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

Md. Sadikur Rahman Shuvo¹, Shahriar Kabir Shakil², Firoz Ahmed¹*  

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

**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 pneumonia¹.

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 failure². 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 microbe³-⁴. 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 pigs⁶. 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 world⁷-⁹.

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...
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 phylogeny14
against 85 other L. pneumophila sequences available in National
Center for Biotechnology Information (NCBI). Gene parameters
of these two strains were analyzed.

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. Schematic 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
as GenBank file format. All protein sequences were retrieved
from the GenBank files using Bio-Python15 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 server16 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 blastKoala19.

Metabolic pathway analysis of essential proteins:
KEGG Automatic Annotation Server KASS20 (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.](image)

**Table 1. Subtractive analytical output**

<table>
<thead>
<tr>
<th>Analytical steps</th>
<th>Total number of proteins</th>
<th>Legionella pneumophila</th>
<th>Legionella pneumophila</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved protein</td>
<td></td>
<td>2931</td>
<td>3020</td>
</tr>
<tr>
<td>Proteins &gt; 100 amino acids</td>
<td></td>
<td>2744</td>
<td>2774</td>
</tr>
<tr>
<td>Non-paralogous proteins</td>
<td></td>
<td>2698</td>
<td>2702</td>
</tr>
<tr>
<td>Human homologous proteins</td>
<td></td>
<td>691</td>
<td>690</td>
</tr>
<tr>
<td>Human non-homologous proteins</td>
<td></td>
<td>2007</td>
<td>2012</td>
</tr>
<tr>
<td>Essential proteins predicted by DEG</td>
<td></td>
<td>301</td>
<td>302</td>
</tr>
<tr>
<td>Essential proteins in metabolic pathway</td>
<td></td>
<td>119</td>
<td>119</td>
</tr>
<tr>
<td>Cell surface essential proteins</td>
<td></td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Proteins involved in unique pathways</td>
<td></td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>
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\(^2\) 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*.

<table>
<thead>
<tr>
<th>Protein accession number of  <em>Legionella pneumophila</em> str. Philadelphia</th>
<th>Protein name</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_010946303.1</td>
<td>desC; stearoyl-CoA desaturase</td>
</tr>
<tr>
<td>WP_010948619.1</td>
<td>transcription termination factor</td>
</tr>
<tr>
<td>WP_010946024.1</td>
<td>MFS transporter, UMFI family</td>
</tr>
<tr>
<td>WP_010948309.1</td>
<td>ftsZ; cell division protein FtsZ</td>
</tr>
<tr>
<td>WP_010947297.1</td>
<td>hydroxymethylpyrimidine kinase</td>
</tr>
<tr>
<td>WP_010946377.1</td>
<td>ATP-dependent HslUV protease</td>
</tr>
<tr>
<td>WP_010946112.1</td>
<td>htrB; Kdo2-lipid IVA lauroyltransferase</td>
</tr>
<tr>
<td>WP_010948046.1</td>
<td>3-deoxy-D-manno-octulosonic-acid transferase</td>
</tr>
<tr>
<td>WP_010946259.1</td>
<td>UDP-N-acetylglucosamine acyltransferase</td>
</tr>
<tr>
<td>WP_010947281.1</td>
<td>UDP-2,3-diacylglycosamine hydrolase</td>
</tr>
<tr>
<td>WP_010946499.1</td>
<td>Fuc2NAc and GlcNAc transferase</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th><em>Legionella pneumophila</em> str. Philadelphia</th>
<th>Protein name</th>
<th>P value</th>
<th><em>Legionella pneumophila</em> str. ATCC43209</th>
<th>Protein name</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_011213753.1</td>
<td>Hypothetical protein (lipid binding protein)</td>
<td>0.90</td>
<td>WP_010946366.1</td>
<td>Pilius assembly protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_011945791.1</td>
<td>Pilius assembly protein</td>
<td>0.99</td>
<td>WP_010946601.1</td>
<td>DUF4156  domain containing protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_014326896.1</td>
<td>Thaumatin domain containing protein</td>
<td>0.99</td>
<td>WP_010948117.1</td>
<td>Hypothetical protein (lipid binding protein)</td>
<td>0.91</td>
</tr>
<tr>
<td>WP_014326828.1</td>
<td>DUF4156  domain containing protein</td>
<td>0.99</td>
<td>WP_010946953.1</td>
<td>Flagellar basal body rod protein FlgG</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010948117.1</td>
<td>Hypothetical protein (Putative transport protein)</td>
<td>0.91</td>
<td>WP_010946952.1</td>
<td>Flagellar basal body rod protein FlgG</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010946953.1</td>
<td>Flagellar basal body rod protein FlgG</td>
<td>0.99</td>
<td>WP_010946896.1</td>
<td>Membrane protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010946952.1</td>
<td>Flagellar basal body rod protein FlgG</td>
<td>0.99</td>
<td>WP_010946948.1</td>
<td>Flagellar basal body rod protein FlgG</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010946896.1</td>
<td>Membrane protein</td>
<td>0.99</td>
<td>WP_010947070.1</td>
<td>Flagellin</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010946948.1</td>
<td>Flagellar basal body rod protein FlgB</td>
<td>0.99</td>
<td>WP_010945781.1</td>
<td>Peptidase M4 family protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010947070.1</td>
<td>Flagellin</td>
<td>0.99</td>
<td>WP_010947197.1</td>
<td>Hypothetical protein (lipid binding protein)</td>
<td>0.91</td>
</tr>
<tr>
<td>WP_010945781.1</td>
<td>Peptidase M4 family protein</td>
<td>0.99</td>
<td>WP_010947059.1</td>
<td>Thaumatin domain containing protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010946850.1</td>
<td>Hypothetical protein (Zinc binding protein)</td>
<td>0.92</td>
<td>WP_010946850.1</td>
<td>Hypothetical protein (Zinc binding protein)</td>
<td>0.92</td>
</tr>
<tr>
<td>WP_010947068.1</td>
<td>Flagellar hook protein FID</td>
<td>0.99</td>
<td>WP_010947068.1</td>
<td>Flagellar hook protein FID</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_010948122.1</td>
<td>Penicillin binding protein</td>
<td>0.99</td>
<td>WP_010948122.1</td>
<td>Penicillin binding protein</td>
<td>0.99</td>
</tr>
<tr>
<td>WP_014326669.1</td>
<td>Type I secretion C-terminal target domain containing protein</td>
<td>0.99</td>
<td>WP_010946382.1</td>
<td>Type I secretion C-terminal target domain containing protein</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>WP_010946381.1</td>
<td>Hypothetical protein (Zinc binding protein)</td>
<td>0.92</td>
</tr>
</tbody>
</table>
In this study a subtractive manner was applied for identification of drug target in \textit{L. pneumophila}. The same approach has been applied for several other pathogenic microorganisms\textsuperscript{25, 26, 27}. This will be helpful for drug development in the further studies.

**Conclusion**

In our study, we involved two different strains of \textit{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 \textit{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**


