Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

Subtractive Proteomics for Identification of Drug Targets in Bacterial Pathogens: A Review

Shalini Maurya[1]
Department of Bioengineering Integral University,
Lucknow Uttar Pradesh, 226026,
India

Salman Akhtar[2]
Department of Bioengineering Integral University,
Lucknow Uttar Pradesh, 226026,
India

Mohammed Haris Siddiqui[3]
Department of Bioengineering Integral University,
Lucknow Uttar Pradesh, 226026,
India

Mohammed Kalim Ahmad Khan[4*]
Department of Bioengineering Integral University,
Lucknow Uttar Pradesh, 226026,
India

Abstract : **The first step in drug and vaccine discovery is the identification of the target. Subtractive proteomics can be widely used for the process. This approach has been used to find drug targets for the pathogen, which are resistant to drugs and for which there is no vaccine. This insilico method can reduce the cost and time for finding the drug targets. It involves subtraction of proteins of host and pathogen which provide a set of proteins, which are essential for pathogen but absent in gut flora and host. A review was carried out on a subtractive proteomics approach on twenty-nine pathogens, as reported from 2008 to 2018. These organisms have been categorized under two categories: Multiple drug-resistant organisms and those which have no effective drug to date. By subtractive proteomics the total proteome undergoes a subtraction process at each step, to narrow down to few drug targets. Tools and databases for subtractive proteomics have also been described.**

Keywords: Subtractive proteomics, drug targets, insilico, vaccine

I. INTRODUCTION

The genome sequencing projects of microbial pathogens have been growing rapidly in the past decade. We also have complete information on the genome of human beings. This has led to the development of new methods for studying the interaction between humans and microbial pathogens. This, in turn, has improved various tools to combat a broad range of pathogens [1]. The process of introducing a drug in the market still takes 10-15 years, despite the use of high throughput techniques and synthetic chemistry. Hence, it increases the cost of developing a drug [2].Using integrated genomics, transcriptomics and metabolomics for target discovery in several diseases is a trend these days because it makes the entire process quick and costs effective. Various bioinformatics tools have been developed, several of which involve a comparative analysis of host and pathogen genome, which reveal drug targets which are non-homologous to human and essential for pathogen survival. *In silico* subtractive proteomics is one such tool that is used to find novel drug target and vaccine candidates in pathogens, by comparing human and pathogen proteome and finding a protein which is essential for the survival of pathogen but absent in humans[3].

The current trends in target discovery for most human diseases are the use of computational approaches, with proteomics, integrated genomics, transcriptomics, metabolomics, interactomics, and signalomics, especially for infectious diseases, cancers, neuroendocrinal diseases, and cardiovascular diseases, thereby making the entire process of drug discovery faster and more cost-effective. Currently, to identify novel drug and vaccine targets, genomics and specially insilico comparative, subtractive and functional genomics are widely used, especially to develop effective antibacterial agents and vaccines against bacterial pathogens that are resistant to existing antibacterial regimes, for which a suitable vaccine is not available [4,5].

Many diseases are caused by a bacterial pathogen. So many people die every year due to these diseases. Despite the presence of antibiotics, many resistant strains have developed which makes them difficult to control. The identification of drug targets is the first step in the drug discovery process. Due to the presence of the host and pathogen sequences, strategies have shifted from generic to genomic, proteomic and metabolomics approaches to identify the novel drug targets. New drug targets are required for designing of new antibiotics against drug-resistant pathogens.

Drugs being used currently to treat diseases caused by pathogens, show less to more side effects in humans. Also, repeated use has led to the development of drug resistance in pathogens, so there is a need to find new drugs and drug targets, to be able to combat pathogens efficiently. A novel approach called “subtractive proteomics” has been found to be able to find novel drug targets in pathogens [99]. This is a step ahead towards developing new and effective drug targets. This review elucidates an overview of insilico subtractive proteomics approaches used to identify drug targets in various pathogenic bacteria. We have also discussed the advantages, disadvantages and future prospects of this approach.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

II. CHARACTERISTICS OF A DRUG TARGET

A drug target should be non-homologous to host and an essential protein for the survival of the pathogen. Although ideally, the target should have four properties : (1) It should be essential protein for survival and pathogenesis of target organism (2) target should be druggable (3) structural and functional characterization, with established assays for screening small molecule inhibition (4) it should be different from current drug targets to avoid cross-resistance between them[6].

III. STRATEGIES USED :

It has been shown in figure 1.

A. Identification of paralogous protein

Paralogous proteins are homologous proteins that have diverged within one species. CD-HIT (Cluster database at high identity with tolerance) is a vastly used program for identification of paralogous proteins, which are finally excluded from the analysis [7]. It is a widely used program for clustering genes/proteins to reduce redundancy during the process of identification of drug targets [8,9,10]. The proteome of the pathogen is usually subjected to CD-HIT analysis with sequence identity cut off between 10-90% depending upon requirement [7]. It is vastly accepted to keep sequence identity cut off 60% in order to maintain rigid criteria for removal of duplicate proteins [9,11,12]. Proteins which are more than 60% identical are considered duplicate or paralog protein, which are excluded. The remaining nonparalog protein which has more than 100 amino acids taken for further analysis. It is usually assumed that protein which is less than 100 amino acids are less likely to be essential, hence excluded [13,14].

B. Essentiality analysis:

Essential genes are those which are required for growth, adaptability, and survival of an organism. A deficiency of its protein can be deadly to an organism. Essential genes are most likely to be evolutionarily conserved across taxa and have a common function[15,16,17]. The Database of Essential Gene (DEG)[18] is an online resource that contains information that essential genes in bacteria, fungi, plant and animal, which are experimentally validated[19]. The database can be analyzed for essentiality using DEG. Cut-off for e-value 10^10 and a bit score >100 is mostly used [20,21]BLASTp against DEG is usually done to find essential proteins. The nonhomologous proteins can be screened out.

C. Identification of human non-homologs

BLASTp of homo sapiens proteome with nonparalog protein is usually done with threshold expectation value>10-3 and bit score<100 to identify human nonhomologous protein [14,22,23]. Human homologs are excluded and the rest of the non-homologs are combined together for further analysis. This step is usually included to avoid cross-reactivity of drugs with a protein of host, to prevent its binding with the active site of host homologous proteins [24]. Proteins having ≤35% and for which no hit was obtained, were called as host non-homologs [24].

D. Identification of orthologs in gut flora

Nonhomology analysis with gut flora is essential. The gastrointestinal tract of healthy human houses about 1014 microorganisms [25]. Gut microbiota lives in a symbiotic relationship with the host.They play an essential role in the metabolism of humans. They help in digestion of food particles and prevent pathogenic bacteria from growing inside gut [26]. If we block the gut microbiota proteins, it may lead to an adverse effect on the host. In order to avoid such circumstances, proteins were subjected to homology search by performing BLASTp[27] with gut flora proteome, keeping e-value threshold 0.0001[28].

E. Pathway analysis

Functional annotation of nonhomologous essential proteins was done with their respective metabolic pathway in KAAS(KEGG Automatic Annotation Server)[29] for specificity. The host and pathogen metabolic pathways can be comparatively analyzed using KEGG (Kyoto Encyclopaedia of Genes and Genomes) to find proteins involved in the pathogen-specific pathway for drug target identification. KAAS can provide functional annotation of genes through BLAST comparison with the KEGG GENES database. The result is KO assignments to metabolic proteins. It can automatically generate a pathway to those metabolic proteins.

F. Druggability analysis

Druggability analysis of all shortlisted protein can be done by searching them in the drug bank database[30]. All non-similar, essential and hypothetical sequences of proteins can be screened by performing BLASTp against Drug bank database [30]. Drug bank database contains a number of protein targets with IDs of drug which are sanctioned by the FDA. Only those drug targets which have bit score >100 and E value < 0.005can be called drug targets

G. Virulence analysis

Bacteria, using virulence factors degrade host defense mechanism with help of adhesion, colonization, and invasion, and thus cause disease. VFDB is a database containing four virulence factors. These are offensive, defensive, nonspecific and virulence-associated regulated proteins [31]. The proteome of bacteria can be subjected to BLASTp against the VFDB database. Cut off bit score>100 and e-value 0.0001 is mostly used.

H. Subcellular localization prediction

In microbes, proteins can be found at five subcellular locations. These are extracellular, periplasm, outer membrane, plasma membrane, and cytoplasm. Finding the subcellular location is essential for their categorization as a vaccine or drug target. Proteins in the cytoplasm can be called a drug target and those on the membrane can vaccine targets [32]. A subcellular localization position can be obtained from the UniProt database [33].But in the absence of experimental information, subcellular localization prediction can be done using CELLO[34] and PSORTb[35].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

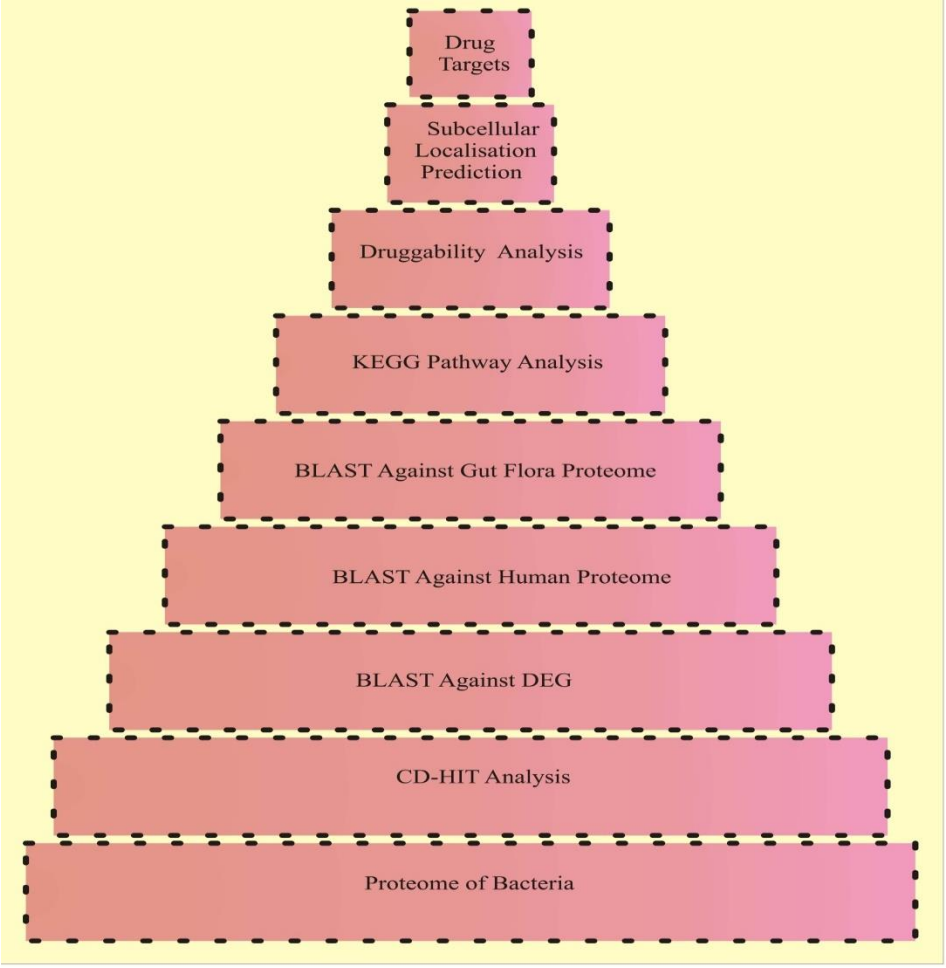

Figure 1: Subtractive proteomics workflow

IV.TOOLS AND DATABASES USED FOR GENOME SUBTRACTION

Table 1 : Tools and databases used for genome subtraction

TOOLS AND DATABASES	UTILITY	WEB ADDRESS	REFERENCES
DATABASES			
UniProt	Reserve of bacterial proteome	https://www.uniprot.org/	[33]
DEG	Contains essential genes	http://www.essentialgene.org/	[18]
KEGG	Pathway comparison and subtraction	http://www.genome.ad.jp/kegg/	[36]
VFDB	Virulence factor resource of various medically significant bacterial pathogens	http://www.mgc.ac.cn/VFs/	[31]
DRUG BANK	Contains known drug targets	http://www.drugbank.ca	[30]
TOOLS			
CD-HIT	Removal of paralogous proteins	http://weizhongli-lab.org/cd-hit/	[7]
CELLO	Subcellular localization prediction	http://cello.life.nctu.edu.tw/	[37]
PsortB	Subcellular localization prediction	https://www.psort.org/psortb/	[38]
NCBI BLAST	Subtraction of nonhuman homolog	https://blast.ncbi.nlm.nih.gov/Blast.cgi	[27]
TID	Platform for subtractive proteomics for drug target identification	http://bmicnip.in/TiD/	[39]
Mgenomesubtractor	Insilico subtractive hybridization	http://202.120.12.134/mGS2/	[40]
TMHMM	Trans membrane domain prediction	http://www.cbs.dtu.dk/services/TMHMM/	[41]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

A. UniProt (https://www.uniprot.org/) :

Swissport, TrEMBL, and PIR protein databases have united to form UniProt (Universal Protein Knowledgebase) consortium. UniProt is a comprehensive and rich database, it is also fully classified. It uses very accurately annotated protein sequence knowledgebase, with extensive cross-references and query interface.TreEMBL and Swissport are its two sections. TrEMBL utilizes extensive cross-references and uses manually curated data at the same time Swissprot utilizes manually curated data. It also provides many non-redundant sequence databases. The Uniprot databases consist of three database layers:

(i)The Uniprot Archive (UniParc) which stores a complete body of publicly available protein sequence data and provides a non-redundant, comprehensive and stable sequence collection.

(ii)The UniProt Knowledgebase (Uniprot) provides a central database of protein sequence with rich, consistent and accurate sequence and functional annotation.

(iii)The Uniprot NREF database (UniRef) provides data collection which is nonredundant and based on the Uniprot knowledge base in order to obtain complete coverage of sequence space at several resolutions.

The proteome is a set of proteins expressed in a cell at a particular time by an organism. Most of UniProt proteomes are based on the translation of a completely sequenced genome, and will normally include sequences that are derived from extra-chromosomal elements like plasmid or organelle genome in an organism where these occur. Some proteomes may also include protein sequences based on high-quality cDNAs that can not be mapped to current genome assembly due to sequencing errors or gaps. Uniprot proteome contains both manually reviewed(UniProtKB/Swiss-Prot) and unreviewed(UniProtKB/TrEMBL) entries. The proportion of reviewed entries varies between proteomes and is obviously greater for the proteomes of intensely curated model organisms. There are about 12,274 reference proteome and 200,686 other proteomes.

Several changes are often incorporated in Uniprot from time to time. The accession number has been expanded to help deal with the increase in sequences. Several protein identifiers and reference proteome have been included to deal with the overwhelming amount of sequencing data that is generated. Annotation scores have been included to identify the proteins with the highest level of functional characterization which will greatly assist in comparative protein sequence analysis. We can search for proteins from UniProt proteome[33].

B. DEG (http://www.essentialgene.org/) :

The genes that are mandatory to support cellular life are essential genes. These minimal genes set are needed for a living cell. Therefore, the functions encoded by these genes set are essential and can be said to be the foundation of life itself [42,43]. The definition of minimal gene set needed to sustain a living cell is of considerable interest as it has significance in practical use and it represents a fundamental question in biology. For example, most of the antibiotics target essential cellular processes, hence essential gene products of microbial cells are a promising target for antibacterial drugs [44].DEG contains all genes that are essential and are continually updated. The proteins encoded by these essential genes are required for basic cell survival and thus present in all cells. A BLAST of the query sequence against DEG can yield homologous genes, which in turn can mean that query sequence is also essential. The essential gene can also be searched using function or name. All records in it can be browsed and extracted. Essential proteins are considered an excellent target for antibacterial drugs.DEG can be accessed online from its website for free.Its website is **http://www.essentialgene.org/**
[18]. DEG boasts of hosting currently available essential genomic elements records, in bacteria like archaea and eukaryotes containing protein-coding genes and non-coding RNAs respectively. Essential genes create a minimal genome, in a bacterium. They form a functional module set that plays a key role in, synthetic biology which is an emerging field.
High throughput sequencing and high-density transposon-mediated mutagenesis have led to significant advancement in research on essential genes under the diverse condition and revised essential gene concept which includes noncoding RNA also [45]. DEG 10 which is a new release of a database of essential genes available at http://www.essentialgene.org, has been developed to accommodate these quantitative and qualitative advancements. In comparison to DEG5 [46], the number of bacteria with saturated genome-wide gene essentiality ahs nearly tripled in DEG10, which has data for 31 bacteria. DEG 10 has nearly 12000 bacterial essential genes, more than twice the number of those in DEG 5.
Performing homologous searches with BLAST program [47] against DEG is common [48-51] and so customizable BLAST tools have been developed.
For the BLAST tools, one can perform BLAST search for a single gene, multiple genes, annotated genomes and unannotated genomes with filters to search to a subset of species or experiments with desirable P-values. A BLAST of DEG with the query sequence can give homologous genes which will mean that query sequence is also essential. These essential proteins are excellent targets for antibacterial drugs. Essential genes can also be searched using function or name. All records can be browsed and extracted.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

C. KEGG (http://www.genome.ad.jp/kegg/):

Kyoto Encyclopedia of Genes and Genomes (KEGG) can perform systemic analysis of gene functions in terms of a network of genes and molecules and in that sense, KEGG has a huge knowledge base. Pathway Database is a major component of KEGG.Pathway Database comprises of graphical diagrams of biochemical pathways, it also includes almost all known metabolic pathways and few of the very known regulatory pathways.

KEGG has multiple objectives. First being the computerization of recent knowledge, this is the knowledge of biochemistry, genetics and other disciplines. This would be done in terms of pathways of interacting molecules and genes.Second being, that KEGG maintains a LIGAND Database[52, 53] i.e. a catalog of chemical, compounds and other substances in the living cell which is linked to the towards predicting biological systems and provide with a new informatics technology.The fourth and final objective is that it can maintain a gene catalog. This catalog will be for all organisms with completely sequenced genome and will also include those selected organisms with partial genomes.

KEGG can be a very useful database. It can help to understand important biological systems, their services and functions like in the cell, organism, and ecosystem in general. It can provide us with very molecular-level information of datasets that are of large scale. These large scale datasets are generated by genome sequencing and other high-throughput experimental technologies and therefore this molecular-level information is important. It is already established in the former paragraphs that KEGG has a huge knowledge base and in that role, it can perform systemic analysis of functions that are done by genes thereby linking this information obtained from genes with the functional information. Within the *insilico* subtractive proteomics, within the metabolic pathways in which they are involved (both host-pathogen and pathogen common) the mapping of identified essential and non-human homolog proteins is done and is present in the KEGG database. It utilizes a comparative method of pathway analysis for both the pathogens and humans. The KEGG database is not only available online for free but is also regularly updated (http://www.genome.ad.jp/kegg/)[36].

D. VFDB (http://www.mgc.ac.cn/VFs/) :

VFDB is a virulence factor database. It is comprehensive and user-friendly. It contains the latest existing knowledge about virulence factors from various bacterial pathogens. Bacterial pathogens pose a great threat to public threat however greater studies on pathogenesis have enhanced our knowledge about disease mechanisms at the molecular level. Forming a database of virulence factors of major bacterial pathogens was necessary for future research. One can access the database using several keywords etc. BLAST search can be used against all VF related genes. VFDB works as a gateway for storing, searching, retrieving and updating information about VFs from bacterial pathogens. We can access the database sat http://www.mgc.ac.cn/VFs/ [31].VFDB was constructed for twin purposes :

To give us thorough and in-depth information about the main virulence factor of the best characterized bacterial pathogen. and second, was to explain the process of bacterial pathogenesis. Thereby bringing in a new approach towards the treatment and prevention of infectious disease.

In order to decode the potential novel or variant pathogens both in emergent outbreaks and in routine clinical practice, the entire genome sequencing is used. It is a big challenge though for the microbiologists or physicians to do an efficient characterization of pathogenomic composition mainly because of their limited bioinformatics skills. Therefore it was introduced to VFDB, an integrated and automatic pipeline, VF analyzer, which systematically identifies known/potential VFs incomplete /draft bacterial genomes. VFanalyzer first constructs within the query genome, orthologous groups and analyzed reference genome from VFDB in order to avoid potential false positives due to paralogs. Then it conducts sequence similarity searches from within the hierarchical prebuilt datasets of VFDB to precisely find untypical/strain-specific VFs.Eventually, the VF analyzer achieves a relatively high specificity and sensitivity without any manual curation with the help of a context-based data refining process for VFs encoded by gene clusters. Apart from this, in order to make it easy for online analysis, an optimized interactive web interface is introduced to present its reports in a comparative pathogenomic style[54-57].

Since the inception of VFDB in 2004, it has always provided up-to-date knowledge of VFs from multiple medically important and significant bacterial pathogens. As such it has become an all in one comprehensive online resource for finding information about the virulence factors of bacterial pathogens.

There were twin reasons for constructing VFDB:

First, providing in-depth coverage: Major virulence factors of bacterial pathogens with their structure features, their functions, and mechanisms used by these pathogens to allow them to win over new niches and to evade host defense mechanisms and eventually cause illness.

Second, to provide knowledge of a wide variety of mechanisms: This helps in the prevention and treatment of infectious diseases. These are the mechanism used by researchers for bacterial pathogens to explain pathogenic mechanisms in bacterial diseases that are not yet characterized properly in order to develop new approaches.

A bacterial pathogen is a bacterium that can cause disease, and its ability to cause disease is called pathogenicity. Virulence gives us a likelihood of causing disease, or a quantitative measure of the pathogenicity. Virulence factors are gene products that enable any microorganism to establish itself within a host and enhance its potential to cause disease. Virulence factor comprises of many things like bacterial toxins, cell surface proteins that mediate the bacterial attachment, cell surface carbohydrates and proteins that provide protection to a bacterium, and hydrolytic enzymes that add to the pathogenicity of the bacterium.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

The **release (R1)** is the core dataset of VFDB. It included only the experimentally validated VFs of main medically important bacterial pathogens from 24 genera.
Comparative genomic approaches were introduced in the **release (R2)** of VFDB, it focuses on intra genera comparison of complete genomes in terms of virulence apparatus. Apart from this it also covers some genome predicted Vfs for comparative purposes.
Release(R3) is dedicated to genetic diversity and molecular evolution of bacterial virulence factors .R3 is better than R2 as it performs inter-genera comparative analysis, R3 includes many add on VFs from the plant and animal pathogens, and sometimes, even from non-pathogens in comparison with previous releases.However, both are better extensions of the first release and are both based on the core dataset. Where R3 is VFs centered while R2 is genome centered, i.e both are extending in different directions.
The release (R4) has helped in the improvement of twin aspects of infrastructural datasets of VFDB, these are (i)Removal of redundancy brought by previous releases, and also generated twin hierarchical datasets, one for experimentally verified VFs only and another consisting of full dataset including all the known and predicted VFs.(ii)with the help of controlled vocabulary.
refining of gene annotation of the core dataset
The **release(R5)** to systematically identify known/potential VFs incomplete/draft bacterial genomes, the R5 includes an integrated and automatic pipeline, Vfanalyzer [54-57].

E. DRUGBANK (http://www.drugbank.ca)

It contains richly annotated drug targets and drugs. It contains wide-ranging data related to ontology, nomenclature, chemistry, structure, function, action, pharmacokinetics, pharmacology, metabolism and pharmaceutical properties of molecule drugs both small and large. It also provides detailed information on the organism on which these drugs act, target diseases, proteins, and also genes. After two years of extensive manual annotation works, the database was made useful for wide-ranging
'omics'(i.e.pharamcogenomics, pharamco-proteomics, pharamcometabolomics an even pharamacoeconomics) DrugBank 3.0 was born. A BLAST search against the Drug Bank database, reveals the druggable targets. It can be accessed from http://www.drugbank.ca [30].It provides comprehensive data related to drugs with detailed drug target information.
The Drug Bank database is a fairly detailed, and readily available, online database that contains information on various drugs and drug targets. Drug Bank can combine detailed drug target (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information, as it is both bioinformatics and a cheminformatics resource.Drug Bank is more analogous to an encyclopedia than a drug database, because of its broad scope, comprehensive and detailed referencing and a detailed data description that is unusual in nature. This is why links to DrugBank are maintained for almost all drugs that are listed in Wikipedia. DrugBank has wide uses for different fields like in the drug industry, used by the medicinal chemists, by pharmacists, by physicians, by students and the general public. DrugBanks has an extensive drug and drug-target data that helps in discovering and repurposing of a number of existing drugs not only that it also helps in treating rare and newly identified illnesses.
The new version of DrugBank (version 5.1.4) has been recently released on 02-07-2019 with 13,339 drug entries including :

a. 2,595 approved small molecule drugs
b. 1,289 approved biotech (protein/peptide) drugs
c. 130 nutraceuticals and
d. over 6,304 experimental drug

This new version also contains additional 5,186 non-redundant protein (i.e. drug target/enzyme/
Transporter/carrier) sequences that are linked to these drug entries. Each and every DrugCard entry contains more than 200 data fields.Half of each of these data fields contains information being devoted to chemical/drug data and the other half has information devoted to drug target or protein data [58-62]
If a BLAST search is done against the DrugBank database, it reveals all the druggable targets. It can be readily accessed from the following link: http://www.drugbank.ca [27]. It has detailed and comprehensive data relating to drugs with detailed drug target information [60].

F. CDHIT (http://weizhongli-lab.org/cd-hit/)

The biological sequence data size is rapidly growing due to genome projects and the emerging field of metagenomics [63]. High throughput sequencing technologies are generating a huge amount of sequencing data, which has generated a need for bioinformatics tools for organizing and analyzing data. Mostly, biological sequences are related and can have homology, therefore clustering them in a group and finding consensus sequences can solve many sequence analysis problems.
CD-HIT is a widely used program to cluster and compare large biological sequence datasets.
CD-HIT stands for Cluster Database at High Identity with Tolerance. CDHIT program is to remove paralogous sequences. It clusters and compares large biological sequence datasets. It clusters the protein clusters which meet user-defined similarity threshold, usually sequence identity. Each cluster has one representative sequence. The input is single file in fasta format whereas output ids two files, a fasta file of representative sequence and text files of the list of clusters.
A CD-HIT suite has been launched, for clustering the user uploaded sequence dataset or comparing to another dataset at different identity levels. A user-friendly web interface is provided by CDHIT suite for clustering and comparing with added visualization tools. It also stores precalculated clusters for many public sequence databases. It is also routinely updated.[7]It accepts input in fasta format sequence and output as non-redundant sequences. It only removes redundant sequences and it does so by removing the overall size of the database without removing any

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

sequence information. It can from a given fasta sequence database, produces a set of closely related protein families.

G. CELLO (http://cello.life.nctu.edu.tw/) :

The subcellular localization of a protein is closely related to its biological functions [64].
Valuable information regarding its function can be obtained from the subcellular localization prediction of a protein. A gram-negative bacteria contains five main subcellular localization sites .i.e. the cytoplasm, the periplasm, the inner membrane, the outer membrane, and extracellular space. It has become extremely important to have an automated and accurate tool for subcellular localization prediction due to the increased number of genomic data. Based on n-peptide compositions, CELLO uses the support vector machines trained by multiple feature vectors. There are 1443 proteins within a standard dataset.It enables an overall prediction of 89%, a never reported prediction rate.[37]
CELLO is an SVM Classification system of the multi-class grade. It utilizes four types of sequence coding schemes: the amino acid composition, the dipeptide composition, the partitioned aminoacid composition and based on physicochemical properties of amino acids, sequence composition [65]

H. psortB (https://www.psort.org/psortb/) :

Computational prediction of subcellular localization is important as it provides information regarding its function. PsortB v1.1 has a bacterial localization tool of the greatest precision. It cannot predict for the gram-positive bacteria but only the gram-negative ones. Thereafter modification was made in its v.2.0. This modified version has a precision of 96%for both grams positive and negative bacteria. PSORTb v.3.0 can return the associated probability value for each and a list of five localization sites. Above the cut-off value of 7.5, which is considered good,single localization can be assigned. Prediction of protein to be on the cell surface is of particular interest because such proteins can be primary drug and vaccine targets. A protein's subcellular localization gets influenced by several features that are present within the protein's primary structure like membrane-spanning alpha-helices, or the presence of signal peptide [38,66].
PSORTb 3.0 continues to be the most accurate subcellular localization predictor, and it has increased recall and predictive coverage. It is the most flexible prediction software for prokaryotes, with both online web server and standalone software. This allows them to be incorporated into any bioinformatics pipeline. It has added the predictive capability of SCL prediction of archaeal protein, prediction for bacteria with typical cell morphologies and addition of new predictive subcategories, which represents first SCL predictor designed to handle a diverse range of all prokaryotes and handle prokaryotic localizations [67]

I. BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) :

BLAST can search for regions of similarity between biological sequences. It calculates statistical significance and can compare nucleotide or protein sequences to a sequence database. Its algorithm is quite simple and robust, so much so that, it can be implemented in a number of ways and in varied contexts that extend to include protein sequence database searches, direct DNA, motif and gene identification searches With the help of heuristics it can produce faster results. It also has the capability to estimate how many matches have occurred a given score by chance, by calculating an 'expect value' which eventually helps in judging the amount of confidence to have in the alignment. BLAST tool has been used to find essential sequences, sequences non-homologous to human and gut flora. BLAST against the DRUG BANK database yields druggable targets. BLAST can be utilized to identify members of gene families and can be used to infer functional and evolutionary relationships between sequences.
The parameters to be considered in analyzing BLAST output are sequence identity, E-value and Bit score. The percentage of matches of the same amino acid residues between two aligned sequences is termed as sequence identity and for a protein sequence to be significant, the identity should be greater than 35%. E-value tells if the sequence matches are purely by chance, hence, lower the E-value, more significant is the sequence match. Zero or negative value of e-value tells if it is significant or not. Another statistical indicator is the Bit score. Higher the bit score, more significant the match is. The bit score measures sequence similarity independent of query sequence length and database size [100].
The blast has four different types :
The BLASTN: nucleotide-nucleotide search looks for more distant sequences.
The BLASTP: type can do a protein-protein sequence comparison, and its algorithm is the basis of many others.
The BLASTX: can search a nucleotide query that can be searched against a protein database, translating the query on the fly.
The TBLASTN: can search a protein query against a nucleotide database, thereby translating the database on the fly [27].

J. TID (http://bmicnip.in/TiD/) :

TID is a standalone software for drug target identification. TID depends on the supposition that protein must be vital for the survival of the pathogen and must be non-homologous to its host for it to be called as a putative drug target. TID removes paralogous sequences, it can select the important ones and eliminate the proteins which are homologous to host organisms. It can perform non-homology analysis on gut flora, to identify targets which will not harm the harmless gut flora.Interactome analysis, chokepoint analysis, pathway analysis, functional annotation, and subcellular localization prediction can be done using TID. It takes less than two hours to find putative drug targets from bacterial proteome with approx. 5000 proteins [39]. Due to these qualities, it becomes, a sought after and a very beneficial tool that is available at http://bmicnip.in/TiD/. Hence it is a useful tool for rational drug design.it is available at http://bmicnip.in/TiD/.
s

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

K. MGENOME SUBTRACTOR (http://202.120.12.134/mGS2/) :

The MGENOME SUBTRACTOR is a tool based on the web for analyzing multiple bacterial genomes for a parallel in silico subtractive hybridization.
It does a mpi BLAST- which is based in insilico "subtractive hybridization" and generates a list of conserved fragments by matching selected closely related genomes. This can provide clues to the specific environmental adaptation, phenotype, or the bacterial disease. The mgenome subtractor is capable of running rapid BLAST searches of the multiple subject genomes at DNA or amino acid level against the segmented reference genome, that too within few minutes mainly because of its parallel computing architecture.
The MGENOME SUBTRACTOR provides a very flexible and sliding window-based genome fragmentation approach that can be used to identify short unique sequences within or between the genes, it can also compare protein-coding sequences within them. It enables searching or exploring identified core and accessory regions, virulence factors or bacterial essential genes, including searches against databases of mobile genetic elements, dinucleotide distribution bias, examination of G+C content and an integrated primer design tool, by providing powerful schematic output for it.mGenome Subtractor can, on the basis of available genomes provides the ready definition of species-specific gene pools. It's a very efficient oligonucleotide design tool, helps in the development of Pan-genomic arrays very easily. This oligonucleotide design tool is a very simple high throughput *insilico* 'subtractive hybridization' analytical tool. The tool supports the ever-increasing number of studies on comparative bacterial genomics that intend to define genomic biomarkers of an evolutionary lineage, their pathotype and phenotype, their environmental adaptation and/or disease-association of different types of bacterial species [40].One does not need to login into the mGenomeSubtractor for usage plus its also available for free at http://202.120.12.134/mGS2/

L. TMHMM: (http://www.cbs.dtu.dk/services/TMHMM/)

TMHMM is based on Hidden Markov Model and is used for the prediction of transmembrane helices in protein. It can forecast97-98% of transmembrane helices accurately. It can also distinguish between the membrane and soluble proteins specifically with a rate of 99%. It also functions with high accuracy that helps researchers predict in the genome the integral membrane protein. On the basis of this, the researcher can estimate that around 20-30% of all genes in all genomes encode a membrane protein. It predicts transmembrane helices and discriminates between soluble and membrane proteins with a high degree of accuracy [41].
Users can submit as many as 4000 protein sequences in FASTA format each time.TMHMM is available at. http://www.cbs.dtu.dk/services/TMHMM/

V. THE ORGANISM IN WHOSE SUBTRACTIVE PROTEOMICS HAS BEEN APPLIED FOR DRUG TARGET IDENTIFICATION SINCE 2008

VI.

Subtractive proteomics has been used in a large number of organisms, Here we are only mentioning those which have been published after 2008. Subtractive proteomics has been used in two types of organisms: multidrug-resistant organisms and those for whom no effective drug is available.

S.No	organism	No. of Drug targets	References
1.	*Acinetobacterbaumanii*	13	[68]
2.	*Mycoplasma hypopneumoniae*	42	[69]
3.	*Vibrio cholera*	16	[70]
4.	*Mycobacterium abscessus*	40	[71]
5.	*Staphylococcus aureus*	12	[72]
6.	*Haemophilusducreyi*	3	[73]
7.	*E.coli*	44	[74]
8.	*Shigellaflexneri*	11	[75]
9.	*Listeria monocytogenes*	46	[76]
10.	*Corneybacteriumdiptheriae*	3	[77]
11.	*Bacillus anthracis*	45	[78]
12.	*Fusobacteriumnucleatum*	35	[79]
13.	*Streptococcus pneumonia strain JJA*	2	[80]
14.	*Salmonella enterica subsp. Entericaserovar*	11	[81]
15.	*Haemophilusinfluenzae*	9	[82]
16.	*Salmonella typhi Ty2*	20	[83]
17.	*Brucellamelitensis*	16	[84]
18.	*Trepanomapallidium*	6	[85]
19.	*Rickettsia rickettsii*	9	[86]
20.	*Mycobacterium tuberculosis*	2	[87]
21.	*Mycoplasma pneumonia*	27	[88]
22.	*Mycobacterium leprae*	8	[89]
23.	*Neisseria gonorrhoeae*	20	[90]
24.	*Neisseria Meningitides Serogroup B*	35	[91]
25.	*Clostridium perfringens SM101*	5	[92]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

26.	*Borreliaburgdorferi ZS7*	15	[93]
27.	*Helicobacter pylori*	10	[94]
28.	*Leptospirainterrogans*	78	[95]
29.	*Klebsiella pneumonia*	105	[96]

Table 2: Bacterial Pathogens to Which In Silico Subtractive proteomics strategy has been applied to identify Drug Targets

VII. AUTHORS INSIGHT ON THE TOPIC

It is a fast and cost-effective method of target discovery. In this method, proteins nonhomologous to host and gut flora is found, which helps in finding those targets that are exclusive to the pathogen. Inhibiting such targets by drugs will help avoid the toxicity issues and will further enable a cost reduction of ADMET evaluation of newly found drugs [97].

It shortens up the time for immune-informatics based epitope prediction, thereby speeding up the time for peptide design [98]. This method just requires proteome data for a start and keeps shortening data at each step. It is a low-cost method and tools mostly used are freely available. Identification of essential proteins in pathogen can be validated via mutagenesis studies [97]

Although this method has several advantages, it also has some disadvantages. This method requires the entire genome and proteome sequence and cannot be used for those organisms whose genome is not sequenced. Targets found using this method requires experimental validation. Essential genes are found using DEG BLAST, this database is continuously updated. With respect to time, the consistency of the number of screened essential genes for a given pathogen is a very noticeable concern. Due to data enrichment of the DEG, this number increases dramatically[98]. Also, we remove proteins <100 amino acid from analysis but some essential proteins are less than 100 amino acids, so we inevitably miss few novel targets. This cannot be true always. The expression of close gene changes, giving us a false positive result related to the essentiality of protein, within a mutagenesis experiment on the insertion mutation, done in a nucleotide sequence of 300bp length. A protein that is essential but not homologous to DEG database can be missed. The above method is based on BLAST and therefore one needs to be careful while interpreting results otherwise a conditional essential protein may be screened and selected during the process. Therefore there is a need to develop a tool that is independent of DEG. One of the findings during the process was that the number of predicted targets gets reduced if we increase the number of different strains within the same genus of the pathogen by using multiple hosts. It is also advised to use all strain-specific hosts in the analysis and use multiple strains of pathogens to identify common targets for broad host range and for all strains as well.

VIII. CONCLUSIONS AND FUTURE PROSPECTS

Insilico subtractive proteomics is a fast, cost-effective and reliable method for drug target discovery, provided information of genome and proteome of host and pathogen is known. The drug target found by this method needs experimental validation. Multiple analysis is done in this method, requiring BLAST.

In order to standardize the method, different BLAST parameters need optimization. It becomes necessary to develop an efficient integrated platform to complete the whole analysis at the same time. and also develop a tool that is independent of DEG(database of essential genes).In order to improve the effectiveness of the original, there is a need for a new approach called ‘insilico mutagenesis’ and a need to develop a computational validation method.

LIST OF ABBREVIATIONS

CD-HIT: Cluster database at high identity with tolerance
DEG: Database of essential genes
BLAST: Basic Local alignment search tool
KAAS: KEGG Automated Annotation Server
KEGG: Kyoto Encyclopedia of Gene and Genome
VFDB: Virulence factor database
UniProt: Universal protein knowledgebase

CONSENT FOR PUBLICATION

Not Applicable

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGMENTS

The author would like to thank DBT INDIA for fellowship to SM. The authors are grateful to the research and development committee of the Integral University for support.

REFERENCES

[1]Miesel L, Greene J, Black TA. Microbial genetics: Genetic strategies for antibacterial drug discovery. Nature Reviews Genetics. 2003 Jun;4(6):442.
[2] Plotkin SA. Why certain vaccines have been delayed or not developed at all. Health Affairs. 2005 May;24(3):631-4.
[3] Huynen M, Dandekar T, Bork P. Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS letters. 1998 Apr 10;426(1):1-5.
[4] Ji Y. The role of genomics in the discovery of novel targets for antibiotic therapy. Pharmacogenomics. 2002 May 1;3(3):315-23.
[5] Pucci MJ. Use of genomics to select antibacterial targets. Biochemical pharmacology. 2006 Mar 30;71(7):1066-72.
[6] Holman AG, Davis PJ, Foster JM, Carlow CK, Kumar S. Computational prediction of essential genes in an unculturableendosymbiotic bacterium, Wolbachia of Brugiamalayi. BMC microbiology. 2009 Dec;9(1):243.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

[7] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010 Jan 6;26(5):680-2.
[8] Sarangi AN, Lohani M, Aggarwal R. Proteome mining for drug target identification in Listeria monocytogenes strain EGD-e and structure-based virtual screening of a candidate drug target penicillin binding protein 4. Journal of microbiological methods. 2015 Apr 1;111:9-18.
[9] Rahman MA, Noore MS, Hasan MA, Ullah MR, Rahman MH, Hossain MA, Ali Y, Islam MS. Identification of potential drug targets by subtractive genome analysis of Bacillus anthracis A0248: An in silico approach. Computational biology and chemistry. 2014 Oct 1;52:66-72.
[10] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Oct 11;28(23):3150-2.
[11] Hasan MA, Khan MA, Sharmin T, Mazumder MH, Chowdhury AS. Identification of putative drug targets in Vancomycin-resistant Staphylococcus aureus (VRSA) using computer aided protein data analysis. Gene. 2016 Jan 1;575(1):132-43.
[12] Mondal SI, Ferdous S, Jewel NA, Akter A, Mahmud Z, Islam MM, Afrin T, Karim N. Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157: H7: an in silico approach. Advances and applications in bioinformatics and chemistry: AABC. 2015;8:49.
[13] Gupta SK, Sarita S, Gupta MK, Pant KK, Seth PK. Definition of potential targets in Mycoplasma Pneumoniae through subtractive genome analysis. J AntivirAntiretrovir. 2010;2(2):38-41.
[14] Haag NL, Velk KK, Wu C. In silico identification of drug targets in methicillin/multidrug-resistant Staphylococcusaureus. Int. J. Adv. Life Sci. 2012;4(1-2):21-32.
[15] Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proceedings of the National Academy of Sciences. 1996 Sep 17;93(19):10268-73.
[16] Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome research. 2002 Jun 1;12(6):962-8.
[17] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F. Essential Bacillus subtilis genes. Proceedings of the National Academy of Sciences. 2003 Apr 15;100(8):4678-83.
[18] Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic acids research. 2004 Jan 1;32(suppl_1):D271-2.
[19] Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic acids research. 2013 Nov 15;42(D1):D574-80.
[20] Amineni U, Pradhan D, Marisetty H. In silico identification of common putative drug targets in Leptospirainterrogans. Journal of chemical biology. 2010 Oct 1;3(4):165-73.
[21] Butt AM, Tahir S, Nasrullah I, Idrees M, Lu J, Tong Y. Mycoplasma genitalium: a comparative genomics study of metabolic pathways for the identification of drug and vaccine targets. Infection, Genetics and Evolution. 2012 Jan 1;12(1):53-62.
[22] Hossain M, Chowdhury DU, Farhana J, Akbar MT, Chakraborty A, Islam S, Mannan A. Identification of potential targets in Staphylococcus aureus N315 using computer aided protein data analysis. Bioinformation. 2013;9(4):187.
[23] Kerfeld CA, Scott KM. Using BLAST to teach "E-value-tionary" concepts. PLoS biology. 2011 Feb 1;9(2):e1001014.
[24] Azam SS, Shamim A. An insight into the exploration of druggable genome of Streptococcus gordonii for the identification of novel therapeutic candidates. Genomics. 2014 Sep 1;104(3):203-14.
[25] Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. Expert review of anti-infective therapy. 2010 Apr 1;8(4):435-54.
[26] Rabizadeh S, Sears C. New horizons for the infectious diseases specialist: how gut microflora promote health and disease. Current infectious disease reports. 2008 Mar 1;10(2):92-8.
[27] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990 Oct 5;215(3):403-10.
[28] Raman K, Yeturu K, Chandra N. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC systems biology. 2008 Dec;2(1):109.
[29] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research. 2007 Jul 1;35(suppl_2):W182-5.
[30] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2010 Nov 8;39(suppl_1):D1035-41.
[31] Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. VFDB: a reference database for bacterial virulence factors. Nucleic acids research. 2005 Jan 1;33(suppl_1):D325-8.
[32] Barh D, Tiwari S, Jain N, Ali A, Santos AR, Misra AN, Azevedo V, Kumar A. In silico subtractive genomics for target identification in human bacterial pathogens. Drug Development Research. 2011 Mar;72(2):162-77.
[33] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ. UniProt: the universal protein knowledgebase. Nucleic acids research. 2004 Jan 1;32(suppl_1):D115-9.
[34] Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins: Structure, Function, and Bioinformatics. 2006 Aug 15;64(3):643-51.
[35] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010 May 13;26(13):1608-15.
[36] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999. doi:10.1093/nar/27.1.29.
[37] Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein science. 2004 May;13(5):1402-6.
[38] Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS. PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics. 2004 Oct 22;21(5):617-23.
[39] Gupta R, Pradhan D, Jain AK, Rai CS. TiD: Standalone software for mining putative drug targets from bacterial proteome. Genomics. 2017 Jan 1;109(1):51-7.
[40] Shao Y, He X, Harrison EM, Tai C, Ou HY, Rajakumar K, Deng Z. mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes. Nucleic acids research. 2010 Apr 30;38(suppl_2):W194-200.
[41] Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology. 2001 Jan 19;305(3):567-80.
[42] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F. Essential Bacillus subtilis genes. Proceedings of the National Academy of Sciences. 2003 Apr 15;100(8):4678-83.
[43] Itaya M. An estimation of minimal genome size required for life. FEBS letters. 1995 Apr 10;362(3):257-60.
[44] Judson N, Mekalanos JJ. TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. Nature biotechnology. 2000 Jul;18(7):740.
[45] Weinberg Z, Perreault J, Meyer MM, Breaker RR. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. Nature. 2009 Dec;462(7273):656.
[46] Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic acids research. 2008 Oct 30;37(suppl_1):D455-8.
[47] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997 Sep 1;25(17):3389-402.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

[48] Holman AG, Davis PJ, Foster JM, Carlow CK, Kumar S. Computational prediction of essential genes in an unculturableendosymbiotic bacterium, Wolbachia of Brugiamalayi. BMC microbiology. 2009 Dec;9(1):243.

[49]Xu P, Ge X, Chen L, Wang X, Dou Y, Xu JZ, Patel JR, Stone V, Trinh M, Evans K, Kitten T. Genome-wide essential gene identification in Streptococcus sanguinis. Scientific reports. 2011 Oct 20;1:125.

[50] Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, Hu LT. Identification of essential genes of the periodontal pathogen Porphyromonasgingivalis. BMC genomics. 2012 Dec;13(1):578.

[51] Juhas M, Stark M, von Mering C, Lumjiaktase P, Crook DW, Valvano MA, Eberl L. High confidence prediction of essential genes in Burkholderiacenocepacia. PloS one. 2012 Jun 29;7(6):e40064.

[52] Goto S, Nishioka T, Kanehisa M. LIGAND: chemical database for enzyme reactions. Bioinformatics (Oxford, England). 1998 Jan 1;14(7):591-9.

[53] Goto S, Nishioka T, Kanehisa M. LIGAND database for enzymes, compounds and reactions. Nucleic acids research. 1999 Jan 1;27(1):377-9.

[54] Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic acids research. 2018 Nov 5;47(D1):D687-92.

[55] Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic acids research. 2015 Nov 17;44(D1):D694-7.

[56] Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic acids research. 2011 Nov 8;40(D1):D641-5.

[57] Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic acids research. 2007 Nov 4;36(suppl_1):D539-42.

[58] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research. 2006 Jan 1;34(suppl_1):D668-72.

[59] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research. 2007 Nov 29;36(suppl_1):D901-6.

[60] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2010 Nov 8;39(suppl_1):D1035-41.

[61] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A. DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research. 2013 Nov 6;42(D1):D1091-7.

[62] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research. 2017 Nov 8;46(D1):D1074-82.

[63] Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS biology. 2007 Mar 13;5(3):e16.

[64] Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Stærfeldt HH, Rapacki K, Workman C, Andersen CA. Prediction of human protein function from post-translational modifications and localization features. Journal of molecular biology. 2002 Jun 21;319(5):1257-65.

[65] Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein science. 2004 May;13(5):1402-6.

[66]Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS. PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics. 2004 Oct 22;21(5):617-23.

[67] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010 May 13;26(13):1608-15.

[68] Solanki V, Tiwari V. Subtractive proteomics to identify novel drug targets and reverse vaccinology for the development of chimeric vaccine against Acinetobacterbaumannii. Scientific reports. 2018 Jun 13;8(1):9044.

[69] Damte D, Suh JW, Lee SJ, Yohannes SB, Hossain MA, Park SC. Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of Mycoplasma hyopneumoniae. Genomics. 2013 Jul 1;102(1):47-56.

[70] Chawley P, Samal HB, Prava J, Suar M, Mahapatra RK. Comparative genomics study for identification of drug and vaccine targets in Vibrio cholerae: MurA ligase as a case study. Genomics. 2014 Jan 1;103(1):83-93.

[71] Shanmugham B, Pan A. Identification and Characterization of Potential Therapeutic Candidates in Emerging Human Pathogen Mycobacterium abscessus : A Novel Hierarchical In Silico Approach 2013;8. doi:10.1371/journal.pone.0059126.

[72] ZAVERI K, PATNALA K. SCREENING OF PUTATIVE THERAPEUTIC CANDIDATES IN SUPERBUG (STAPHYLOCOCCUS AUREUS): A SYSTEMATIC IN SILICO APPROACH. SCREENING. 2016;9:2.

[73] de Sarom A, Kumar Jaiswal A, Tiwari S, de Castro Oliveira L, Barh D, Azevedo V, Jose Oliveira C, de Castro Soares S. Putative vaccine candidates and drug targets identified by reverse vaccinology and subtractive genomics approaches to control Haemophilusducreyi, the causative agent of chancroid. Journal of The Royal Society Interface. 2018 May 23;15(142):20180032.

[74]Mondal SI, Ferdous S, Jewel NA, Akter A, Mahmud Z, Islam MM, Afrin T, Karim N. Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157: H7: an in silico approach. Advances and applications in bioinformatics and chemistry: AABC. 2015;8:49.

[75] Hossain MU, Khan M, Hashem A, Islam M, Morshed MN, Keya CA, Salimullah M. Finding potential therapeutic targets against Shigellaflexneri through proteome exploration. Frontiers in microbiology. 2016 Nov 22;7:1817.

[76] Hossain M, Mosnaz AT, Sajib AM, Roy PK, Shakil SK, Ullah SM, Prodhan SH. Identification of putative drug targets of Listeria monocytogenes F2365 by subtractive genomics approach. Journal of BioScience& Biotechnology. 2013 Jan 1;2(1).

[77] Khalid Z, Ahmad S, Raza S, Azam SS. Subtractive proteomics revealed plausible drug candidates in the proteome of multi-drug resistant Corynebacteriumdiphtheriae. Meta Gene. 2018 Sep 1;17:34-42.

[78] Rahman MA, Noore MS, Hasan MA, Ullah MR, Rahman MH, Hossain MA, Ali Y, Islam MS. Identification of potential drug targets by subtractive genome analysis of Bacillus anthracis A0248: An in silico approach. Computational biology and chemistry. 2014 Oct 1;52:66-72.

[79]Kumar A, Thotakura PL, Tiwary BK, Krishna R. Target identification in Fusobacteriumnucleatum by subtractive genomics approach and enrichment analysis of host-pathogen protein-protein interactions. BMC microbiology. 2016 Dec;16(1):84.

[80] Wadood A, Jamal A, Riaz M, Khan A, Uddin R, Jelani M, Azam SS. Subtractive genome analysis for in silico identification and characterization of novel drug targets in Streptococcus pneumonia strain JJA. Microbial pathogenesis. 2018 Feb 1;115:194-8.

[81] Hossain T, Kamruzzaman M, Choudhury TZ, Mahmood HN, Nabi AH, Hosen M. Application of the subtractive genomics and molecular docking analysis for the identification of novel putative drug targets against Salmonella enterica subsp. entericaserovar Poona. BioMed research international. 2017;2017.

[82] N NR. Identification of Novel Therapeutic Drug Targets By Subtractive Genomics Approach In HaemophilusInfluenzae 2012;5:4466–8.

[83] Batool N, Waqar M, Batool S. Comparative genomics study for identification of putative drug targets in Salmonella typhi Ty2. Gene. 2016 Jan 15;576(1):544-59.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

[84] Pradeepkiran JA, Sainath SB, Kumar KK, Bhaskar M. Complete genome-wide screening and subtractive genomic approach revealed new virulence factors, potential drug targets against bio-war pathogen Brucellamelitensis 16M. Drug design, development and therapy. 2015;9:1691.

[85] Pradeepkiran JA, Sainath SB, Kumar KK, Bhaskar M. Complete genome-wide screening and subtractive genomic approach revealed new virulence factors, potential drug targets against bio-war pathogen Brucellamelitensis 16M. Drug design, development and therapy. 2015;9:1691.

[86] Maurya PK, Singh S, Mani A. Comparative genomic analysis of Rickettsia rickettsii for identification of drug and vaccine targets: tolC as a proposed candidate for case study. Acta tropica. 2018 Jun 1;182:100-10.

[87] Uddin R, Siddiqui QN, Azam SS, Saima B, Wadood A. Identification and characterization of potential druggable targets among hypothetical proteins of extensively drug resistant Mycobacterium tuberculosis (XDR KZN 605) through subtractive genomics approach. European Journal of Pharmaceutical Sciences. 2018 Mar 1;114:13-23.

[88] Gupta SK, Sarita S, Gupta MK, Pant KK, Seth PK. Definition of potential targets in Mycoplasma Pneumoniae through subtractive genome analysis. J AntivirAntiretrovir. 2010;2(2):38-41.

[89] Shanmugam A, Natarajan J. Computational genome analyses of metabolic enzymes in Mycobacterium leprae for drug target identification. Bioinformation. 2010;4(9):392.

[90] Barh D, Kumar A. In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae. In silico biology. 2009 Jan 1;9(4):225-31.

[91] Sarangi AN, Aggarwal R, Rahman Q, Trivedi N. Subtractive genomics approach for in silico identification and characterization of novel drug targets in Neisseria Meningitides Serogroup B. J Comput SciSyst Biol. 2009;2(5):255-8.

[92] Chhabra G, Sharma P, Anant A, Deshmukh S, Kaushik H, Gopal K, Srivastava N, Sharma N, Garg LC. Identification and modeling of a drug target for Clostridium perfringens SM101. Bioinformation. 2010;4(7):278.

[93] Madagi S, Patil VM, Sadegh S, Singh AK, Garwal B, Banerjee A, Talambedu U, Bhattacharjee B. Identification of membrane associated drug targets in Borreliaburgdorferi ZS7-subtractive genomics approach. Bioinformation. 2011;6(9):356.

[94] Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D. In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. In silico biology. 2006 Jan 1;6(1, 2):43-7.

[95] Amineni U, Pradhan D, Marisetty H. In silico identification of common putative drug targets in Leptospirainterrogans. Journal of chemical biology. 2010 Oct 1;3(4):165-73.

[96]Georrge JJ, Umrania V. In silico identification of putative drug targets in Klebsiella pneumonia MGH78578.

[97] Sakharkar KR, Sakharkar MK, Chow VT. A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. In silico biology. 2004 Jan 1;4(3):355-60.

[98] Barh D, Misra AN. In silico identification of membrane associated candidate drug targets in Neisseria gonorrhoeae. Int J Integr Biol. 2009;6:65-7.

[99] Pratheek J Madabhavi, V G Shanmugapriya, Rakesh N R, Preeti S Honagudi, SurekhaJiddagi. SUBTRACTIVE GENOMICS – A Promising way To Combat Pathogens (A Review). *International Research Journal of Engineering and Technology (IRJET)* 2015; 02(03): .

[100] Preeti Dora , V.G. ShanmugaPriya , Preeti.S.H , U.M. Muddapur , Rakesh.N.R. IMPORTANT DATABASES AND TOOLS TO IDENTIFY PROMISING DRUG TARGETS BY SUBTRACTIVE GENOMICS APPROACH – A REVIEW. International Journal of Research in Engineering and Technology 2015; 04(06):