Bangladesh J Microbiol, Volume 35, Number 2, December 2018, pp 102-107

Original Article

Potential Drug Target Identification of *Legionella pneumophila* by Subtractive Genome Analysis: An *In Silico* Approach

Md. Sadikur Rahman Shuvo[1], Shahriar Kabir Shakil [2], Firoz Ahmed[1*]

[1]Department of Microbiology, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.
[2]Department of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.

Though *Legionella pneumophila* is an opportunistic pathogen, recent reports about multi drug resistance in *L. pneumophila* is alarming. Annotated whole genome provides a pool of information which is applied for therapeutic drug targets identification in pathogenic bacteria. Subtractive genomic analysis is a pragmatic approach to screen the essential proteins present in pathogen but absent in host. Phylogenetically closely related *L. pneumophila* str. Philadelphia and *L. pneumophila* str. ATCC43209 protein profiles were analyzed to identify putative drug targets. Paralogous duplicate profiles were primarily discarded using CD-hit suit. Six hundred and ninety one *L. pneumophila* str. Philadelphia and 690 *L. pneumophila* str. ATCC43209 human homologous proteins were excluded using blstP. Among the human non-homologous proteins, the essential proteins for bacteria were separated using DEG tool. For both strains, one hundred and nineteen essential proteins were marked which participate in various metabolic pathways. Among them 11 unique proteins were found. Beside it, 15 and 16 exposed surface proteins were present in strain Philadelphia and ATCC43209 respectively. These unique and cell surface proteins can be utilized for effective drug and vaccine targets.

Keywords: Phylogeny, Essential genes, Drug target, Unique pathway

Introduction

The Legionellaceae are Gram-negative bacteria found in aquatic environments all over the world. Usually they are intracellular parasites of free-living protozoa. They are also distributed in manmade water systems where they survived freely in biofilms. The family Legionellaceae consists of a single genus, *Legionella*. More specifically, this genus includes the species *L. pneumophila*, which are non-encapsulated, aerobic bacilli. *L. pneumophila* is an opportunistic pathogen that causes infections in immunocompromised individuals. The bacterium is most notable as the causative agent of Legionnaires' disease, a potentially fatal pneumonia[1].

This organism was identified in 1976 in Philadelphia when an outbreak of a serious pneumonia occurred in individuals attending an annual convention. Approximately 200 people developed pneumonia within the first few days after the convention and about 2 dozen succumbed to respiratory failure[2]. This outbreak of pneumonia in the hotel was initially suspected to cause by toxic substances or some other environmental problems.

After rigorous investigations from Centers for Disease Control and Prevention (CDC), a unique bacterium was identified several months later and thought to be a new microbe[3-4]. The organism was named Legionella because the disease was identified first in those attending the Legionnaire's convention. Some previous outbreaks of pneumonia had become evident caused few decades back [5].

By subculturing into a rich artificial medium, *L. pneumophila*, was first isolated by inoculation of postmortem lung tissue into guinea pigs[4]. By indirect immunofluorescent antibody assay, it was discovered that a number of unexplained respiratory disease were associated with seroconversion to *L. pneumophila*, a "rickettsia-like" organism, isolated by guinea pig inoculation from the blood of a feverish patient in 1947, which today is recorded as the earliest known isolate of L. pneumophila[6].

In the moderately brief time frame, *L. pneumophila* was first distinguished as a human pathogen, in excess of 50 types of Legionella have been perceived, and no less than 24 of these have been related with human disease. It is conceivable that under the fitting conditions, immunocompromised individuals can be tainted with any types of Legionella. The incredible larger part of Legionnaires' ailment, around 90%, is caused by *L. pneumophila*, and notwithstanding the portrayal of somewhere around 15 serogroups, *L. pneumophila* serogroup 1 is in charge of over 84% of cases around the world[7-9].

It has been accounted for that *L. pneumophila* is getting to be impervious to specific anti-infection agents, for example, erythromycin, rifampicin and quinolones derivatives [10-13]. The development of medication obstruction of *L. pneumophila* has prompted the look for novel medication targets. The accessibility of finish genome successions of *L. pneumophila* strain Philadelphia has cleared the better approach to recognize the

Corresponding author:
Dr. Firoz Ahmed, Associate Professor, Department of Microbiology, Noakhali Science & Technology University, Mobile: +8801713195658, Email: firoz19701016@gmail.com.

novel medication targets. The goal of present work was to distinguish putative medication focuses in *L. pneumophila* strain Philadelphia through metabolic pathway examination, to play out the homology displaying and to play out the atomic elements reenactment of a candidate drug target.

Materials and Methods:

Sequences of methodology are summarized in Figure 1.

Phylogenetic analysis:

Phylogenetic analysis of *L. pneumophila* str. Philadelphia (Accession: NC_002942) and *L. pneumophila* str. ATCC43209 (Accession: NC_016811) was performed by PanX phylogeny[14] against 85 other *L. pneumophila* sequences available in *National Center for Biotechnology Information (NCBI). Gene parameters of these two strains were analyzed.*

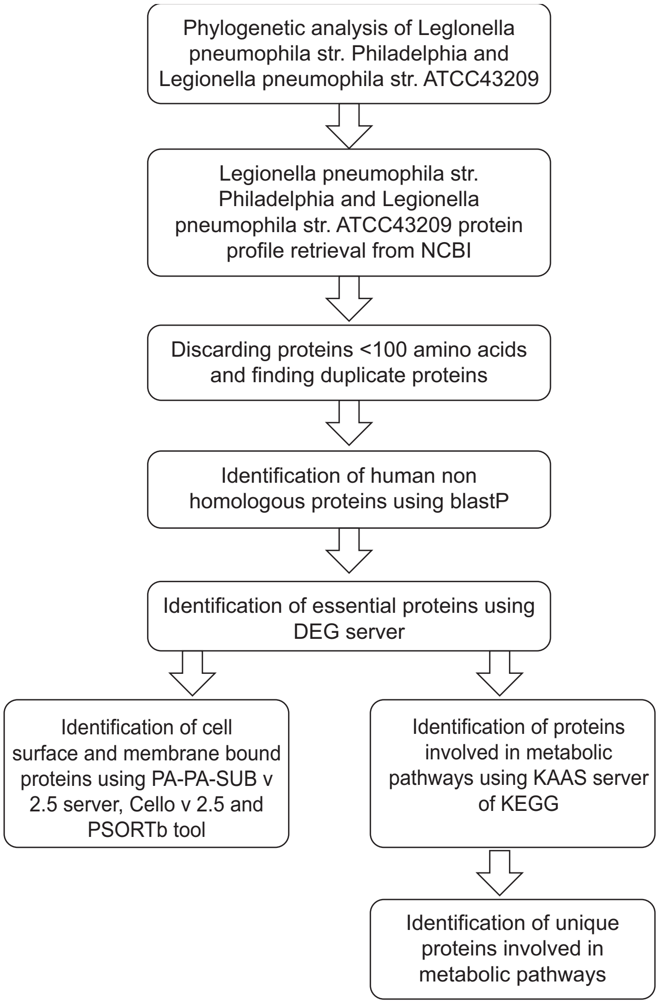
Phylogenetic analysis of Leglonella pneumophila str. Philadelphia and Legionella pneumophila str. ATCC43209
Legionella pneumophila str. Philadelphia and Legionella pneumophila str. ATCC43209 protein profile retrieval from NCBI
Discarding proteins <100 amino acids and finding duplicate proteins
Identification of human non homologous proteins using blastP
Identification of essential proteins using DEG server
Identification of cell surface and membrane bound proteins using PA-PA-SUB v 2.5 server, Cello v 2.5 and PSORTb tool
Identification of proteins involved in metabolic pathways using KAAS server of KEGG
Identification of unique proteins involved in metabolic pathways

Figure 1. *Schemetic diagram of methodology.*

Retrieval of protein sequences:

Complete information of *L. pneumophila* str. Philadelphia (Accession: NC_002942) and *L. pneumophila* str. ATCC43209 (Accession: NC_016811) was retrieved from *National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) as GenBank file format. All protein sequences were retrieved from the GenBank files using Bio-Python[15] module and saved as fasta file format for the following assays. Fasta comments were trimmed from the multiple fasta files in such a way that they contain only the accession number of respective protein sequences for efficient tagging and easy recognition.*

Paralog proteins identification:

The duplicate protein sequences of *L. pneumophila* str. Philadelphia and str. ATCC43209 were removed from the master fasta files. The multiple fasta files were subjected to CD-hit suit web server[16] which uses sequence identity at cut off value 0.60%[17]. After the duplicate proteins were separated from the master fasta file, it only contained non-paralog proteins. The sequences were then screened to have at least 100 amino acids in their chains. The protein sequences which had less than 100 amino acids in their chains were excluded from the master database.

Human non homolog protein identification:

The non paralog proteins found in the previous stage were then subjected to NCBI blastP against *Homo sapiens.* The purpose was to find out the proteins those are homolog to the human genome. At this stage the threshold expectation value was 0.0001. The proteins those showed similarity at this threshold value were removed from the database. The rest of the proteins remained in the database are human non-homologous proteins which will be ultimately used for the identification of drug target in the subsequent stages.

Search for essential proteins in the database:

Database of essential genes (DEG)[18], an web server, providing the facility of essential protein identification, was used to find the essential genes in *L. pneumophila* str. Philadelphia and str. ATCC43209. In this case maximum threshold value was 10^{-100} and bit score cut off value was 100. Output from this analysis gives the essential human non-homologous essential proteins of *L. pneumophila* str. Philadelphia and str. ATCC43209. The essential proteins were then categorized based on the metabolic activity by blastKoala[19].

Metabolic pathway analysis of essential proteins:

KEGG Automatic Annotation Server KASS[20] (https://www.genome.jp/tools/kaas/) was used for the analysis of metabolic pathways of essential proteins. This is extremely helpful to identify the unique targets. In this analysis functional annotation of genes are performed where manually arranged KEGG gene database is used for template search. The output result comprises of KEGG Ontology assignments and KEGG pathways.

Unique pathway analysis:

KEGG Genome Database was used for the identification of unique metabolic pathway prevailing in *L. pneumophila* str. Philadelphia and str. ATCC43209. The metabolic pathways of *Homo sapiens* and the two pathogenic strains of *L. pneumophila* were compared. The pathway map generated in this step represents the unique metabolic machineries in *L. pneumophila* str. Philadelphia and str. ATCC43209.

Identification of cell surface proteins:

Function of a protein can easily be portended if location of it is predicted. Bacterial surface proteins are always good objective for drug and vaccine.

PA-SUB server[21] (Proteome Analyst Specialized Subcellular Localization Server v2.5) (http://webdocs.cs.ualberta.ca/~bioinfo/PA/Sub/) was used for the prediction of particular localization of the necessary proteins identified in the previous stages. Results were verified by doing the same analysis in two other web tools, PSORTb[22] and CELLO v2.5[23]

Result and Discussion:

Prime purpose of this study is to look into the effective drug target by systematic subtraction of genes from genome. In this study we worked with two different strains of *L. pneumophila.* At first their evolutionary relationship was verified with other 85 strains of *L. pneumophila* by whole genome phylogenetic analysis. The phylogenetic study (Figure 2) shows that the two strains considered in this study have high similarities with most of the other *L. pneumophila* strains*.* Through *in-silico* analysis, such genes were identified those are absent in host *Homo sapiens.* The result of subtractive analytical steps is summarized in Table 1.

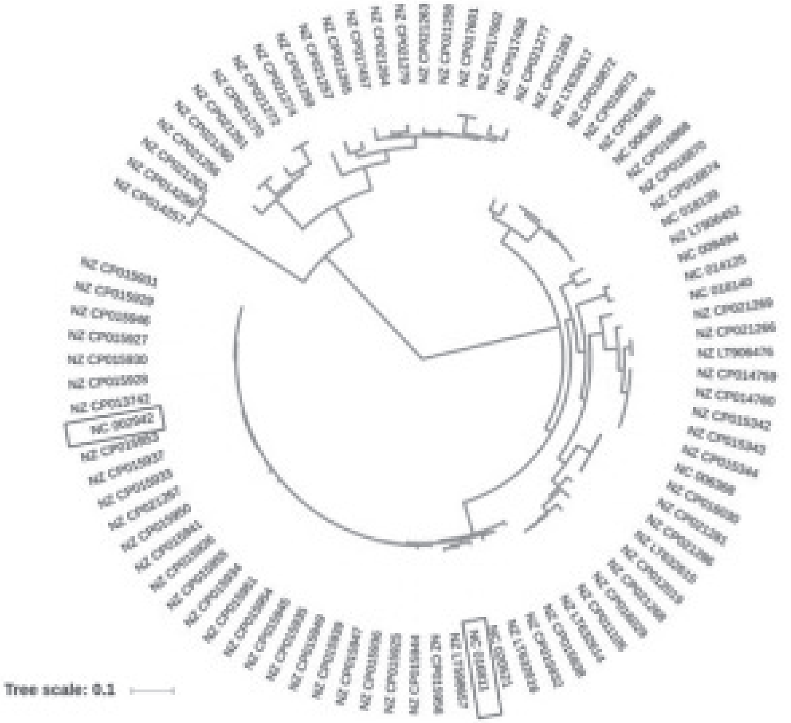

Figure 2. *Whole genome phylogenetic analysis of Legionella pneumophila str. Philadelphia (Accession: NC_002942) and Legionella pneumophila str. ATCC43209 (Accession: NC_016811) against other 85 strains of Legionella pneumophila.*

Table 1. *Subtractive analytical output*

Analytical steps	Total number of proteins	
	Legionella pneumophila str. Philadelphia	*Legionella pneumophila* str. ATCC43209
Retrieved protein	2931	3020
Proteins > 100 amino acids	2744	2774
Non-paralogous proteins	2698	2702
Human homologous proteins	691	690
Human non-homologous proteins	2007	2012
Essential proteins predicted by DEG	301	302
Essential proteins in metabolic pathway	119	119
Cell surface essential proteins	15	16
Proteins involved in unique pathways	11	11

A total of 2931 *L. pneumophila* str. Philadelphia and 3020 *L. pneumophila* str. ATCC43209 protein sequences were retrieved. Less than 100aa sequence length proteins were excluded. After exclusion 2744 *L. pneumophila* str. Philadelphia and 2774 *L. pneumophila* str. ATCC43209 proteins were used for next step. After filtering in CD hit, 2698 *L. pneumophila* str. Philadelphia and 2702 *L. pneumophila* str. ATCC43209 remained. Threshold value for CD-hit was 60%. The blast search result in NCBI showed 691 *L. pneumophila* str. Philadelphia and 690 *L. pneumophila* str. ATCC43209 proteins which are human homologous. Ultimately 2007 *L. pneumophila* str. Philadelphia and 2012 *L. pneumophila* str. ATCC43209 proteins identified which are human non homologs. Evolutionary consequences dictate the sharing of some common genes in host and bacteria. Even they are involved in similar cellular systems. In this case proteins having insignificant similarity were considered as non-homologous proteins.

Database of Essential Genes (DEG) screened 302 *L. pneumophila* str. Philadelphia and 301 *L. pneumophila* str. ATCC43209 proteins which are essential proteins in pathogen. BlastKoala categorized the essential proteins according to their functions (Figure 3). All these proteins are human non-homologous and at the same time they have key role in cellular activities. These proteins were again subjected to KAAS server at KEGG to find which metabolic pathways they are involved. KAAS is a BLAST based analysis tool which analyzes prokaryotic proteome against host proteome. KAAS analysis gave an output of 119 essential proteins which are directly involved in metabolic pathways. This is a vital stage of screening, because the proteins found in this step are involved in major metabolic activities of bacteria. So, targeting these proteins we can design drug which can deactivate one or more metabolic pathways and making the bacteria susceptible to that drug.

By the comparative analysis of the metabolic pathway of host and the pathogen (*L. pneumophila*) in KEGG, both *L. pneumophila* str. Philadelphia and *L. pneumophila* str. ATCC43209 showed five uncommon metabolic pathways have been found which are not present in human. Eleven proteins of *L. pneumophila* str. Philadelphia are involved in these pathways which can be used for drug target Table 2. Same proteins were found in *L. pneumophila* str. ATCC43209 while performed BLAST search against *L. pneumophila* str. Philadelphia. These proteins play vital role for survival and regulatory mechanism in

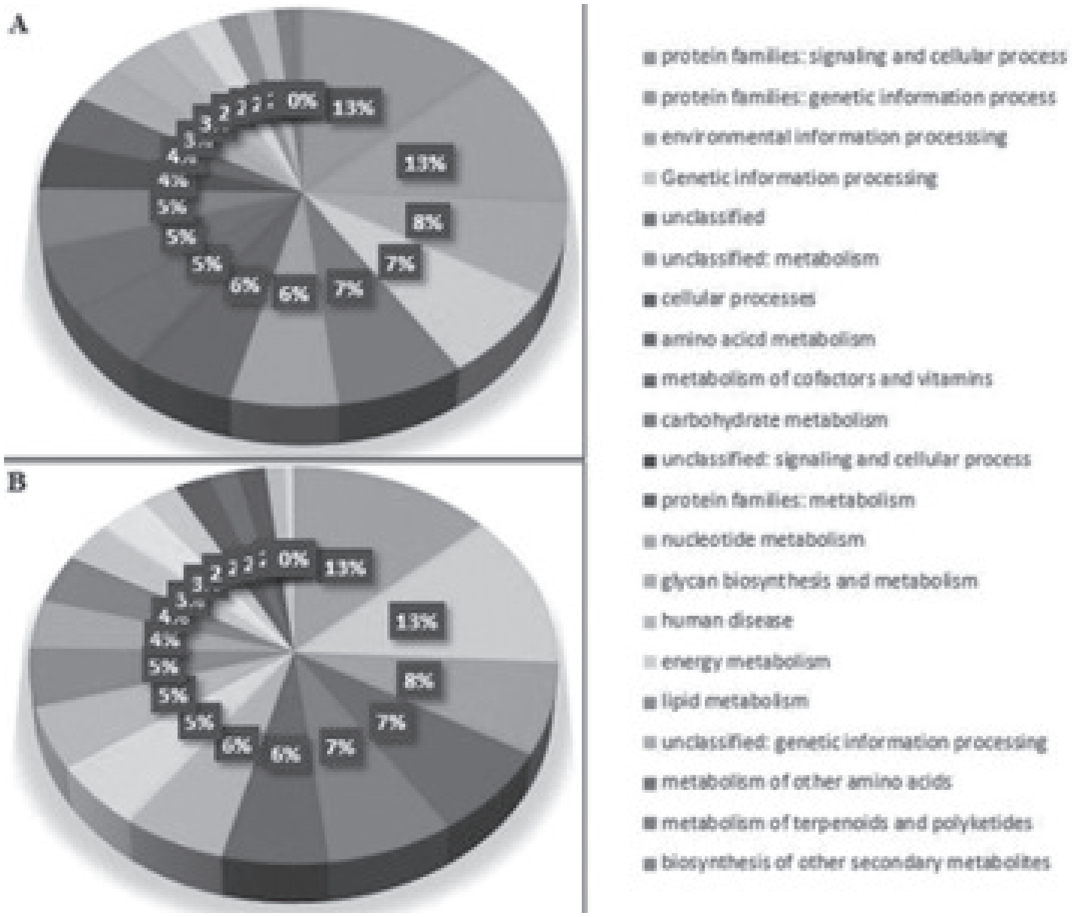

Figure 3. *Categories of essential proteins based on their functions in Legionella pneumophila str. Philadelphia (A) and Legionella pneumophila str. ATCC43209 (B).*

bacteria. Thus these proteins must have high possibility of being good drug target.

Prediction of the position of proteins in bacteria was done by PA-SUB v 2.5 server, CELLO v2.5 and PSORTb server. Those proteins which showed cell surface localization by the entire three prediction tool were taken consideration as cell surface protein. List of the proteins as well as their function is listed in Table 3. Both strains of *L. pneumophila* showed same type of cell surface protein. The hypothetical protein's function was predicted using SVMProt web server[24] based on P value. Among the hypothetical proteins, 2 are zinc binding protein and 2 are lipid binding protein. One of the lipid binding proteins is putative transport protein. All these proteins can be good target for vaccine design to control disease caused by *L. pneumophila.*

Table 2. *List of proteins involved in unique pathways in Legionella pneumophila.*

Protein accession numberof *Legionella pneumophila* str. Philadelphia	Protein name
WP_010946303.1	desC; stearoyl-CoA desaturase
WP_010948619.1	transcription termination factor
WP_010946024.1	MFS transporter, UMF1 family
WP_010948309.1	ftsZ; cell division protein FtsZ
WP_010947297.1	hydroxymethylpyrimidine kinase
WP_010946377.1	ATP-dependent HslUV protease
WP_010946112.1	htrB; Kdo2-lipid IVA lauroyltransferase
WP_010948046.1	3-deoxy-D-manno-octulosonic-acid transferase
WP_010946259.1	UDP-N-acetylglucosamine acyltransferase
WP_010947281.1	UDP-2,3-diacylglucosamine hydrolase
WP_010946499.1	Fuc2NAc and GlcNAc transferase

Table 3. *List of cell surface proteins and their functions.*

Legionella pneumophila str. Philadelphia	Protein name	P value	*Legionella pneumophila* str. ATCC43209	Protein name	P value
WP_011213753.1	Hypothetical protein (lipid binding protein)	0.90	WP_010946366.1	Pilus assembly protein	0.99
WP_011945791.1	Pilus assembly protein	0.99	WP_010946601.1	DUF4156 domain containing protein	0.99
WP_014326896.1	Thaumatin domain containing protein	0.99	WP_010948117.1	Hypothetical protein (lipid binding protein)	0.91
WP_014326828.1	DUF4156 domain containing protein	0.99	WP_010946953.1	Flagellar basal body rod protein FlgG	0.99
WP_010948117.1	Hypothetical protein (Putative transport protein)	0.91	WP_010946952.1	Flagellar basal body rod protein FlgF	0.99
WP_010946953.1	Flagellar basal body rod protein FlgG	0.99	WP_010946896.1	Membrane protein	0.99
WP_010946952.1	Flagellar basal body rod protein FlgF	0.99	WP_010946948.1	Flagellar basal body rod protein FlgB	0.99
WP_010946896.1	Membrane protein	0.99	WP_010947070.1	Flagellin	0.99
WP_010946948.1	Flagellar basal body rod protein FlgB	0.99	WP_010945781.1	Peptidase M4 family protein	0.99
WP_010947070.1	Flagellin	0.99	WP_010947197.1	Hypothetical protein (lipid binding protein)	0.91
WP_010945781.1	Peptidase M4 family protein	0.99	WP_010947059.1	Thaumatin domain containing protein	0.99
WP_010946850.1	Hypothetical protein (Zinc binding protein)	0.92	WP_010946850.1	Hypothetical protein (Zinc binding protein)	0.92
WP_010947068.1	Flagellar hook protein FID	0.99	WP_010947068.1	Flagellar hook protein FID	0.99
WP_010948122.1	Penicillin binding protein	0.99	WP_010948122.1	Penicillin binding protein	0.99
WP_014326669.1	Type I secretion C-terminal target domain containing protein	0.99	WP_010946382.1	Type I secretion C-terminal target domain containing protein	0.99
			WP_010946381.1	Hypothetical protein (Zinc binding protein)	0.92

In this study a subtractive manner was applied for identification of drug target in *L. pneumophila.* The same approach has been applied for several other pathogenic microorganisms[25, 26, 27]. This will be helpful for drug development in the further studies.

Conclusion

In our study, we involved two different strains of *L. pneumophila.* Relatedness of the two strains were verified by phylogenetic analysis. Subtractive genome analysis finally found 11 unique proteins in both strains which are involved in unique metabolic pathways of *L. pneumophila*. These proteins are non-homologous to human genome. The unique proteins can be analyzed by laboratory experimental analysis for drug target in future. We also found 15 and 16 cell surface proteins in Philadelphia and ATCC43209 respectively that will be useful for vaccine target identification.

References

1. Brenner DJ, Steigerwalt AG and McDADE JE. 1979. Classification of the Legionnaires' disease bacterium: *Legionella pneumophila*, genus novum, species nova, of the family Legionellaceae, familia nova. *Ann Intern Med.*, **90**(4): 656-658.
2. Newton HJ, Ang DK, van Driel IR and Hartland EL. 2010. Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clin Microbiol Rev.*, **23**(2): 274-298.
3. Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE and Shepard CC. 1977. Legionnaires' disease: description of an epidemic of pneumonia. *New Engl J Med.*, **297**(22): 1189-1197.
4. McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA, Dowdle WR and Laboratory Investigation Team*. 1977. Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *New Engl J Med.*, **297**(22): 1197-1203.
5. Fang G, Yu VL and Vickers RM. 1989. Disease due to the Legionellaceae (other than *Legionella pneumophila*). Historical, microbiological, clinical, and epidemiological review. *Medicine*, **68**(2): 116-132.
6. Hébert GA, Moss CW, McDougal LK, Bozeman FM, McKinney RM and Brenner DJ. 1980. The rickettsia-like organisms Tatlock (1943) and HEBA (1959): bacteria phenotypically similar to but genetically distinct from *Legionella pneumophila* and the WIGA bacterium. *Ann Intern Med.*, **92**(1): 45-52.
7. Marston BJ, Plouffe JF, File TM, Hackman BA, Salstrom SJ, Lipman HB, Kolczak MS and Breiman RF. 1997. Incidence of community-acquired pneumonia requiring hospitalization: results of a population-based active surveillance study in Ohio. *Arch Intern Med.*, **157**(15): 1709-1718.
8. Muder RR and Victor LY. 2002. Infection due to Legionella species other than L. pneumophila. *Clin Infect Dis.*, **35**(8): 990-998.
9. Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, Summersgill J, File T, Heath CM, Paterson DL and Chereshsky A. 2002. Distribution of Legionella species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. *J Infect Dis.*, **186**(1): 127-128.
10. Moffie BG and Mouton RP. 1988. Sensitivity and resistance of *Legionella pneumophila* to some antibiotics and combinations of antibiotics. *J Antimicrob Chemoth.*, **22**(4): 457-462.
11. Walz A, Nichterlein T and Hof H. 1997. Excellent activity of newer quinolones on *Legionella pneumophila* in J774 macrophages. *Zbl Bakt.*, **285**(3): 431-439.
12. Tsakris A, Alexiou-Daniel S, Souliou E and Antoniadis A. 1999. In-vitro activity of antibiotics against *Legionella pneumophila* isolates from water systems. *J Antimicrob Chemoth.*, **44**(5): 693-695.
13. Fong DH, Lemke CT, Hwang J, Xiong B and Berghuis AM. 2010. Structure of the antibiotic resistance factor spectinomycin phosphotransferase from Legionella pheumophila. *J Biol Chem.*, pp.jbc-M109.
14. Ding W, Baumdicker F and Neher RA. 2017. panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**(1): e5-e5.
15. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and De Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11): 1422-1423.
16. Huang Y, Niu B, Gao Y, Fu L and Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**(5): 680-682.
17. Kumar GS, Sarita S, Kumar GM, Pant KK and Seth PK. 2010. Definition of potential targets in Mycoplasma Pneumoniae through subtractive genome analysis. *J. Antivir. Antiretrovir*, **2**: 038-041.
18. Zhang R, Ou HY and Zhang CT, 2004. DEG: a database of essential genes. *Nucleic Acids Res*, **32**(suppl_1): D271-D272.
19. Kanehisa M, Sato Y and Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.*, **428**(4): 726-731.
20. Moriya Y, Itoh M, Okuda S, Yoshizawa AC and Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**(suppl_2): W182-W185.
21. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C and Eisner R. 2004. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**(4): 547-556.
22. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ and Brinkman FS. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**(13): 1608-1615.
23. Yu CS, Lin CJ and Hwang JK. 2004. Predicting subcellular localization of proteins for Gram negative bacteria by support vector machines based on peptide compositions. *Protein Sci.*, **13**(5): 1402-1406.
24. Cai CZ, Han LY, Ji ZL, Chen X and Chen YZ. 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**(13): 3692-3697.
25. Uddin R and Saeed K. 2014. Identification and characterization of potential drug targets by subtractive genome analyses of methicillin resistant Staphylococcus aureus. *Comput Biol Chem.*, **48**: 55-63.
26. Galperin MY and Koonin EV. 1999. Searching for drug targets in microbial genomes. *Curr Opin Biotech.*, **10**(6): 571-578.
27. Sakharkar KR, Sakharkar MK and Chow VT. 2004. A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. *In sili Biol.*, **4**(3): 355-360.