*Available online at http://www.journalcra.com*

*International Journal of Current Research*
*Vol. 4, Issue, 07, pp.017-021, July, 2012*

**INTERNATIONAL JOURNAL OF CURRENT RESEARCH**

**ISSN: 0975-833X**

## RESEARCH ARTICLE

# SUBTRACTIVE GENOME ANALYSIS FOR *IN SILICO* IDENTIFICATION AND CHARACTERIZATION OF NOVEL DRUG TARGETS IN *C. trachomatis* STRAIN D/UW-3/Cx

**Khalida Shoukat[1]*, Nadia Rasheed[1] and Mohammad Sajid[2]**

[1]Baqai Institute of Information Technology, Baqai Medical University, Karachi, Pakistan
[2]Department of Biochemistry, Hazara University, Mansehra, Pakistan

**ARTICLE INFO**

***Article History:***
Received 14th April, 2012
Received in revised form
15th May, 2012
Accepted 17th June, 2012
Published online 30th July, 2012

***Key words:***
Subtractive genomics analysis,
*C. trachomatis* strain D/UW-3/Cx,
Essential protein,
Druggablility.

**ABSTRACT**

Studies based on subtractive genomics approach could facilitate the selection, processing and development of strain-specific drugs against various pathogens. The current study based on complete proteome information of *Chlamydia trachomatis* strain D/UW-3/Cx (ocular-urogenital pathogen of human) revealed 623 proteins; which were non-homologous to human genome. Subjecting this set of non homologous proteins against the Database of Essential Genes 203 proteins were screened out as essential proteins of the *C. trachomatis*. Among 203 proteins; around 39 essential proteins were found to be part of membrane of pathogen using PSORT tool at Expasy server. All the non homologous essential genes were characterized for differential metabolic pathways using the KEGG Automated Annotation Server and 182 proteins were found to be involved in major metabolic pathway of bacterium. The 12 hypothetical essential proteins were functionally annotated through SVMProt server. Druggability of each of the identified drug targets was also evaluated by the DrugBank database. Moreover, metabolic pathway analysis of the identified druggable essential proteins also revealed 7 proteins that participate in unique pathways of *C. trachomatis* strain D. Enzymes from Peptidoglycan and Riboflavin biosynthesis were identified as attractive candidates for drug development.

## INTRODUCTION

*Chlamydia trachomatis* is globally accepted as the most common cause of bacterial sexually transmitted infections (STIs) that leads to further complications such as Pelvic Inflammatory Diseases (PIDs), ectopic pregnancies, cervicitis, chronic pelvic pain and infertility (Morre *et al*., 2006). *C. trachomatis*, an obligate human pathogen includes three human biovars: serovars A, B, Ba or C cause ocular blindness or trachoma whereas various strains of serovar D-K are associated with the urogenital infections. Lymphogranuloma venereum (LGV) serovars L1, 2 and 3 are involved in the infection of urogenital lymph nodes and hemorrhagic proctitis (Fredlund *et al*., 2004). *C. trachomatis* strain D/UW-3/Cx is a trachoma biovar, serovar D strain which is a well known intracellular pathogen of human that causes venereal diseases (VD) or STIs and trachoma. The strain has also been reported to be involved in infections such as bronchitis, pharyngitis, and respiratory infection. The complete genome of *C. trachomatis* strain D/UW-3/Cx is accessible at NCBI (National Center for Biotechnology Information) server under the accession number NC_000117.1. The size of the genome is around 1.04 Mbp which codes for total 895 proteins (Stephens *et al*., 1998). According to World Health Organization estimates (1994) more than 90 million new cases of genital *C. trachomatis* infection are reported each year. Due to asymptomatic nature of chlamydial infection most of the people remain unaware of their condition and do not seek treatment unless they reach to a complicated and chronic stage. Though various antibiotics are available for the treatment, yet a better strain specific infection control and management strategy is needed to eradicate the prevalence of the disease. The idea of interpretation of the genome sequences, piling up at various databases and the analysis of the conceptually translated data through various bioinformatics tools has revolutionized the drug discovery process. Various *in silico* based methodologies have gained importance over laboratory experiments for the identification of potential drug targets. Use of subtractive genomics approach helps in the identification of those essential genes of pathogen which must not be homologous to the host genome. These non homologous essential genes ensure the survival of the pathogen and therefore can be targeted for drug development. Drugs against such potential targets are assumed to hold fewer side effects and chances of drug resistance are also expected to be low in future. Various novel drug targets have already been successfully identified in *P. aeruginosa* (Perumal *et al*., 2007), *S. typhi* (Rathi *et al*., 2009) and *N. meningitides* serogroup B (Barh and Kumar, 2009) using this approach. The current study encompasses the high through put screening of complete proteome of *C. trachomatis* strain D/UW-3/Cx using BLAST tool against human genome for the

****Corresponding author:*** khalidanaveed@baqai.edu.pk

identification of non homologous sequences. Subtractive genome analysis also constitutes the use of Database of Essential genes (DEG) for the identification of potential drug targets. Moving one step ahead, differential pathway analysis and subcellular localization of identified prospective drug candidates have also been carried out. In present study, subtractive genomics approach has been utilized to identify the potential druggable non homologous essential proteins of *C. trachomatis* strain D/UW-3/Cx.

## MATERIALS AND METHODS

Subtractive Genomics approach was implemented for the identification of essential proteins in the *C. trachomatis* strain D/UW-3/Cx which were then analyzed for the identification of potential drug targets. The identified drug targets were then screened through DrugBank database to evaluate their druggability scope. The overall work flow is described in Figure 1.

### A. Complete Proteome Retrieval

The complete proteome of *C. trachomatis* strain D/UW-3/Cx was retrieved from NCBI (www.ncbi.nlm.nih.gov/) (Tatusova *et al*., 1999).

### B. Identification of Essential genes in *C. trachomatis*

Paralogous sequences were excluded from the complete proteome of *C. trachomatis* by using CD-HIT (Li *et al*., 2001) at 80% threshold. BLASTp was performed for the remaining proteins against *H. sapiens* using threshold expectation value (E- value) $10^{-3}$ as parameter. This exercise screened out those sequences which were not homologous to *H. sapiens* and were compiled. These non homologous protein sequences were then subjected to BLASTp against the Database of Essential Genes (DEG) accessed at http://tubic.tju.edu.cn/deg/ using E-value cut-off of $10^{-5}$, to screen out essential gene proteins. The resultant data set represents the non homologous essential proteins of *C. trachomatis* strain D/UW-3/Cx of serovar D.

### C. Functional Categorization of the Non Homologous Essential Proteins

Functional family prediction of the non homologous hypothetical essential proteins was done by using the SVMProt (http:/ /jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.) (Cai *et al*., 2003).

### D. Metabolic Pathway Analysis

KEGG Automatic Annotation Server (KAAS) (www.genome.jp/tools/kaas/) was accessed to analyze the metabolic pathways of the essential proteins of *C. trachomatis* strain D/UW-3/Cx for the identification of potential drug targets. The server performs BLAST comparisons of the query protein against Kyoto Encyclopedia of Genes and Genomes (KEGG) Genes database (Moriya *et al*., 2007).

### E. Sub Cellular Localization Analysis

Non homologous essential surface membrane proteins of bacteria illustrate their potential of becoming the possible vaccine targets in future. Therefore, PSORT tool at expasy server was utilized to identify the subcellular localization of non homologous essential protein sequences.

### F. Drug Targets Prioritization

The modulation of the activity of a protein target with a small molecule of a drug accounts for its prospective druggability. In order to prioritize our drug targets DrugBank database (Knox *et al*., 2011) containing 6796 drug entries; was accessed to calculate the druggability potential of each identified drug target. BLASTp with default parameters was used to align the potential drug targets from *C. trachomatis* serovar D against the list of the protein targets of compounds found within the Drug Bank.

## RESULTS AND DISCUSSION

The complete proteome of *C. trachomatis* strain D/UW-3/Cx retrieved from NCBI comprised of 895 number of protein sequences. No redundancy was found as analyzed by the CD-HIT program at 80% identity. Therefore no sequence was excluded from the study plan. Overall summary of the project is depicted in Table 1. These 895 non-paralogous sequences were then subject to BLASTp against the human genome. The results showed that 623 proteins were non homologous with the host genome. Sequences which are homologous to host genome when treated as drug targets might lead to certain undesirable side effects and cytotoxic reactions in the patients. Genes essential for the survival of any pathogen along with the fact that those genes should also be non homologous with the host genome; hold great promise to become specie-specific drug targets. Hence, the 623 human non homologous proteins were subjected to BLASTp against DEG with an E-value cutoff score of $10^{-5}$ to obtain a set of 203 numbers of proteins that were essential to *C. trachomatis* strain D/UW-3/Cx. Among 203 proteins around 39 essential proteins were found to be part of membrane of pathogen using PSORT tool at Expasy server. Functional annotation of 12 hypothetical proteins which were found to be essential was done through SVMProt web server (Table 2). The predicted transmembrane

**Table 1: Subtractive proteomic and metabolic pathway analysis result for *C. trachomatis* strain D/UW-3/Cx**

| Total Number of Proteins | 895 |
|---|---|
| Duplicates (>80% identical) in CD-HIT | No paralogous sequences were found |
| Number of proteins without hits in *H. sapiens* (E-value $10^{-3}$) | 623 |
| Essential proteins in DEG (E-value $10^{-5}$) | 203 |
| Essential proteins involved in metabolic pathways | 182 |
| Number of hypothetical protein as essential proteins | 12 |
| Number of essential membrane proteins | 39 |
| Essential proteins found to be druggable | 11 |

**Table 2: Functional annotation of non-homologous essential genes of *C. trachomatis* strain D/UW-3/Cx**

| GENE ID | Predicted protein family name | R-value | P-value (%) |
|---|---|---|---|
| gi\|15604730 | Transmembrane | 3.9 | 97.5 |
| gi\|15604766 | Zinc-binding | 2.3 | 88.1 |
| | Transmembrane | 1.8 | 80.4 |
| gi\|15604775 | Zinc-binding | 3.9 | 97.5 |
| | Transferases - Glycosyltransferases | 2.2 | 86.8 |
| gi\|15604796 | Iron-binding | 4.8 | 98.6 |
| | Zinc-binding | 3.0 | 94.2 |
| | All lipid-binding proteins | 3.0 | 94.2 |
| gi\|15604804 | All lipid-binding proteins | 2.7 | 92.1 |
| | Hydrolases - Acting on Ester Bonds | 2.0 | 83.9 |
| | Iron-binding | 1.9 | 82.2 |
| gi\|15604870 | Transmembrane | 6.1 | 99.0 |
| | Transferases - Glycosyltransferases | 2.6 | 91.3 |
| gi\|15604931 | Manganese-binding | 3.0 | 94.2 |
| | All lipid-binding proteins | 2.2 | 86.8 |
| gi\|15604976 | Magnesium-binding | 1.0 | 58.6 |
| gi\|15605058 | Motor protein 1.0 58.6 | 1.0 | 58.6 |
| gi\|15605200 | Transmembrane | 1.5 | 73.8 |
| gi\|15605424 | Transferases - Transferring Phosphorus-Containing Groups | 2.0 | 83.9 |
| | Iron-binding | 1.5 | 73.8 |
| gi\|15605496 | All DNA-binding | 1.1 | 62.2 |
| | Zinc-binding | 1.1 | 62.2 |

**Table 3: Non-homologous essential proteins of *C. trachomatis* strain D/UW-3/Cx similar to binding partners of FDA approved drugs against DrugBank database using BLASTP**

| GENE ID | Protein name | DrugBank ID | DrugBank organism |
|---|---|---|---|
| gi\|15604991 | cell division protein FtsI | (1)DB00535<br>(2)DB01327; DB01413; DB00267; DB00274; DB01328; DB01329; DB01331; DB00430; DB01416; DB00438; DB01415; DB01332; DB00303 | (1) *N. gonorrhoeae*<br>(2) *E. coli* |
| gi\|15605022 | serine/threonine protein kinase, bacterial | (1) DB00482 (2)DB01169; DB00995; DB00244; DB00795 | *H. sapiens* |
| gi\|15605023 | valyl-tRNA synthetase | (1) DB00161<br>(2) DB00410<br>(3) DB00167 | (1) *H. sapiens*<br>(2) *S. aureus*<br>(3) *H. sapiens* |
| gi\|15605028 | V-type H+-transporting ATPase subunit B | (1) DB05260<br>(2) DB00630; DB01077; DB01133 | (1) *H. sapiens*<br>(2) *H. sapiens* |
| gi\|15605130 | Riboflavin synthase | DB00140 | *E. coli* (strain K12) |
| gi\|15605182 | UDP-N-Acetylglucosamine Transferase (murA) | DB00828 | *E. coli* |
| gi\|15605236 | RNA Polymerase Alpha (rpoA) | DB00615 | *E. coli* |
| gi\|15605252 | L22 Ribosomal Protein (rl22) | DB00207, DB00199, DB01369 | *E. coli* |
| gi\|15605415 | PBP2-transglycolase/transpeptidase (pbpB) | (1) DB00671, DB00303 (2) DB01327, DB01413, DB01328, DB01329, DB00438, DB01415, DB00303, DB01598, DB00948 | *(1) H. influenzea*<br>(2) *E. coli* |
| gi\|15605495 | D-alanine-D-alanine ligase | DB00260 | *E. coli* |
| gi\|15605544 | Large subunit ribosomal protein L32 | DB01361 | *D. radiodurans* |

**Table 4: Unique pathways of the selected druggable targets of *C. trachomatis* strain D/UW-3/Cx**

| GENE ID | Protein name | Unique Pathway |
|---|---|---|
| gi\|15604991 | cell division protein FtsI (penicillin-binding protein 3) | Glycan Biosynthesis and Metabolism(Peptidoglycan biosynthesis) |
| gi\|15605130 | Riboflavin synthase | Metabolism of Cofactors and Vitamins : Riboflavin metabolism |
| gi\|15605182 | UDP-N-Acetylglucosamine Transferase (murA) | Peptidoglycan Synthesis |
| gi\|15605236 | RNA Polymerase Alpha (rpoA) | Transcription |
| gi\|15605252 | L22 Ribosomal Protein (rl22) | Translation |
| gi\|15605415 | PBP2-transglycolase/transpeptidase (pbpB) | Peptidoglycan biosynthesis |
| gi\|15605495 | D-alanine-D-alanine ligase | D-Alanine metabolism Peptidoglycan biosynthesis |

proteins and transporter protein may possibly be taken as potential drug targets. Metabolic pathway analysis of 182 proteins resulted that 4 proteins are involved in Carbohydrate metabolism, 11 in amino acid metabolism, 11 lipid metabolism, 39 in signal transduction pathways, 23 in membrane component synthesis, 12 in energy metabolic pathways, 25 in genetic information processing (replication), 2 in steroid biosynthesis, 2 in nucleotide metabolism, 47 in protein synthesis and processing and 6 in the metabolism of cofactors and vitamins. Around 25 proteins take part in the replication of *C. trachomatis* strain D/UW-3/Cx (Figure 2). If inhibitor for these particular proteins will be designed then there is a better chance of interrupting the replication of *C. trachomatis* strain D/UW-3/Cx; and in this way pathogenesis can be controlled. We expanded our approach and analyzed the drug ability potential of each non homologous essential 203 proteins of *C. trachomatis* serovar D. This led to the identification of 11 *C. trachomatis* serovar D proteins that shared similarity with the available targets of FDA-approved drugs at the Drug Bank database. The details of the proteins

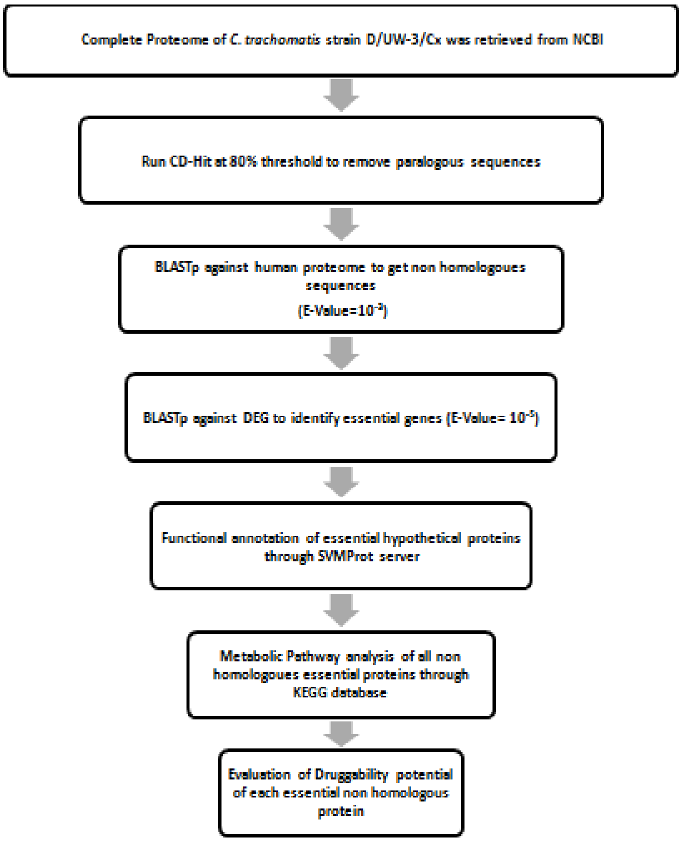
Complete Proteome of C. trachomatis strain D/UW-3/Cx was retrieved from NCBI
Run CD-Hit at 80% threshold to remove paralogous sequences
BLASTp against human proteome to get non homologoues sequences
(E-Value=$10^{-3}$)
BLASTp against DEG to identify essential genes (E-Value= $10^{-5}$)
Functional annotation of essential hypothetical proteins through SVMProt server
Metabolic Pathway analysis of all non homologoues essential proteins through KEGG database
Evaluation of Druggability potential of each essential non homologous protein

**Figure 1: Overall Work flow of the Study**

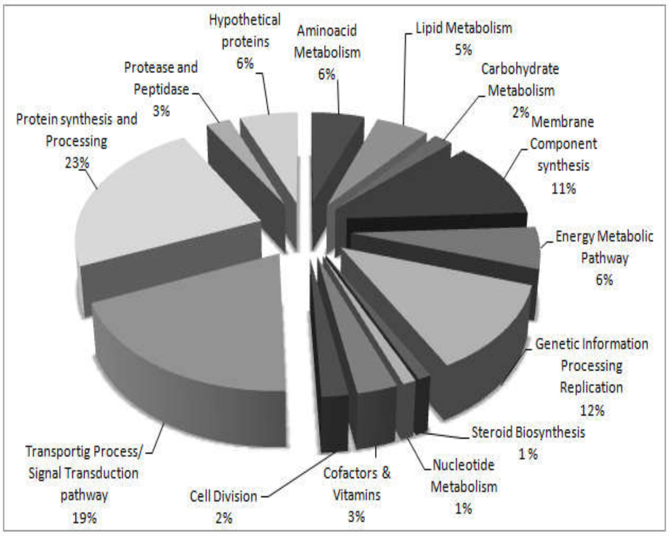

**Figure 2: Distribution of non-homologous essential proteins of *C. trachomatis* strain D/UW-3/Cx in different metabolic pathways**

along with targets ID are available in Table 3. Comparative analysis of the metabolic pathways of these identified 11 druggable proteins against pathogen (*C. trachomatis* strain D/UW-3/Cx) and the host (*H. sapiens*) by using Kyoto Encyclopedia of Genes and Genomes (KEGG) disclosed 7 druggable proteins to be involved in the unique pathways of *C. trachomatis* serovar D (Table 4). These unique non homologous essential proteins might be the good drug targets due to their involvement in various metabolic pathways which are indispensible for bacterial survival and growth.

The Peptidoglycan layer of bacteria plays an imperative role in the pathogenesis as it helps the pathogen to withstand the osmotic lysis. Peptidoglycan is a network of alternating chains of N-acetylglucosamine and N-acetylmuramic acid cross-linked by peptides attached to the N-acetylmuramic acid (Ghuysen, 1968; Schleifer and Kandler, 1972). The biosynthsis of peptidoglycan involves various ADP forming ligases such as MurA, MurC, MurD, MurE and MurF which catalyze the successive additions of L-alanine, D-glutamate, a diamino acid and D-alanine- D-alanine to UDP N-acetylmuramic acid (Rogers *et al*., 1980). The peptide cross linking provides tremendous strength to the bacteria. In this study, UDP-N-Acetylglucosamine Transferase (MurA) and D-alanine-D-alanine ligase enzyme were found to be essential proteins which did not show any homology with human. They have also been identified to be involved in the unique pathways of *C. trachomatis* serovar D and could become probable drug targets. Enzymes catalyzing the riboflavin biosynthetic pathway are evolutionarily conserved in bacteria and absent in humans (Bacher *et al*., 1996). Riboflavin acts as a precursor in the synthesis of flavin mononucleotide and flavin adenine di-nucleotide, which are both essential cofactors involved in the basic energy metabolism of the cell. Being non homologous and essential protein riboflavin synthase enzyme; can be an attractive candidate for drug development.

## Conclusion

The idea of interpretation and the analysis of the available genome and proteome sequences of various pathogens at different databases through a number of bioinformatics tools have revolutionized the drug discovery process. The increase in the drug resistance cases supports the utilization of *in silico* methodologies for the identification of better potential drug targets which should not display any homology with host proteome. Subtractive genomics approach facilitates this process in the characterization of the non homologous essential proteins that could be targeted for the discovery of potential therapeutic compounds against *C. trachomatis* strain D/UW-3/Cx. Strain specific drugs targeting the non homologous essential proteins of pathogen ensures the eradication of disease with fewer side effects to the host.

## REFERENCES

1. Bacher, A., Eberhardt, S., and Richter, G. 1996. In Escherichia coli and Salmonella: Cellular and Molecular Biology; 2nd ed.; Neidhardt, F. C., Ed.; ASM Press: Washington, DC, pp. 657-664.
2. Barh, D., and Kumar, A. 2009. In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae. *In Silico Biol.,* 9: 225-231.
3. 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:*** 3692-3697.
4. Fredlund, H., Falk, L., Jurstrand, M., and Unemo, M. 2004. Molecular genetics methods for diagnosis and characterization of Chlamydia trachomatis and Neisseria gonorrhoeae: impact on epidemiological surveillance and interventions. *APMIS.,* 112: 11-12.

5. Ghuysen, JM. 1968. Use of bacteriolytic enzymes in determination of wall structure and their role in cell metabolism. *Bacteriol Rev.* 32: 425-464.
6. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, AC., and Wishart, DS. 2011. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.,* 39: D1035-1041.
7. Li, W., Jaroszewski, L., and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics.,* 17: 282-283.
8. Morré, SA., Spaargaren, J., Ossewaarde, JM., Land, JA., Bax, CJ., Dörr, PJ., Oostvogel, PM., Vanrompay, D., Savelkoul, PH., Pannekoek, Y., van Bergen, JE., Fennema, HS., de Vries, HJ., Crusius, JB., Peña, AS., Ito JI, and Lyons, JM. 2006. Description of the ICTI consortium: an integrated approach to the study of Chlamydia trachomatis infection. *Drugs Today (Barc).,* 42: 107-114.
9. 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: W182- W185.
10. Perumal, D., Lim, CS., Sakharkar, KR., and Sakharkar, MK. 2007. Differential genome analyses of metabolic enzymes in Pseudomonas aeruginosa for drug target identification. *In Silico Biol.,*7: 453-465.
11. Rathi, B., Sarangi, AN., and Trivedi, N. 2009. Genome subtraction for novel target definition in Salmonella typhi. *Bioinformation.,* 4: 143-150.
12. Rogers, HT., Perkins, HR., and Ward, J B. 1980. Microbial cell walls and membranes. Chapman & Hall Ltd, London., pp. 239–297.
13. Schleifer, KH., and Kandler, O. 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriol Rev.* 36: 407-477.
14. Stephens, RS., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, RL., Zhao, Q., Koonin, EV., and Davis, RW. 1998. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science.,* 282: 754-759.
15. Tatusova, TA., Karsch-Mizrachi, I., and Ostell, JA. 1999. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics.,* 15: 536–543.

*******