing diarrheal disease. Clin In-fect Dis 1997;27 Suppl 1: S102-5/>

[5] Barh, D., et al. (2011). Drug development research, 72, 162-177

[6] Ying Huang, Beifang Niu, Ying Gao, Limin Fu and Weizhong Li, CD HT suite: a web server for clustering and comparing biological sequences. Bioinformatics, 2010 (26): 680-682

[7] Hao Luo, YanLin, Feng Gao, Chun –Ting Zhang and Ren Zhang, (2014) DEG10, an update of the protein-coding genes and non-coding genomic elements Nucleic AcidsResearch 42, D574-D580

[8] Yu CS, Chen YC, Lu CH, Hwang JK: Prediction of Sub-cellular localization. Proteins: Structures, Function and Bioinformatics 2006, 64: 643-651

[9] www.genome.jp/kegg/

[10] http://www.ncbi.nlm.nih.gov/blast/BLAST.cgi

[11] Barh, D., et al. (2011). Drug development research, 72, 162-177

[12] Bhattacharya D, Bhattacharya H, Sayi DS, Bharadwaj AP, Roy S. Changing patterns and widening of antibiotic resistance in *Shigella*

[13] CD HIT (cd hit.org/)

[14] DEG and Cello (www.cello.life.nctu.edu.tw/) tools