Journal of Bioinformatics and Sequence Analysis Vol. 2(3), pp. 36-41, July 2010
Available online at http://www.academicjournals.org/JBSA
ISSN 2141-2464 ©2010 Academic Journals

Full length Research Paper

In silico identification of potential horizontal gene transfer events between archaea and pathogenic bacteria

Hasan Bilal Mirza*, Maryam Anwar and S. Habib Bokhari

Department of Biosciences, COMSATS Institute of Information Technology, Chak Shehzad Campus, Islamabad, Pakistan.

Accepted 21 June, 2010

Horizontal gene transfer plays a potent role in the evolution of prokaryotes. A rigorous sequence and phylogenetic analysis was carried out using the robust ClustalW, motifs/domains finding suites and neighbor-joining based ProtDist and BioNJ. This paper reports a few cases of horizontal gene transfer events between archaea and bacteria. Some of these events have been found to be unique to the bacterial pathogenic members and have not been observed in respective non-pathogenic counterparts. Two cases have been shown to exhibit particular importance. The first one is a Cps4I gene that codes for capsule polysaccharide biosynthesis protein in *Streptococcus pneumoniae*. The other gene has been detected in *Streptococcus agalactiae* that codes for N-acetyl neuramic acid synthetase, which is involved in the synthesis of N-acetyl neuramic acid or sialic acid. We believe that these genes, having been retained in the genome through selective advantage, have key functions in the organism's biology and may play a role in pathogenesis.

Key words: Horizontal gene transfer, BLAST, ClustalW, conserved domain, motif, bootstrap value, bit score.

INTRODUCTION

In prokaryotes, the principal mode of genetic flexibility is the natural genetic transformation. However, in order to adapt to new environments, microbes acquire novel genes through horizontal gene transfer (HGT) from the inhabitants of the environment (Doolittle, 1999). It has also been observed that horizontal gene transfer events play a more important role as compared to alterations in gene functions mediated by point mutations, in the adaptation of microbes to new environments (Wiezer et al., 2005). For the same reason, certain microbial species may look similar to each other although there is no ancestor-descendent relationship between them.

Like bacteria, archaea are widely distributed. While they resemble bacteria in their shapes and various cell structures, they differ immensely in the chemical composition of their structures. Extensive analysis on DNA and biochemical features of archaea has revealed many differences in comparison to bacteria including that of ribosomal RNA, cell wall composition, types of lipids used in the cell membrane, and the way DNA is packaged and transcribed (Rossi et al., 2003). Archaea consist of molecular features that are encoded by two different groups of genes. One group is eukaryotic in nature and is called the group of informational genes,

*Corresponding author. E-mail: hasan_bilal@comsats.edu.pk. Tel: +92-331-5581783.

Abbreviations: HGT, Horizontal gene transfer; **NCBI,** national centre for biotechnology information; **BLAST,** basic local alignment search tool; **MEME,** multiple Em for motif elicitation; **MAST**, motif alignment and search tool; **VFDB,** virulence factor database; **MVirDB,** a microbial database of protein toxins, virulence factors and antibiotic resistance genes; **LLNL,** lawrence livermore national laboratory; **CD-SEARCH,** converved domain search; **ABC-** Transporter, **ATP-**binding cassette transporter; **CPS4I,** gene that codes for capsule polysaccharide biosynthesis protein in Streptococcus pneumonia; **NeuB,** gene that codes for N-acetyl neuramic acid synthetase in Streptococcus agalactiae.

which are involved in transcription, translation and other such processes. The other group of genes is similar to bacterial genes and includes the operational genes that encode housekeeping functions (Garcia-Vallve et al., 2000; Jain et al., 1999).

Recent studies have indicated that horizontal transfer events are greater for operational genes than the informational genes. It has been suggested that the reason for this partiality is the fact that informational genes are associated to large complex systems as compared to operational genes making HGT of informational genes difficult (Jain et al., 1999). Despite extensive genome sequencing and DNA analysis, there are certain features of archaea that are still unknown due to the fact that they are difficult to isolate and culture.

Archaea have been found to be capable of colonizing in the human host as the normal flora. There are anaerobic archaea in the human colon, vagina and oral cavity (Eckburg et al., 2003). However, no virulence genes have been identified in archaea till now. Recent studies have led to the identification of certain characteristics in archaea that are common to known pathogens indicating towards the possibility of a probable role of archaea in causing virulence. Since Griffith, HGT has been thought of as a mode to acquire novel virulence genes in pathogens. Recent sequencing of bacterial and archaeal genomes has shown that inter-domain transfer is common. For instance, a large fraction of *Thermotoga maritima* genes appear to be of archaeal origin (Mongodin et al., 2005). Since archaea are not directly involved in causing a disease but there are similarities between pathogenic bacterial genome and archaeal genome, archaea might be linked to virulence as donors of virulence-promoting genes to pathogenic bacteria through the process of lateral gene transfer (Faguy, 2003).

To explore the issue of horizontal gene transfer between archaea and bacteria and to determine the probable direction of transfer, we have analyzed and compared the genomes of archaea and pathogenic bacteria through the use of web-based computational and statistical tools including specialized softwares such as MEME (Multiple Em for motif elicitation) and MAST (Motif alignment and search tool) for the identification of highly conserved motifs and protein function and Neighbor-Joining packages for phylogenetic analysis. In order to establish if the transfer events between archaea and bacteria also contribute to virulence, we searched and compared the probable candidates of horizontal transfer in pathogenic strains with their non-pathogenic counterparts to examine if they met specific criteria for archaea to bacteria gene transfer or vice versa.

MATERIALS AND METHODS

In order to find the horizontal gene transfer of potential virulence candidate proteins between archaea and pathogenic bacteria, the following scheme was implemented:

Retrieval of pathogenecity-associated sequences

In the first part of this step, updated organism lists of bacterial pathogens, bacterial non-pathogens and archaeal species/strains were obtained from National centre for biotechnology information (NCBI) (http://www.ncbi.nlm.nih.gov/.html). Only those bacterial pathogens were selected for further study whose non-pathogenic counterparts were available at the genus level. About 20 bacterial pathogens were selected in the first part. In the second part of this step, sequences of pathogenecity-associated proteins of the 20 selected bacterial pathogens were retrieved from two major databases: The Virulence Factors Database (VFDB) (Chen et al., 2005) and MvirDB (a microbial database of protein toxins, virulence factors and antibiotic resistance genes) Virulence Database at Lawrence Livermore National Laboratory (LLNL) (Zhou et al., 2007).

Sequence BLASTs

The pathogenecity-associated sequences obtained in the first step were subjected to a series of BLASTs (Altschul et al., 1990) against the genomes and proteomes of bacterial pathogens (excluding pathogens of the same species and genus), bacterial non-pathogens and archaea. From the Blast results, Protein sequences from pathogenic bacteria were selected as probable HGT candidates if they matched the following criteria:

(A) A bit score of at least 105 in BLAST with Archaea and at most 90 in BLAST with Non-Pathogenic counterparts.
(B) At least 2 matches in BLAST with Archaea.
(C) If BLAST with bacterial pathogens resulted in matches with a bit score value greater than that of the best match in archaea, then the candidate protein was selected only if the best matches in bacterial pathogens were from the same species or genus as the candidate. In cases where the number of matches of a specific protein sequence exceeded 15, only first 15 matches were considered.

Multiple sequence alignment

The sequences selected in the second step were subjected to multiple sequence alignment to check for the presence of conserved regions. For this purpose the robust ClustalW tool (Thompson et al., 1994), present at the EBI web server was used.

Identification of conserved motifs and domains

In the first part of this step, each set of similar sequences (HGT candidate, archaeal, bacterial) was given as input to two collaborative web-based softwares MEME and MAST (http://meme.sdsc.edu) to identify highly conserved regions in the sequences. In the second part of this step, the consensus sequences of conserved regions in each set were given as input to CD-Search (Conserved domain search, Marchler et al., 2004) to identify the domains to which the conserved motifs corresponded.

Construction of phylogenetic trees

To ensure optimality of results, the ClustalW alignments of the selected proteins were subjected to refinement using the JALVIEW alignment editor (Clamp et al., 2004). The edited multiple sequence alignments were used to build Distance based Neighbor-Joining Trees using ProtDist and BioNJ packages of Phylip. PhyloDraw (Jeong et al., 2000) was used to display phylogenetic

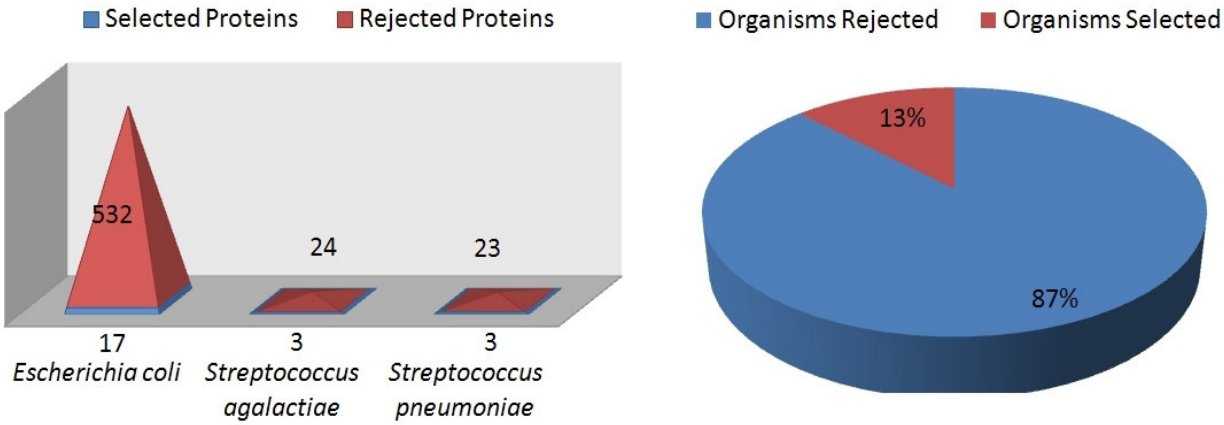

Figure 1. Statistics of the selected and rejected bacteria and their proteins. Of all the bacteria, only 13% were selected for the study as their non-pathogenic counterparts were available at genus level. After BLAST results, proteins from only three bacteria, that is, *Escherichia coli*, *Streptococcus agalactiae* and *Streptococcus pneumonia* fit into selected criteria.

Table 1. Categorical division of the selected proteins**.** The majority of selected proteins belong to 5 protein classes except for three individual proteins. Here ABC ~ ATP-binding cassette transporter.

Protein groups	Individual proteins
ABC transporters family Bacterial capsular proteins	DNA methylase
Transposases and helper proteins Reverse transcriptases	Helicase
Restrictions enzymes	Lysyl t-RNA synthetase

trees in graphical form by manipulating the shape of phylogenetic trees and interactively by using several control parameters.

RESULTS AND DISCUSSION

A total of 23 proteins were selected based on the pre-defined criteria. Their division among bacterial pathogens and further categorical divisions are given in Figure 1 and Table 1, respectively.

Few of the HGT candidates in our analysis were directly or indirectly related to virulence. It appears so that the genes acquired by horizontal gene transfer from archaea are more involved in normal biosynthetic functions rather than in pathogenesis of bacteria. Primarily we identified HGT events for genes involved in capsule biosynthesis and DNA regulatory or modification functions. Two events are of particular importance.

The first one is a protein coded by Cps4I gene in *Streptococcus pneumoniae*. *S. pneumoniae* is a human pathogen that causes a variety of diseases including otitis media, pneumonia, sepsis and meningitis (Roche et al., 2000). Cps4I, a capsule polysaccharide biosynthesis protein is also known as UDP-N-acetylglucosamine-2-epimerase. It catalyzes the reversible epimerization of UDP-N-acetylglucosamine (UDP-GlcNAc) at carbon-2, providing bacteria with UDP-N-acetylmannosamine (UDP-ManNAc), which is the activated donor of ManNAc residues. ManNAc is one of the major virulence factors and plays a very important role in several bacterial processes such as formation of the antiphagocytic capsular polysaccharide in pathogens (Campbell et al., 2000). A homologue of UDP-N-acetyl glucosamine in mammals is a bifunctional enzyme involved in the initiation and regulation of biosynthesis of sialic acids that play a role in cell-cell and cell-matrix interactions (Astrid et al., 2004). According to our analysis, Cps4I appears to have been transferred horizontally from archaea. The relationship between Cps4I of *S. pneumoniae* and N-acylneuraminate-9- phosphate synthase of several archaea was clearly depicted in our phylogenetic tree as shown in Figure 2.

MEME and MAST results showed three highly conserved motifs in CPS4I and its archaeal counterparts as shown in Table 2. Based on these results, two main domains WecB (involved in cell envelope biogenesis) and Epimerase_2 (UPD-N-acetylmannosamine kinase activity) were found to be conserved among Cps4I and methanococcal sequences (Table 3).

The HGT candidate Cps4I matched best with the N-acylneuraminate-9- phosphate synthase of *Methanospirillum hungatei JF-1*, which was also its closest neighbor in the archaeal cluster. Though the function of Cps4I is quite clear, the exact function of N-acylneuraminate-9-phosphate synthase in archaea is unknown as yet (Wilson et al., 2005). Based on our phylogenetic analysis, we infer that an HGT event involving capsular proteins has taken place from archaea to bacteria and may have contributed to virulence in bacteria.

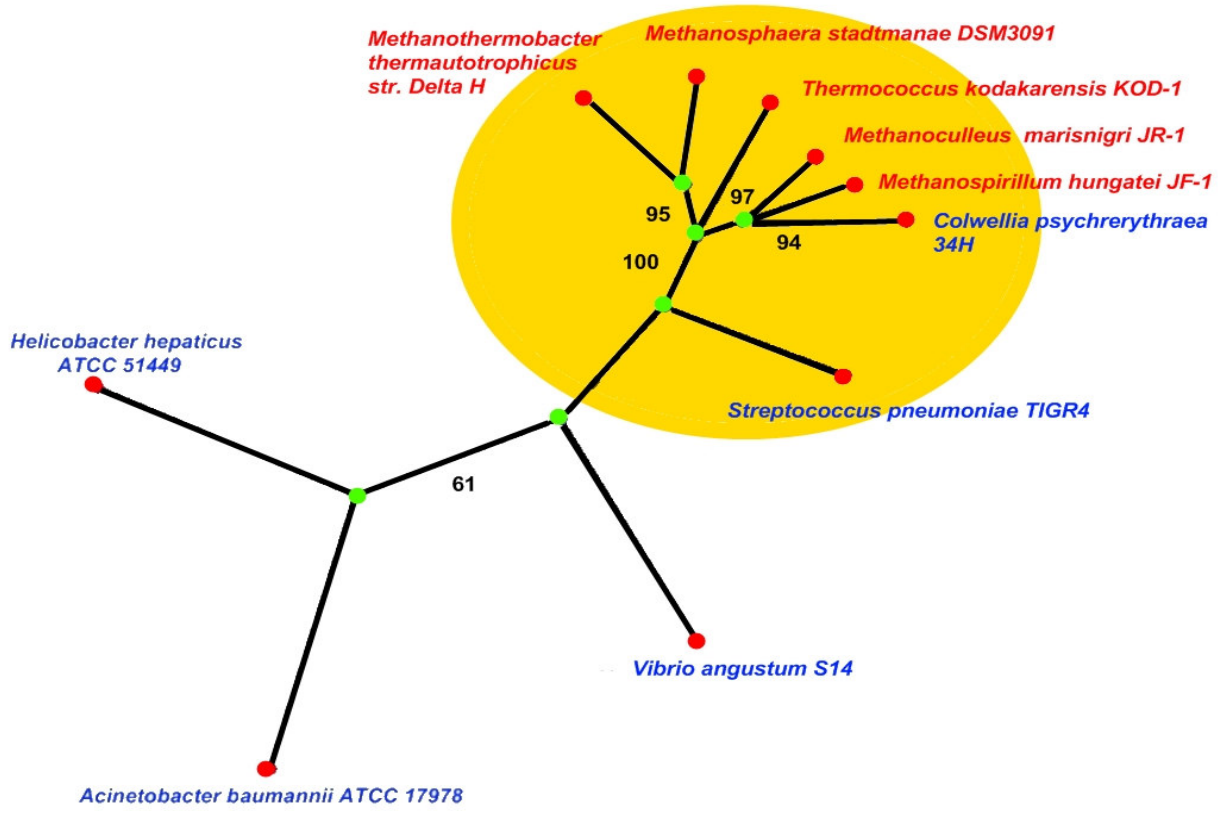

Figure 2. Phylogenetic tree of Cps4I, its archael and bacterial matches. The tree clearly depicts the grouping of *Streptococcus pneumonia* among archaeal sequences.

Table 2. Cps4I motif sequences.

Motif	Width	Consensus sequence
1	50	LIKPLGYLDFLQLLSNAFLVLTDSGGIQEEACTFGVPCVTLRYNTERPET
2	29	QEKPDCVLVQGDTNTVFAGALAAFKLQIP
3	15	GHVEAGLRSYDRYMP

Table 3. Cps4I domains and their functions.

Motifs	Matches with domains	Function of domains
MOTIF 1	WecB	Cell envelope biogenesis
	Epimerase_2	UPD-N-acetylmannosamine kinase activity
MOTIF 2	Epimerase_2	UPD-N-acetylmannosamine kinase activity
MOTIF 3	No significant results	

Our second HGT candidate is a gene for capsular protein and was detected in the pathogen *Streptococcus agalactiae*. *S. agalactiae* is the causative agent of a multitude of infections including sepsis (Maeland et al., 2005), meningitis, bacterimia (Férez et al., 1991) and osteoarticular infections (Gómez et al., 1995). The candidate capsular protein called NeuB (N-acetyl neuramic acid synthetase) is involved in the synthesis of N-acetyl neuramic acid or sialic acid, which is displayed on the cell surface (Haft et al., 1994). In microbes it is antiphagocytic in nature and therefore weakens immune recognition to enhance pathogenecity (Haft et al., 1994).

MEME and MAST results showed three highly conserved motifs in NeuB and its archaeal counterparts (Table 4). These motifs correspond to a highly conserved and important domain 'NeuB' in the NeuB protein of *S.*

Table 4. NeuB motif sequences.

Motif	Width	Consensus sequence
1	50	VGYSDHTLGIYVPIAAVAMGACVIEKHFTLDRNMPGPDHKASLEPDEFRT
2	32	WKIPSGEITNYPYLRKIGRQQQPVILSTGMAT
3	50	RRSIVAKCDIQKGEIFSEDNLTVKRPGTGISPMYWDQWCGRQARRDYQED

Table 5. NeuB domains and their functions.

Motifs	Matches with domains	Function of domains
MOTIF 1	NeuB	Catalyses the direct formation of Neu5Ac (the most common sialic acid)
MOTIF 2	No significant results	
MOTIF 3	SAF: UxaA/GarD-like hexuronate dehydratases	Present in antifreeze proteins, flagellar FlgA proteins, and CpaB pilus proteins.

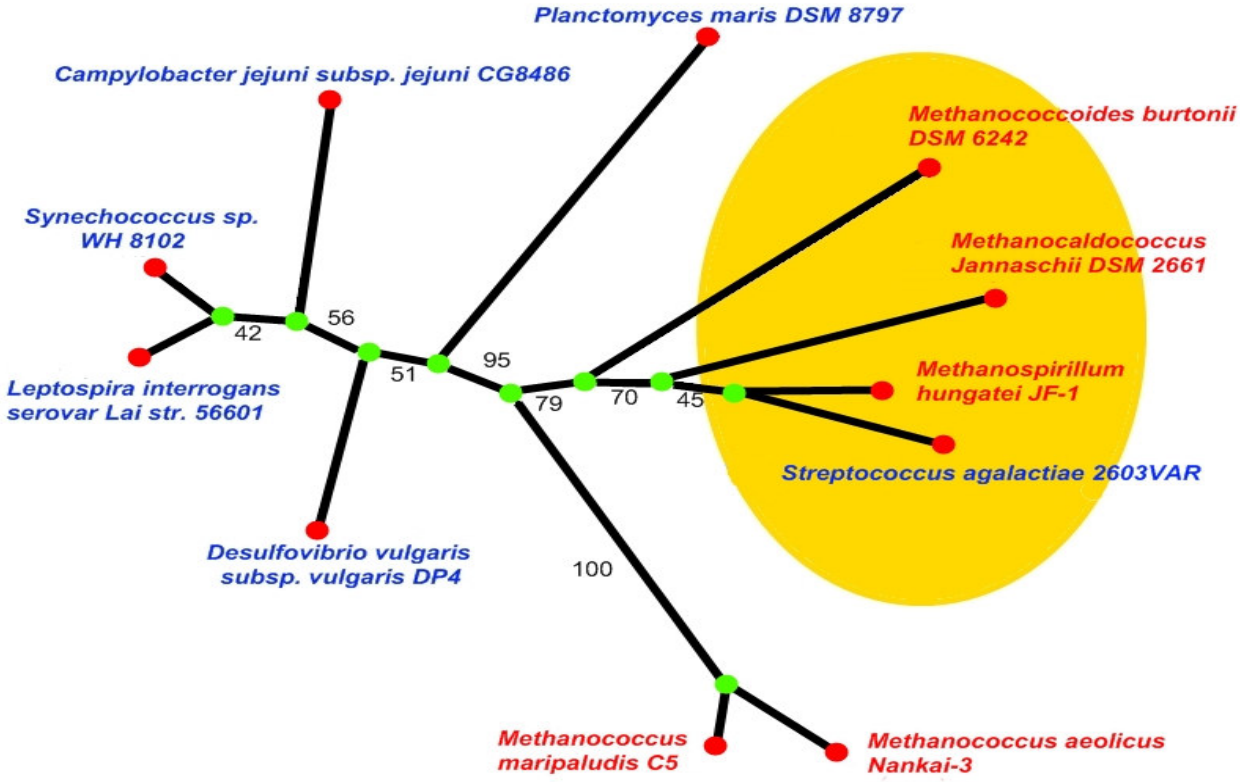

Figure 3. Phylogenetic tree of NeuB, its archael and bacterial matches. The tree clearly depicts the grouping of Streptococcus agalactiae among archaeal sequences.

agalactiae and in N-acylneuraminate-9-phosphate synthases of Methanococci (Table 5).

As the name of the domain 'NeuB' indicates, it is vital to the function of the protein NeuB. The closest match to the NeuB protein of *S. agalactiae* that was determined through phylogenetic analysis is N-acylneuraminate-9-phosphate synthase of *Methanospirillum hungatei JF-1* (Figure 3). As mentioned earlier, N-acylneuraminate-9-phosphate synthase is categorized as an orphan protein and its function is not well understood in archaea (Wilson et al., 2005). Microarray experiments involving comparison of genomes of *S. agalactiae* and other streptococci revealed differences in the polysaccharides of the capsule as well as in several of the metabolic pathways and transport systems (http://www.innovations-report.com). Such a genetic diversity indicates a probable

acquisition of genes horizontally.

Conclusion

The findings suggest that a sequential acquisition of archaeal genes is an important feature in the evolutionary history of bacteria. Our study depicted that the genes acquired by horizontal gene transfer from archaea are more involved in normal biosynthetic functions rather than in pathogenesis of bacteria. Apart from the good cases including capsule biosynthesis proteins, we identified genes involved in DNA regulatory or modify-cation functions. The important aspect of this research is the identification of the role of archaea as a reservoir of a variety of metabolic innovations for bacteria. Such metabolic innovations not only enable the bacteria to adapt to new environment but may also contribute to a better survival in or on the host.

ACKNOWLEDGEMENTS

This work is a part of an undergraduate thesis supported by the department of biosciences, COMSATS Institute of Information Technology. We are grateful to our super-visor, Dr. S. Habib Bokhari for his impetus guidance, to the Chairman of the department of biosciences, Dr. Raheel Qamar for providing us with the facilities to work on this project. We are also thankful to the 'development and maintenance teams' of all the free software tools utilized in this research.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol., 214: 1-8.

Astrid B, Wenke W, Ulrich S, Erich EW, Lothar L, Peter D, Werner R, Rüdiger H, Stephan H (2004). Domain-specific characteristics of the bifunctional key enzyme of sialic acid biosynthesis, UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase. Biochem. J., 384: 599-607.

Campbell RE, Mosimann SC, Tanner ME, Strynadka NC (2000). The structure of UDP-N-acetylglucosamine 2-epimerase reveals homology to phosphoglycosyl transferases. Biochemistry, 39(49): 14993-5001.

Chen LH, Yang J, Yu J, Yao ZJ, Sun LL, Shen Y, Jin Q (2005). VFDB: A reference database for bacterial virulence factors. Nucleic Acids Res., 33: 325-328.

Clamp M, Cuff J, Searle SM and Barton GJ (2004). The Jalview Java Alignment Editor. Bioinformatics, 20: 426-427.

Doolittle WF (1999). Phylogenetic classification and the universal tree. Sciences, 284: 2124-2129.

Eckburg PB, Lepp PW, Relman DA (2003). Archaea and their potential role in human disease (Mini Review). Infect. Immun., 71(2): 591-596.

Faguy DM (2003). Lateral Gene Transfer between Archaea and Escherichia coli is a contributor to the emergence of novel infectious disease. BMC Infect. Dis., 3(13): 1471-2334.

Férez A, Fajardo MT, Estellés MA, Moreno R, Esteban A, Martín C, Royo G (1991). Bacteremia caused by Streptococcus agalactiae in adults. Enfermedades Infecciosasy Microbiol. Clin., 9(6): 354-6.

Garcia-Vallve S, Romeu A, Palau J (2000). Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. Genome Res., 10(11): 1719-1725

Gómez RN, Ferreiro JL, Willisch A, Muñoz LR, Formigo E, González MG (1995). Osteoarticular infections caused by Streptococcus agalactiae. Enfermedades Infecciosasy Microbiol. Clin., 13(2): 99-103.

Haft RF, Wessels MR (1994) Characterization of CMP-N-acetylneuraminic acid synthetase of group B streptococci. J. Bacteriol., 176(23): 7372-7374.

Jain R, Rivera MC, Lake JA (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings Nat. Acad. Sci. USA. 96(7): 3801-3806.

Jeong HC, Ho YJ, Hey SK, Hwan G (2000). PhyloDraw: A phylogenetic tree drawing system. Bioinform., 16(11): 1056-1058.

Maeland JA, Bevanger L, Lyng RV (2005) Immunological markers of the R4 protein of Streptococcus agalactiae. Clinical and diagnostic laboratory immunol., 12(11): 1305-10.

Marchler BA, Bryant SH (2004). CD-Search: protein domain annotations on the fly. Nucleic Acids Res., 32: W327-331.

Mongodin EF, Hance IR, DeBoy RT, Gill SR, Daugherty S, Huber R, Fraser CM, Stetter K, Nelson KE (2005). Gene Transfer and Genome Plasticity in Thermotoga maritima, a Model Hyperthermophilic Species. J. Bacteriol., 187(14): 4935-4944.

Roche AM, King SJ, Weiser JN (2007) Live attenuated streptococcus pneumoniae strains induce serotype-independent mucosal and systemic protection in mice. Infect. Immun., 75(5): 2469-75.

Rossi M, Ciaramella M, Cannio R, Pisani FM, Moracci M, Bartolucci S (2003). Extremophiles. J. Bacteriol., 185(13): 3683-3689.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22: 4673-4680.

Wiezer A, Merkl R (2005). A comparative categorization of gene flux in diverse microbial species. Genomics. 86: 462-475.

Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D (2005) Orphans as taxonomically restricted and ecologically important genes. Microbiology., 151: 2499-2501.

Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T (2007). MvirDB-a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res., 35: 391-394.