International Journal of Genetics and Genomics
2014; 2(4): 62-67
Published online August 20, 2014 (http://www.sciencepublishinggroup.com/j/ijgg)
doi: 10.11648/j.ijgg.20140204.13

In silico identification of novel candidate drug targets in *Haemophilus influenzae* Rd KW20

Ranjith Kumavath[1, *], Swaraj Prasad[1], Pratap Devarapalli[1, 2], Debmalya Barh[3]

[1]Department of Genomic Sciences, School of Biological Sciences, Central University of Kerala, P.O. Central University, Kasaragod-671314, India

[2]Genomics & Molecular Medicine Unit, Institute of Genomics and Integrative Biology Council of Scientific and Industrial Research, Mathura Road, New Delhi-110025, INDIA

[3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, PurbaMedinipur, West Bengal-721172, INDIA

Email address:

RNKumavath@gmail.com (R. N. Kumavath), ranjith_kumavath@yahoo.com (R. N. Kumavath)

To cite this article:

Ranjith Kumavath, Swaraj Prasad, Pratap Devarapalli, Debmalya Barh. *In Silico* Identification of Novel Candidate Drug Targets in *Haemophilus Influenzae* Rd KW20. *International Journal of Genetics and Genomics.* Vol. 2, No. 4, 2014, pp. 62-67.
doi: 10.11648/j.ijgg.20140204.13

Abstract: Background: Globally, respiratory diseases cause an estimated 1.9 million deaths per year. One of the most important aetiological organisms of both adult and childhood respiratory disease is Non-Typeable *Haemophilus influenzae* (NTHi). NTHi is frequently isolated from the respiratory tract during episodes of sinusitis, Otitis Media and pneumonia and is the most common cause of Chronic Obstructive Pulmonary Disease (COPD) and bronchiectasis exacerbations. Methods: The work has been effectively complemented with the compilation of the Database of Essential Genes (DEG) of *H. influenzae* Rd KW20. Each protein is subjected to BLASTP in NCBI server *http://www.ncbi.nlm.nih.gov/blastp.* The candidate drug targets are determined by KAAS (KEGG Automatic Annotation Server), KEGG ORTHOLOGY and KEGG GENES. Results: The given gram negative bacteria *H. influenzae* Rd KW20 has six distinguished domains i.e., cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane, extracellular and unknown domains. Out of 642 essential genes, the predicted non-human Homologous are 412 in the different domain of given bacteria. With the help of KAAS (KEGG Automatic Annotation Server), KEGG ORTHOLOGY and KEGG GENES, we successfully identified 35 novel drug targets which have common metabolic pathways both in host and pathogen & pathogen specific metabolic pathways. Conclusion: The novel drug targets suggest those genes which are active in both the host and pathogen metabolic pathway and should be a pathogen specific metabolic pathway. The important drug target regions are vacJ, lepB, emrB, MurG & dgkA. vacJ is present in outer membrane while lepB, emrB, MurG&dgkA are present in cytoplasmic membrane. All these genes are fully sequenced and specifically annotated in KEGG PATHWAY.

Keywords: NTHi (Non-Typeable *H. Influenzae*), NP (Non-Typeable) , DEG (Database of Essential Genes), Rd (Rough Derivative), ChoP (Phosphorylcholine), COPD (Chronic Obstructive Pulmonary Disease), SVM (Support Vector Machine), SCL (Subcellular Localization)

1. Introduction

Haemophilus influenzae Rd KW20 strain is an avirulent laboratory strain of *H. influenzae* which lacks adhesins commonly found in disease causing NTHi (Non-Typeable *H. influenzae*) [1]. Respiratory tract infections associated with Non-Typeable *H. influenzae* are major causes of morbidity and mortality in both developed and non-industrialized nations. The vaccine for *H. influenzae* type b is directed against its type-specific polysaccharide capsule. It has no ability to prevent infections caused by the non-encapsulated NTHi. Lower respiratory tract infections, associated with NTHi are major causes of mortality in both infants and children in developing countries. Sometimes it affects the innate mucosal immune system, such as Chronic Obstructive Pulmonary Disease (COPD) and cystic fibrosis. The prevalence of Otitis Media during the first 3 years of life has enormous effects on intellectual ability, school achievement, speech, and language [2].There are several main virulence factors in *H. influenzae* such as: capsule, fimbrialadhesin, HMW1 and HMW2, Hap

Adhesin, Hia and Hsf Adhesins, Haemocin, IgA protease, Lipooligosaccharide, Outer Membrane Proteins (OMPs). Capsule: The type b capsule is Polyribosyl-Ribitol-Phosphate (PRP) which is composed of linear teichoic acid containing ribose, ribitol - an alcohol containing a five carbon sugar and a phosphate linked by phosphodiester bonds and is a critical determinant of virulence [3]. Fimbrial adhesion: Fimbriae facilitate adherence to human cells by binding to glycoproteins and glycolipids present on the respiratory mucin proteins. Fimbriae composed of another protein called P5-fimbrin, have been identified on Non-Typeable *H. influenzae* strains [4]. HMW1 and HMW2 adhesins: The function of HMW1 (160 kDa) and HMW2 (155 kDa) is adherence to host epithelial cells. These adhesive proteins are present in almost 80% of Non-Typeable *H. influenzae* but are absent from typeable strains. Hap adhesin: Hap belongs to the auto-transporter family of proteins and is synthesized as a precursor protein with 3 functional domains, including an N-terminal signal sequence, an internal 110-kDa domain called Haps, and a C-terminal 45-kDa domain called Hapβ. Hapsharbors the adhesive activity responsible for bacterial interaction with epithelial cells and for bacterial aggregation [5]. Hia and hsf adhesins: Hsf shares significant homology with Hia, a 1,098-amino-acid auto-transporter protein that is present in ~25% of NTHi strains and Mediates efficient attachment to human epithelial cells. Haemocin: This small heat-stable protein is a type of bacteriocin produced by over 90% of type b *H. influenzae* strain and its function is to inhibit the growth of other bacteria belonging to the same or similar species in the site of the infection. HMC-producing strains of *H. influenzae* can invade the cells much earlier than the HMC-deficient isogenic mutants [6]. IgA protease: The IgA1 protease of *H. influenzae* is encoded by two genes, the iga, present in most *H. influenzae* strains, and the igaB gene, that is present in one-third of *H. influenzae* strains. Strains containing both genes have been correlated with significantly higher levels of IgA1 protease activity as compared to strains containing only the igagene. Lipooligosaccharides: Non-Typeable *H. influenzae* attacks the host cell by binding to PAF (platelet-activating factor) receptor via their LOS glycoforms that contains phosphorylcholine (ChoP). For its ability to attract opsonization and phagocytosis, *H. influenzae* LOS has also been investigated as a potential antigen for Non-Typeable *H. influenzae* vaccine development. Outer membrane proteins (OMPS): More than 36 different OMPs from *H. influenzae* have been isolated and characterized. The first proteins that were described were in the order of decreasing molecular weight. These are considered to be major OMPs and include P1, P2, and P4-P6. Other proteins such as the transferrin binding protein 1, 2 (Tbp1/Tbp2) and protein D belong to the minor OMPs of *H. influenzae*. Protein 1 (P1): P1 is a heat-modifiable protein of 35-50 kDa that accounts for approximately 10% of the OMP content.P1 is highly immunogenic and has been shown to induce protective antibodies against NTHi-induced Otitis Media in chinchillas. Protein 2 (P2): P2 is a porin that accounts for 50% [7],of the OMP content and is thus the most abundant outer membrane protein of *H. influenzae*. During the course of an infection, specific regions of P2 vary at high frequency, in particular a surface exposed loop which undergoes mutations at high rate. Protein 4 (P4): P4 is a lipoprotein that is highly conserved among NTHi and Hib strains, and has been shown to be important for the bacterial growth. P4 has been found to be one component in the heme-acquisition pathway uptake in *H. influenzae*. Protein 6 (P6): P6 triggers high release of IL-8 and TNF-α in macrophages. In respiratory epithelial cells, P6 upregulates the mucin gene transcription leading to overproduction of mucin, which is a hallmark of diseases such as COPD. Transferrin binding proteins 1 and 2 (Tbp1 and Tbp2): The transferrin-binding protein 1and 2 (Tbp1 and Tbp2) are the most important ones since the lack of either protein severely impairs the bacterial growth [8]. Tbp2 serves as surface receptor for transferrin and Tbp2 is responsible of the transferrin transport. Protein D (PD): Due to the properties of PD, such as surface localization, high degree of antigenic conservation, wide distribution, pathogenicity, and promising preclinical trials, it was decided to use PD as antigenically active carrier protein in a new 11-valent pneumococcal conjugate vaccine. In a randomized double-blind efficacy study, it was shown that by using PD as a carrier protein for pneumococcal polysaccarides, the vaccine induced protection both against pneumococcal Otitis and acute Otitis Media due to NTHi [9].

2. Materials &Methods

2.1. Genomes, Databases and BLAST Parameters for Selection of Essential Genes

Identification of bacterial genes that are non-homologous to human genes and important for the survival of bacteria is one of the promising means to identify novel drug targets. The target should be essential for growth and viability of the organism, should provide selectivity, and should yield a drug which is highly selective against pathogen with respect to human host. The work has been effectively complemented with the compilation of the Database of Essential Genes (DEG) of H .influenzae Rd KW20.The DEG is a freely available database which can be accessed by http://tubic.tju.edu.cn/deg. The DEG which is used here has 6.8 version and updated on November, 2011. Each functional gene and corresponding protein sequence of the given bacteria was subjected to standard BLASTP, respectively against DEG. We have set some parameters as Expected threshold value, matrix, max target sequences and mask are considered to be 0.00001, BLOSUM45, 10 & lower case letters.

2.2. Identification of Subcellular localization of H. Influenzae Rd KW20

Subcellular localization plays a key role to elucidate the functions of a protein. Therefore, proteins that cooperate towards a common biological function are located in the same subcellular compartment. Prokaryotes (Gram-negative

bacteria) have 5 major subcellular localizations (outer membrane, cytoplasmic membrane, periplasmic, cytoplasm, and extracellular) specialized in distinct biochemical process. Prediction of protein localization is important to identify the surface membrane proteins which could be feasible vaccine target. Subcellular localization analysis of the essential protein sequences has been done by PSORTb v3.0 server (http:/db.psort.org/). PSORTdb is a database of SCL for bacteria that contains both information determined through laboratory experimentation (ePSORTdb dataset) and computational predictions (cPSORTdb dataset). PSLpred has been made to develop a SVM based method for the prediction of subcellular localization of prokaryotic proteins. Amino acid composition based SVM module can predict cytoplasmic, extracellular, inner-membrane, outer-membrane, periplasmic localization with 87%, 78%, 87%, 94% & 80% accuracy respectively. It has been assessed with reliability index, the higher the percentage of accuracy, the higher RI value is predicted. Combined value of reliability index and accuracy gives the accurate subcelllular localization of the given protein. CELLO is a multi-class SVM classification system. CELLO uses 4 types of sequence coding schemes: the amino acid composition, the di-peptide composition, the partitioned amino acid composition and the sequence composition based on the physico-chemical properties of amino acids.

2.3. Selection of Non-Human Homologous Essential Genes

Non-human homologous can eradicate possibilities of cross contamination that might be harmful to the human host. The subtractive genomics approach is subtractive because we focus on the complement of the genome of the pathogen that is essential for the viability of the pathogen but is not present in the human. Each protein is subjected to BLASTP in NCBI server *http://www.ncbi.nlm.nih.gov/blastp.* Sequences which don't give any hit against Human BLASTP server were selected as non-human homologous essential genes. Non-human Homologous essential genes are important for identifying drug targets because non-human Homologous won't interfere in the pathway of human as it solely works only for respective pathogens.

2.4. Identification of Novel Drug Targets

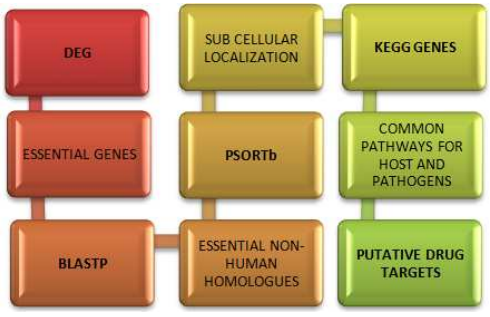

Fig. 1*. Pictorial flow chart presentation of in silico identification of novel drug target*

List of selected membrane proteins from various pathways were prepared based on essential non-human homologs those are involved in pathways common in both the host and the pathogen and from pathogen specific metabolic pathways by using the KEGG database. The candidate drug targets are determined by KAAS (KEGG Automatic Annotation Server), KEGG ORTHOLOGY and KEGG GENES. KAAS, KEGG ORTHOLOGY & KEGGGENES can be accessed by online server *http://www.genome.jp/tools/kaas,http://www.genome.jp/kegg/ko.html & http://www.genome.jp/kegg/genes.html.* The KEGG GENES database provides a single resource for cross-species annotation of all available genomes by a standardized mechanism, called the KEGG Orthology (KO) system. The essence of the KO system is that it is a pathway based definition of orthologous genes. The KO entry represents an ortholog group that is linked to a box (gene product) in the KEGG pathway diagram. Thus, once the KO identifiers (K numbers) are assigned to genes in the genome organism-specific pathways can be computationally generated.

3. Results

3.1. Prediction of Sub Cellular Localization of Essential Genes in H. Influenzae Rd KW20

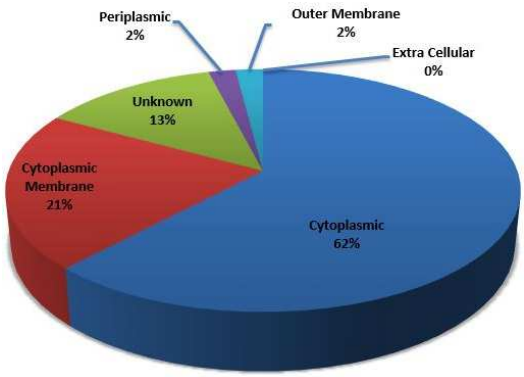

Fig. 2. *Subcellular localization of H. influenzae Rd KW20. (Cytoplasm contains 62% of total available area. Remaining subcellular localized area constitutes of 21% of cytoplasmic membrane, 13% of unknown region, 2% of periplasmic & outer membrane while extracellular subcellular localized area is negligible).*

With the help of PSORTbv3.0 and PSLpred SCL prediction tools, we are able to predict the accurate SCL of proteins in *H. influenzae* Rd KW20. The given gram negative bacteria *H. influenzae* Rd KW20 has six distinguished domains i.e., cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane, extracellular and unknown domains. The highest number of essential genes are present in cytoplasmic region (62%). The other regions are as; cytoplasmic membrane (21%), Unknown region (13%), Periplasmic (2%), Outer Membrane (2%) and Extracellular (0%) (Fig. 2) Table 1 confirms the number of essential genes in all the subcellular localized regions of the given bacteria. The highest no. of essential genes present in the cytoplasmic region (396) while

there is only one essential gene present in extracellular region. The cytoplasmic membrane contains 135 essential genes, unknown region contains 83 essential genes, periplasmic region contains 16 essential genes and outer membrane contains essential 11 genes.

3.2. Essential Non-Human Homologous in H. Influenzae Rd KW20

The genes which have taken for studies is solely based on DEG database of *H. influenzae* Rd KW20. The objective of the work was to find and locate those essential genes of *H. influenzae* Rd KW20 that play important roles in the normal functioning of the bacterium within the host and to shortlist them in the view of drug targeting. Out of 642 essential genes, the predicted non-human Homologous are 412(Table 1) in the different domain of given bacteria. The highest percentage of non-human Homologous genes is in cytoplasmic region (51%) (Fig.3) while other regions are as; cytoplasmic membrane (25%), unknown region (17%), periplasmic (4%), outer membrane (3%) and extracellular (1%).

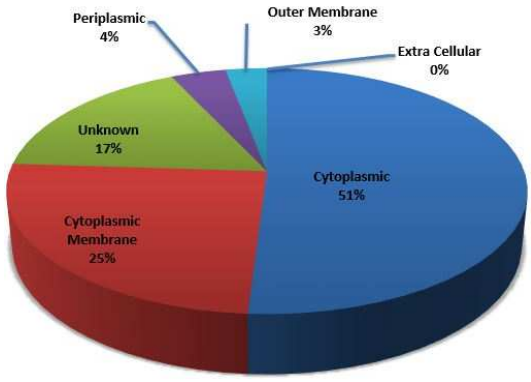

Fig. 3. *Non-human Homologous genes in the respective domains of H. influenzae Rd KW20.51%of the cytoplasmic domain constitutes of non-homologous genes while 25% of the cytoplasmic membrane constitutes of non-homologous genes. 17% of the unknown region, 4% of periplasmic, 3% of the outer membrane and negligible amount of extracellular constitutes non-homologous genes.*

3.3. Identification of Drug Targets in Subcellular Localized Domains of H. Influenzae Rd KW20

With the help of KAAS (KEGG Automatic Annotation Server), KEGG ORTHOLOGY and KEGG GENES, we are able to identify 35 (Table 1) novel drug targets which have common metabolic pathways both in host and pathogen & pathogen specific metabolic pathways. Since many genes in the unknown regions are putative and hypothetical proteins. So, we have taken only the cytoplasmic, cytoplasmic membrane, periplasmic and outer membrane where cytoplasmic region has the maximum number of drug targets (24 out of 35). The percentage analysis tells that abundant drug target regions are laid in the cytoplasmic (68%) (Fig.4) while cytoplasmic membrane contains 26% and outer membrane contains 6% of drug target regions.

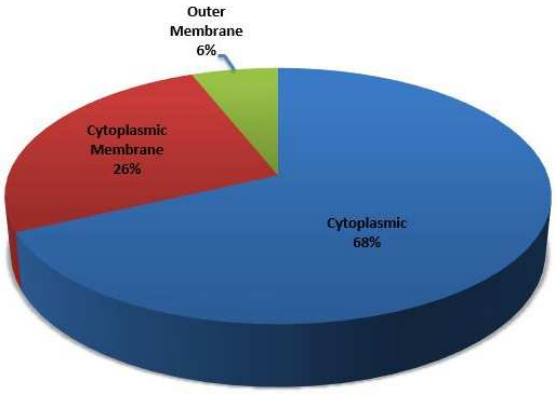

Fig. 4. *Drug target regions in the above specified localizations in H. influenzae Rd KW20.In the above specified pie chart it shows that cytoplasmic region contains 68% of drug target regions while cytoplasmic membrane and outer membrane contains 26% and 6% drug target regions. Others don't possess significant drug target regions.*

Table 1. *Number of essential genes, non-human Homologous genes and drug targets in specified domains of H. influenzae Rd KW20.*

Domains	Essential Genes	Non-Human Homologous Genes	Drug Targets
Cytoplasmic	396	210	24
Cytoplasmic Membrane	135	105	09
Unknown	83	70	00
Periplasmic	16	15	00
Outer Membrane	11	11	02
Extracellular	01	01	00
Total	642	412	35

4. Discussions

An essential gene can be defined as a gene without which a cell is unable to survive and whose deletion or disruption results in the death of the organism. For this reason, essential genes could furnish novel drug targets for the therapy of bacterial infections. As the consideration of novel drug targets, is about 5% of genes are important for identifying novel drug targets. *H. influenzae* Rd KW20 is the virulent form of *H. influenzae* NTHi which has lost the virulence characters due to lack of adhensins. We have considered here Rd instead of NP because the database of essential genes (DEG) doesn't contain database related to *H. influenzae* 86-028NP (a virulent nontypeable form of *H. influenzae*). So, it is better to predict the non-human homologous of essential genes in *H. influenzae* Rd KW20 instead of NP which contains whole genome of non-typeable strain. The most severe form of NTHi infection is Otitis Media. There are different strains available for specific Otitis Media disease. For COME (Chronic Otitis Media with effusion) the causative NTHi strain is *H. influenzae* PittEE and AOM (Acute Otitis Media) the causative NTHi strain is *H. influenzae* PittGG which are isolated from the external ear discharge of a spontaneously perforated tympanic membrane of a child in Pittsburgh who had been diagnosed with Otorrhea. So, if anyone wants to

discover drug target region in specific Otitis Media, anyone can go for specific strain PittEE and PittGG which contain 1613 and 1661 protein coding genes. There are different localization tools are available for predicting subcellular localization. It depends upon which analytical method we are using for predicting subcellular localization. We have used PSORTb, CELLO and PSLpred where PSROTb uses multi-component analytical method while CELLO and PSLpred uses SVM analytical method. There are several other prediction tools which use multi-component, annotation keywords and SVM. Proteome Analyst uses Annotation Keywords as analytical method and SubLoc, LOCtree and P-CLASSIFIER use SVM as analytical method. However, none of the system is perfect. So, analyzing subcellular localization is a tough task. Instead of using three, verify genes with all the prediction tools [10].The process of BLASTP is quite lengthy and analyzing each sequence manually will take a prolong time to assess the data for essential non-human Homologous. Cd-hit is a useful tool for analyzing millions of database at a very short span of time. Several new programs using the same algorithm including cd-hit-2d, cd-hit-est and cd-hit-est-2d. Cd-hit-2d compares two protein datasets and reports similar matches between them; cd-hit-est clusters a DNA/RNA sequence database and cd-hit-est-2d compares two nucleotide datasets. All these programs can handle huge datasets with millions of sequences and can be hundreds of times faster than methods based on the popular sequence comparison and database search tools, such as BLAST [11].

5. Conclusion

Membrane associated drug targets [12], are considered as potential drug targets which are the important binding site for bacteria. The important membrane associated drug targets are vacJ, rec2, dgkA, emrB, murG,etc.rec2 is a recombination protein which is important for genetic recombination and recombinational repair, while dgkA is an enzyme responsible for binding DAG & ATP-binding sites. emrB is a multidrug resistance protein which is important for forming novel antibiotic [13].*MurG* is an essential bacterial glycosyltransferase that is involved in the biosynthesis of peptidoglycan. The enzyme is found in all organisms that synthesize peptidoglycan and is a target for the design of new antibiotics [14]. vacJ is an important membrane associated protein found in outer membrane of *H. influenzae*. Cloning and sequencing of the *vacJ* region Indicated that the *vacJ* gene encoded a 28.0 kDa protein possessing a signal peptide at the N-terminus which contained the motif characteristic of lipoproteins. The analysis of the *vacJ* product indicated that VacJ was exposed on the bacterial surface. The *vacJ* gene was distributed among shigellae and enteroinvasive *Escherichia coli*, and the constructed *vacJ* mutants failed to spread intercellularly, indicating that *vacJ* is a chromosomal gene essential for the pathogenicity of shigellae [15]. Although experimental and computational methods have been previously employed for the study of essential genes to our knowledge, this is the first report of essential gene identification as probable drug targets in *H. influenzae.*

References

[1] Martin K, Morlin G, Smith A, Nordyke A, Eisenstark A, Golomb M. The tryptophanase gene cluster of Haemophilus influenzae type b: evidence for horizontal gene transfer. J Bacteriol 1998; 80:107–18.

[2] Teele DW, Klein JO, Chase C, Menyuk P, Rosner BA, the Greater Boston Otitis Media Study Group. Otitis Media in infancy and intellectual ability, school achievement, speech, and language at age 7 years. J Infect Dis 1990; 162:685–94.

[3] Sukupolvi SP, Grass S, Geme St III JW. The *H. influenzae* Type b hcsA and hcsB gene products facilitate transport of capsular polysaccharide across the outer membrane and are essential for virulence. J Bacteriol 2006; 188: 3870-7.

[4] Van SMH, Van LA, Mooi FR, Van JP. Contribution of the major and minor subunits to fimbria-Mediated adherence of *H. influenzae* to human epithelial cells and erythrocytes. Infect Immun 1995; 63: 4883-9.

[5] Hendrixson DR, Geme III JWS. Haemophilus influenzae Hap serine protease promotes adherence and microcolony formation, potentiated by a soluble host protein. Mol Cell 1998; 2: 841-50.

[6] Murley YM, Edlind TD, Plett PA, LiPuma JJ. Cloning of the haemocin locus of *H. influenzae* type b and assessment of the role of Haemocin in virulence. Microbiology 1998; 144:2531-8.

[7] Duim B, Vogel L, Puijk W, Jansen HM, Meloen RH, Dankert J, et. al. Fine mapping of outer membrane protein P2 antigenic sites which vary 55 during persistent infection by Haemophilus influenzae. Infection and immunity 1996; 64:4673-79.

[8] Gray-Owen SD, Loosmore S, Schryvers AB. Identification andcharacterization of genes encoding the human transferrin-binding proteins from Haemophilus influenzae. Infection and immunity 1995; 63:1201-10

[9] Prymula R, Peeters P, Chrobok V, Kriz P, Novakova E, Kaliskova E, et al. Pneumococcal capsular polysaccharides conjugated to protein D for prevention of acute Otitis Media caused by both Streptococcus pneumonia and non-typable *H. influenzae*: a randomised double-blind efficacy study. Lancet 2006; 367:740-8.annotation and pathway reconstruction server. Nucleic Acids Research 2007; 35:W182–5.

[10] Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization. Nature Publishing Group 2006; 4:741-51.

[11] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22(13):1658-9.

[12] Barh, D.&Misra, N.A. In silico identification of membrane associated candidate drug targets in Neisseria gonorrhoeae. IJIB 2009; 6(2):65-7.

[13] Haung J, O'Tole WP, Shen W, Amrine-Madsen H, Jiang X, Lobo N, et al. Novel chromosomally encoded multidrug efflux transporter mdea in Staphylococcus aureus. Antimicrob Agents Chemother 2004; 48(3):909–17.

[14] Ha S, Gross B, Walker S. E. Coli MurG: A Paradigm for a Superfamily of Glycosyltransferases. Current Drug Targets Infectious Disorders 2001; 1(2):201-13.

[15] Suzuki T, Murai T, Fukuda I, Tobe T, Yoshikawa M, Sasakawa C. Identification and characterization of a chromosomal virulence gene, vacJ, required for intercellular spreading of Shigella flexneri. Molecular Microbiology 1994; 11 (1):31-41.