Available online at www.elixirpublishers.com (Elixir International Journal)

Bio Technology

Elixir Bio Tech. 69 (2014) 23105-23110

ISSN: 2229-712X

Comparative Genome Analysis of *Streptococcus pyogenes* and *Streptococcus equi*

Rohit Lall[1], Nitendra[1] and Satendra Singh[2]

[1]Department of Molecular and Cellular Engineering, JSBB, SHIATS, Allahabad-211007.

[2]Department of Computational Biology & Bioinformatics, JSBB, SHIATS, Allahabad-211007.

ARTICLE INFO

Article history:

Received: 4 February 2014;

Received in revised form:

29 March 2014;

Accepted: 10 April 2014;

Keywords

UTR,

CNS,

Childhood meningitis.

ABSTRACT

The comparative genome analysis of *Streptococcus pyogenes* and *Streptococcus equi* sps equi *4047* depict exon, Untranslated Region (UTR), conserved noncoding sequences (CNS), contigs, mobile genetic elements, Insertion sequence (IS) elements, transposons, plasmids and shows high degree of heterogeneity evidenced by the large number of single nucleotide polymorphisms (SNPs). Thus UTR, CNS, and IS elements can be used as drug target.

© 2014 Elixir All rights reserved

Introduction

The *Streptococcus pyogenes 370* [4] is a gram positive [2, 3, 12], nonmotile, facultative, mesophilic obligate parasitic bacterium with cocci in shape. The genome is circular with a size of 1,852,442 base pairs and an average G+C content of 38.5%. The average G+C content of the protein-coding sequences is 39.1%. The bacteriophage genome contains SLP, PAG encoding virulence factors. The transcription starts in both directions from *oriC* [8]. The presence of alternate transcription signals allows the streptococcus to respond to environmental changes [1, 7]. The salivaricin A (bacteriocin) was present in 90% of *S. pyogenes* strains [13]. The major virulence regulon was controlled by Mga and located upstream to immunogenic secreted protein gene, *isp* [11]. The *Streptococcus pyogenes M1GAS* contains 1,752 protein encoding genes (PEG), 40 putative virulence associated genes, prophage associated genes (PAG), superantigens like proteins (SLP) [9] causing toxic shock syndrome, cellulitis, and rheumatic fever [15-18]. The six potential virulence factors are responsible for horizontal gene transfer (HGT) and generate new strains with increased pathogenic potential. *S.equi*.subsp.equi. *4047* cause childhood meningitis [5] due to rupture of abscesses in retropharyngeal lymph nodes. The functional loss, pathogenic specialization, and genetic exchange results in evolution [10] of *S.equi*.subsp.equi. *4047*.

Comparative genome analysis added value of complete genomes in more sequences, large scale pattern detection in genomes, function/orthology prediction by bi-directional best hit approaches, presence/absence/variation of pathways and prediction of new pathway. Analysis and Comparison of genomes at various levels of DNA (e.g. GC content) (we actually do not need sequencing for that), dinucleotide frequencies, coding densities of leading/lagging strands. GC skew etc.), protein coding potential (e.g. coding density), presence/absence/size of Protein families, presence/absence of genes/comparing at the level of orthologs and Gene Order evolution. Evolution of gene content is identified by Quantitative approaches. Count the number of genes that two genomes share (orthology) and relate to phylogenetic distance. The rate of gene content reconstructs genome evolution. Qualitative approaches: interpret the differences between two genomes in terms of the functions of the encoded proteins to explain the differences between the phenotypes in terms of the genomes' gene content. Genome phylogeny is based on gene content. Count the number of shared orthologs between genomes using the bi-directional best, significant, hit approach (include fusion/fission). Create a similarity matrix by dividing number of shared orthologs by the genome size of the smallest genome. Create a distance based phylogeny from the similarity matrix [14].

Materials and methods

Comparative analysis of DNA sequences was done by selecting *Streptococcus pyogenes* as reference organism and *Streptococcus equi* as compared organism in microbial vista. The microbial vista displays conserved region identity alignment, synteny viewer, and dot plot. The conserved region describes the exon, UTR, conserved noncoding sequences, contigs, and protein coding region alignment between *Streptococcus pyogenes* and *Streptococcus equi*.

The synteny viewer represents the alignment density at particular location in chromosomes. The dot plot describes the chromosome boundaries/ scaffold, the blue color line describes the forward strand and red color line describes the opposite strand. The grey color predict the *Streptococcus pyogenes* chromosome and colored as *Streptococcus equi* chromosome.

The VISTA tools allows to align DNA sequences, quickly visualize conservation levels between them, identify highly conserved regions, and analyze sequences of interest through one of the following approaches: Precomputed whole-genome sequences of reference and compared organisms were browsed and submitted to Genome VISTA or mVISTA to align them with each other (a variety of alignment programs with several distinct capabilities are made available). The web page http://genome.lbl.gov/vista/ serves as a portal for access to all VISTA tools.

Results and discussion

The visual comparative analysis of *Streptococcus pyogenes* and *Streptococcus equi* genome assemblies were done at

Tele:

E-mail addresses: nitendra_20011@rediffmail.com

© 2014 Elixir All rights reserved

different levels of resolution, using pairwise and multiple large-scale alignments. For *Streptococcus pyogenes* the resolution was 39, whereas in *Streptococcus equi* 15 alignments were observed on 146 region having conservation identity 70% and 100bp. The conserved regions, conservation and annotation coloring shows the conservation curves where purple, grey, pink, red and black represents the exon, contigs, conserved non-coding sequence, gaps, and mRNA respectively. The light blue color represents Untranslated Regions. The average identity score for the alignment between *Streptococcus pyogenes* and *Streptococcus equi* was calculated by VISTA curve. The "peaks and valleys" in graph represents present conservation between aligned sequences at a given coordinate on the base genome. The scores for each base pair describes conserved region as "Minimum Conserved Width" (default value 100bp) and "Conserved Identity" (default value 70%) in the genomic interval. The region was conserved if the region had conservation identity greater than or equal to 70% and had minimum length of "Minimum Conserved Width". Regions of high conservation were colored as exons (purple), UTR (light blue) or non-coding (pink). The threshold determines color as well as minimum and maximum percentage. The quality of alignment and alignment overlap was evaluated by Vista plot. The conservation identity of exon was more than and 0% at 17-29, 64-79, 80-85, and 96-100kbp. The conserved non-coding sequence regions were between 58-62kbp (Figure 1).

The conservation identity reveals the gene regulation and alternate splicing which will depend on the position of exon. The protein coding gene contains the regulation genes which control the up regulation and down regulation of protein. The contigs and gap increased the similarity or identity by making the gap between the gene sequences. In mRNA the slicing of gene takes place that makes the replica copy of genes which can be done by sequencing. The replication of gene can be regulated.

The Synteny Browser enables visual comparative analysis of complete genome assemblies at different levels of resolution, ranging from genome-scale comparison of chromosomes to comparisons of individual regions of alignment at the nucleotide level. Synteny in the Synteny Browser was calculated referenced on pair-wise whole genome alignment. Genome Synteny Browser was displayed in three collapsible panels as Genome Panel, Chromosome Panel and Comparison Panel. The color intensity in the Genome Panel indicates the alignment density of gene on chromosomes (or scaffolds, or contigs, or draft assemblies between *Streptococcus pyogenes* and *Streptococcus equi*. Alignment density describes synteny between *Streptococcus pyogenes* and *Streptococcus equi* which was denser at corner. Darker region in the image had higher density of coverage. The Synteny at Chromosome Panel depicts block representing an alignment of *Streptococcus pyogenes* and *Streptococcus equi* genome sequences. The position of the block indicates the alignments location on the *Streptococcus pyogenes* genome and the color of the block indicates the chromosome where the match was found on the *Streptococcus equi* genome. The gene number 1550-2450 in genome had maximum match represented by green blocks. Legends reveal the color-coding scheme. The blocks stacked on top when a fragment of the *Streptococcus pyogenes* genome had synteny with multiple locations in the *Streptococcus equi* genome. Predicted gene appears as black lines with exonic regions indicated in purple blocks. The alignment between specific regions on *Streptococcus pyogenes* and *Streptococcus equi* were displayed in Comparison Panel. The pair of blocks on *Streptococcus pyogenes* chromosome (grey) and other on *Streptococcus equi* chromosomes (colored) were connected by a line. The Chromosome Panel and Comparison Panel display different regions of the *Streptococcus pyogenes* genome. Maximum numbers of synyteny were observed between 300-500kbp (Figure 2).

From the Vista viewer and Synteny viewer it was clear that the conservation identity of *Streptococcus equi* were more than 70% and genome alignment were 80% similar to *Streptococcus pyogenes* genome. The Synteny viewer represents the origin of replication, gene density, Synteny location (where recombination occurs by crossing over) between *Streptococcus pyogenes* and *Streptococcus equi* chromosomes. It annotates the point where replication of gene occurs. The density of gene can be identified by the color intensity. Thus the virulence of target gene can be identified.

The Vista Dot plot describes the DNA conservation between *Streptococcus pyogenes M1 GAS* and *Streptococcus equi* subspecies equi *4047* genome assemblies at different levels of resolution and across multiple chromosomes/in descending orders by size. The DNA coordinates of *Streptococcus pyogenes M1 GAS* were present on the X-axis and *Streptococcus equi* subspecies equi *4047* genome were present on the Y-axis. The diagonal lines depict the homologous region between *Streptococcus pyogenes M1 GAS* and *Streptococcus equi* subspecies equi *4047* genome. The blue lines in the regions were on the same strand and red lines were on opposite strand of chromosomes. The grid in black lines displays chromosome boundaries. The cutoff control filters alignment to show syntenic region greater than specified length. The genome coordinates were represented in yellow box at the bottom left. The first set of coordinates corresponds to *Streptococcus pyogenes* and the second set of coordinates refers to *Streptococcus equi* genome. Right-clicking on the segment of the alignment corresponds to region of the alignment to *Strptococcus pyogenes* and *Streptococcus equi* 1852441 and 2253793bp respectively (Figure 3).

Table 1: Tools used

Vista Tools:	The VISTA tools are used for annotation of DNA through visualization of genome alignment and multiple comparative genome analysis for identification of functional elements, such as exons or enhancers which exhibit sequence similarity. Vista tools are used for full scaffold alignments, for locating genes, transcription factor motifs and other putative control regions between genomes.

Conclusion:

Comparative genomics of *Streptococcus pyogenes M1 GAS* and *Streptococcus equi* sps equi *4047* states that dnaA genes were involved in chromosomal replication. The comparative genome analysis of *Streptococcus pyogenes* and *Streptococcus equi* sps equi *4047* depict exon, Untranslated Region (UTR), conserved noncoding sequences (CNS), contigs, mobile genetic elements, Insertion sequence (IS) elements, transposons, plasmids and shows high degree of heterogeneity evidenced by the large number of single nucleotide polymorphisms (SNPs).

Thus UTR, CNS, and IS can be used as drug target.

Acknowledgement

The authors are deeply grateful to Sam Higginbottom Institute of Agriculture, Technology and Sciences for providing the necessary guidance, facilities and support required for completing the research work.

References:

[1] Bender, G.; Sutton, S.; and Marquis, R. (1986): Acid tolerance, proton permeabilities, and membrane ATPases of oral streptococci. *Infectitious Immunology*, (53): 331–338.

[2] Cunningham, M. W. (2000): In Gram-Positive Pathogens. *American Society of Microbiology*, 66–77.

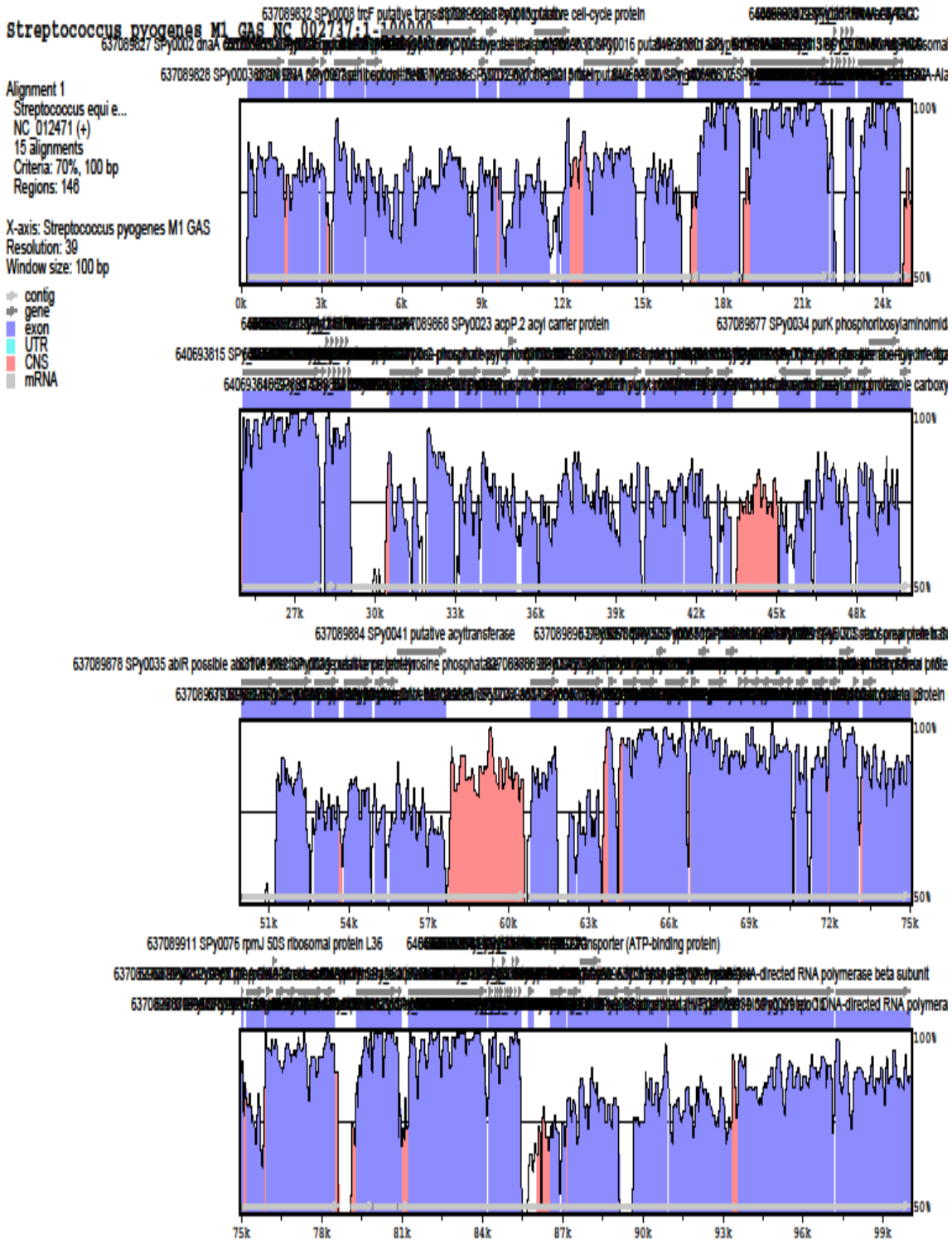

Figure 1: Alignment of *Streptococcus equi* sps equi *4047* and *Streptococcus pyogenes M1 GAS*

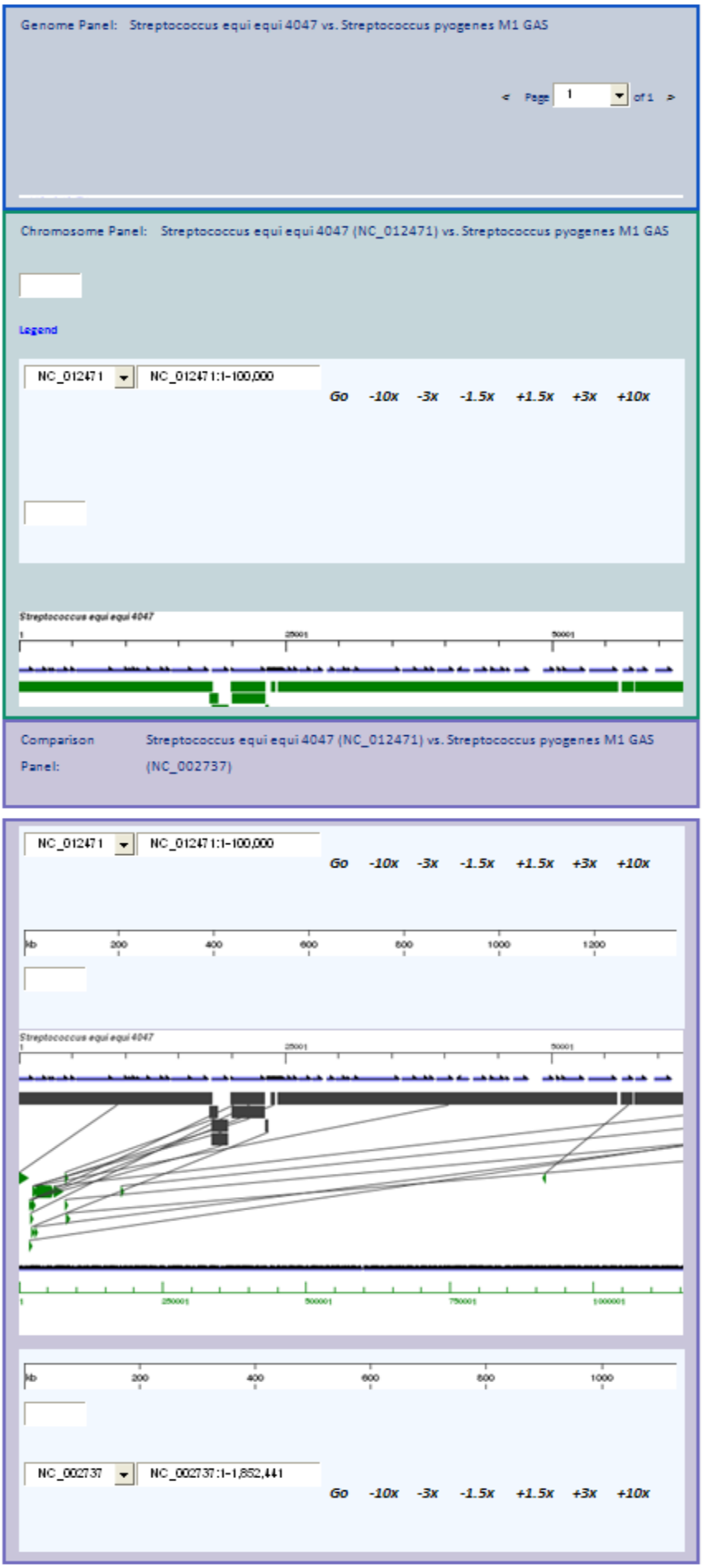

Figure 2: The genome, chromosome and comparison panel predict the Synteny at particular loci

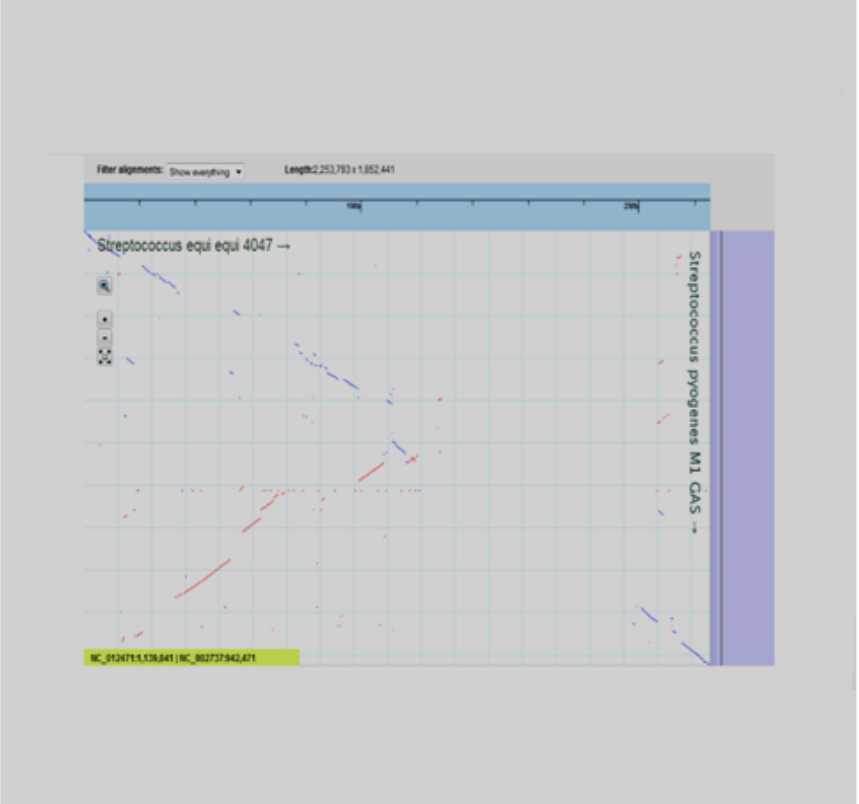

Figure 3: Dot plot represent the DNA conservation of *Streptococcus pyogenes* and *Streptococcus equi*

[3] Dubnau, D. (1993): In *Bacillus subtilis* and Other Gram-Positive Bacteria. *American Society of Microbiology*, 555–584.
[4] Ferretti, J.J.; Mc Shan, W.M.; Ajdic, D.; Savic, D.J.; Savic, G.; and Lyon, K. (2001): Complete genome sequence of an M1 Strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA*, (98): 4658-4663.
[5] Fischetti, V. A. (2000): In Gram-Positive Pathogens. *American Society of Microbiology*, 96–104.
[6] Foster, J. W. (2000): In Bacterial Stress Responses. *American Society of Microbiology*, 99–116.
[7] Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004): VISTA: computational tools for comparative genomics, Nucleic Acids Res., (32): 273-9.
[8] Gross, C. A. (1996): In *Escherichia coli* and *Salmonella:* Cellular and Molecular Biology. *American Society of Microbiology*, 1382–1399.
[9] Heyningen, T. V.; Fogg, G.; Yates, D.; Hanski, E.; and Caparon, M. G. (1993): Adherence and fibronectin binding are environmentally regulated in the group A Streptococci. *Mol Microbiol*, (9): 1213–1222.
[10] Hill, T. M. (1996): In *Escherichia coli and Salmonella:* Cellular and Molecular Biology, *American Society of Microbiology*. (2): 1602–1614.
[11] Koonin, E. V., Aravind, L. & Galperin, M. Y. (2000): In Bacterial Stress Responses, *American Society of Microbiology*. 417–444.
[12] Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A.; Alloni, G.; Azevedo, V.; Bertero, M.; Bessieres, P.; Bolotin, A.; and Borchet, S. (1997): The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature, London*, (390): 249–256.
[13] McIver, K. S.; Subbarao, S.; Kellner, E. M.; and Scott, J. R. (1996): Identification of isp, a locus encoding an immunogenic secreted protein conserved among group A Streptococci. *Infectitious Immunology*, (64): 2548–2555.
[14] Simpson, W.; Ragland, N.; Ronson, C.; and Tagg, J. (1995): A lantibiotic gene family widely distributed in *Streptococcus salivarius* and *Streptococcus pyogenes*. *Dev Biol Stand*, (85): 639–643.
[15] Snel(1999): Horizontal gene transfer. *Nature General*, (21):108.
[16] Stevens, D. L. (1995): Infectitious Disease. *Emergence*, (1): 69–78.
[17] Stevens, D. L.; Tanner, M. H.; Winship, J.; Swarts, R.; Ries, K. M.; Schlievert, P. M.; and Kaplan, E. (1989): Medicine. *Nature England Journal*, (321)**,** 1–7.
[18] Veasy, L. G.; Wiedmeier, S. E.; and Orsmond, G. S. (1987): Medicine. *Nature England Journal*, (316)**,** 421–427.

Sequence used:

The genomic sequence and protein sequence of dnaA gene product was retrieved from ncbi/uniprotkb.

SPY_0002_dnaA >gi|15674251|ref|NP_268424.1| chromosomal replication initiation protein [Streptococcus pyogenes M1 GAS]
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
KHIATIYLDQMKELFWEKNLKDVILTAGFEVYNAQISVD
YVFEEDLMIEQNQTKINQKPKQQALNSLPTVTSDLNSKY
SFENFIQGDENRWAVAASIAVANTPTTYNPLFIWGGPGL
GKTHLLNAIGNSVLLENPNARIKYITAENFINEFVIHIRLD
TMDELKEKFRNLDLLLIDDIQSLAKKTLSGTQEEFFNTFN
ALHNNNKQIVLTSDRTPDHLNDLEDRLVTRFKWGLTVNI
TPPDFETRVAILTNKIQEYNFIFPQDTIEYLAGQFDSNVRD
LEGALKDISLVANFKQIDTITVDIAAEAIRARKQDGPKMT
VIPIEEIQAQVGKFYGVTVKEIKATKRTQNIVLARQVAMF
LAREMTDNSLPKIGKEFGGRDHSTVLHAYNKIKNMISQD
ESLRIEIETIKNKIK
>gi|225869488|ref|YP_002745435.1| dnaA gene product [Streptococcus equi subsp. equi 4047]
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLIKVDN
HVATIFLDQMKELFWEKNLKDVILTAGFEVY
NAQIAVDYVYEDDLMIEQQHQGQQGYTEQAFQQLPAVQ
SDLNPKYSFDNFIQGDENRWAVAASIAVANTPGTTYNPL

FIWGGPGLGKTHLLNAIGNSVLLENPNARIKYITAENFINE
FVVHIRLDTMDELKEKFRNLDLLLIDDIQSLAKKTLSGTQ
EEFFNTFNALHNNNKQIVLTSDRTPDHLNDLEDRLVTRF
KWGLTVNITPPDFETRVAILTNKIQEYNFIFPQDTIEYLAG
QFDSNVRDLEGALKDISLVANFKQIDTITVDIAAEAIRAR
KQDGPKMTVIPIEEIQAQVGKFYGVTVKEIKATKRTQDIV
LARQVAMFLAREMTDNSLPKIGKEFGGRDHSTVLHAYN
KIKNMIGQDESLRIEIETIKNKIK