International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

GENOME ANNOTATION OF THE COMPLETE GENOME OF STREPTOCOCCUS MUTANS LJ23 SEROTYPE K BY IN-SILICO APPROACH

Shruti Kaushik1 and Rahul Anand*
1Department of Biotechnology; Madhav Institute of Technology and Science, Gwalior.
*Assistant Professor, Department of Biotechnology; Madhav Institute of Technology and Science, Gwalior

Abstract

The aim of high-quality annotation is to identify the key features of the genome in particular, the genes and their products. The tools and resources used for annotation are developing rapidly and the scientific community is becoming increasingly reliant on this information for all aspects of biological research. *S. mutans* is considered to be the main cause of dental caries and also cause bacteremia and infective endocarditis. In this study, we had analyzed the complete genome of *S. mutans* LJ23 serotype k through identification and prediction of the functional regions & regulatory elements in promoters by two automated methods of annotation using GLIMMER (HMM and IMM algorithm) and the second is MEME to develop a complete set of domains and motifs that are manually analyzed by *in-silico* approach. Furthermore, genes basically non-coding regions and lactose operon genes were demonstrated in the genome of *Streptococcus mutans* LJ23. The distribution of new motifs prevelant in putative promoter were also defined in the complete genome of serotype k *S.mutans* LJ23.

Background

Streptococcus mutans is gram-positive cocci, anaerobic, acidogenic, chemo-organotrophic and aciduric bacterium commonly found in the human oral cavity and is a significant contributor to tooth decay and plaque (Toda *et al*., 1987). *Streptococcus* is a genus of spherical Gram-positive bacteria belonging to the phylum Firmicutes and the lactic acid bacteria group [Loesche *et al*., 1986]. *S. mutans* is widely recognized as the main etiological agent of dental cavities. The complete genome sequence of *S. mutans* LJ23 was deposited in the GenBank databases under accession no. AP012336. *S. mutans* is composed of circular DNA that consist of plasmids of 5.6 kilobase(kb). These plasmids play an important role in *S. mutans* because of their functions that includes bacteriocin production and immunity, accessory catabolic pathways and mechanisms for conjugation-like transfer activities (Hamada *et al*., 1986). The complete genome sequence of serotype k *S. mutans* strain LJ23 was recently isolated from the oral cavity of a Japanese patient (Aikawa *et al*, 2012). *Streptococcus mutans* is a major pathogen of dental caries and is classified into serotypes c,e,f, and k. Aikawa *et al*, also demonstrated the genome of *S.mutans* LJ23 that contains a single circular chromosome having 2,015,626 bp length. After the genome of an organism is sequenced and assembled, comprehensive and accurate initial gene prediction and annotation by computational analysis have become the necessary first step towards understanding of the functional content of the genome. The elements of the annotation process are gene finding, homology searches, functional assignment, ORF management and data availability. Gene annotation provided by Ensembl includes both automatic annotation and manual annotation that includes genome-wide determination of transcripts and reviewed determination of transcripts on a case-by-case basis. The "unit" of genome annotation is the description of an individual gene and its protein (or RNA) product, and the focal point of each such record is the function assigned to the gene product.

Basically, we focus on the Structural Annotation that helps us in finding the genes in genomic DNA. Here, two main types of data used in defining and annotating the gene structure:

- Prediction based algorithms are focussed to find genes/gene structures based on nucleotide sequence and composition.
- Sequence similarity (DNA and protein): alignment to mRNA sequences (ESTs) and proteins from the same species or related species; identification of domains and motifs.

The complete structure of mRNA can be derived from sequence alignment of full- length cDNA with the genome. ESTs can be sequenced easily but contain only partial 5'-3' end structures. Computational prediction is now very useful for finding possible targets such as transcript structure and transcription factor binding sites (TFBSs). Coding regions are more conserved than non coding regions so conserved regions are important part for functional elements in the genome, comparison between all types of genome annotations will be useful, especially for target screening before analysis.

Gene identification and prediction programs can be divided into two categories: an empirical category which are based on sequence similarity and *ab-initio* which uses signal and content sensors. Empirical method search similarity in the genome; they identify genes based on homologies with known database, such as genome DNA, cDNA, dbEST and proteins. Gene identification is the most dynamic stage of process as new algorithm are developed and more database become

available that frequently enhance the annotation process. Gene identification can be characterised by taking individual homology search matches and also *ab-initio* computational prediction; aligning them to a particular genomic sequence and then making prediction of genes structure. Gene identification method can be differentiated as *ab- initio* method and consensus method, depending on whether they need to be trained on a set of genes in an order to evaluate whether a query sequence is coding or not. Relatively; few non-consensus method of gene identification are either on universal measure for differentiating between coding and non-coding DNA or on some self-consistent model of gene structure. In the case of putative genes, the genes identified can be indicative of orfs which are ultimately non-coding. Hence, the identification of promoters using *in-silico* approach is very important for improving genome annotation and understanding transcriptional regulation.

Methodology

Working platform & data set

The platform is one of the important parameter for the bioinformatics project. Windows 7 operating system and Linux Kernel version 3.02 installed with 2.00 GB RAM & GNU Bourne Again Shell static version 4.2-2 Ubuntu 12.04 was used for *in-silico* approach. The complete genome sequence of *S. mutans* LJ23 was downloaded from GenBank database under accession no.AP012336**.**

A. Static view of annotation by GLIMMER version 3.02

The FASTA genome sequence of *S. mutans* LJ23 was uploaded in the NCBI-NIH GLIMMER webpage. By briefly focussing on Markov models in the context of DNA sequence analysis the GLIMMER version 3.02 system was used to identify regions that are likely to be gene that consists of two programs - *build-imm* takes an input sequences and builds and outputs the Interpolated Markov Model for them (sequence and partial orf) (Delcher *et al*., 1999). The methods and algorithms of GLIMMER generally use interpolated context model, Markov model and resolving overlapping genes (Salzberg *et al*., 1998).

B. Protein level annotation nBLAST:

The FASTA genome sequence was retrieved from GenBank database of the different organisms such as *S.pyogenes, S.pneumoniae, S.agalactiae, S.salivarius and S. thermophilus.* Further, the location was determined using sequence comparison of various strains by nBLAST using a protein query (Altschul *et al*., 1990). Only the results with the low E-value and high score were selected.

C. Dynamic view of annotation MEME:

The -40 and +15 regions of positive and negative frame GLIMMER putative ORFs were separated on Dotnet platform on-line. The positive frames were uploaded in the Multiple EM for motif Elicitation tool web-page for discovering and analysis of functional motifs. Through the same web server, users can also access the motif alignments and search tool to search sequence database for matches to motifs encoded in several formats (Bailey et al., 2009). MEME also discovered the binding sites for shared transcription factor in set of promoter.

MAST (Motif alignment and search tool) search nucleotide database with protein motifs of the query nucleotide sequences. Here, non-redundant and upstream database was selected as a supported database category for the *S. mutans.* (Bailey *et al.*,1998). Then we selected the MEME putative promoter (motifs) and compared with a promoter database search using MAST.

5s ribosomal RNA database

5s ribosomal RNA is an integral component of the large subunit of all cytoplasmic and most organellar ribosomes. The database is available on line through the World Wide Web at http://biobases.ibch.poznan.pl/5SData/. The sequences for *S. mutans* were retrieved as single files using a taxonomic browser or in multiple sequence structural alignments. Using the various 5S RNA sequence of the different species of *Streptococcus* as Query, the respective location were searched in the genome of interest (GenBank AP011236.1) using nBLAST search [database nucleotide collection (nr/nt)] and organisms N.C.B.I. TaxoID 1309 of *Streptococcus mutans.* Similarly, location of 16s and 23s rRNA sequences was analyzed by finding various queries from GenBank by tBLASTn.

GC profile:

GC-Profile provides a quantitative and qualitative view of genome organization. It shows that GC-Profile would be an appropriate starting point for analyzing the isochore structure of higher prokaryotes genomes, and an intuitive tool for identifying genomic islands in prokaryotic genomes. GC-Profile is freely available at the website http://tubic.tju.edu.cn/GC-Profile/. (Gao and Zang, 2006).

Results and discussion

Identification of genes using GLIMMER version 3.02

GLIMMER showed a total of 1962 orfs in the genome of *Streptococcus mutans* LJ23. The GLIMMER also found the start, end location, frame along with their score of the predicted orfs. The maximum length of the predicted genes was found to be 17124 bp. The total numbers of 3 frames were found using GLIMMER i.e. +1, +2, +3, -1,-2 and -3. The average of the orfs score was found to be 8.67875. We subtracted and added -40 & +15 to the start region of the orfs of genome of *Streptococcus mutans* LJ23. Finally, the Dot net platform was used to find location and the sequences of the orfs in the genome of *Streptococcus mutans*. The total number of positive frames was found to be 1036 and the negative frames were found to be 925. The positive and negative set of the sequences were separated manually. Using only the genome sequence as input, a training set of orfs that were greater than 500nt were selected from the GLIMMER and the resulting IMM model was then compared to the annotated set of genes identified for *S. mutans* manually. 1958 of 1962 genes found to be correctly identified, while some of them were eliminated by removing those that conflict with rRNA and tRNA. This implies that minimal false result negative rate of 0.44% for GLIMMER.

Table1: Results of Open reading frames of the predicted genes using GLIMMER

Frames	No. of ORF
Total no. of +1 frames	344
Total no. of +2 frames	351
Total no. of +3 frames	341
Total no. of -1 frames	326
Total no. of -2 frames	306
Total no. of -3 frames	293

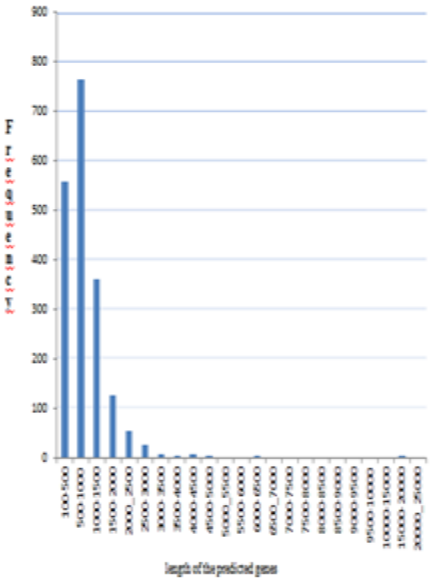
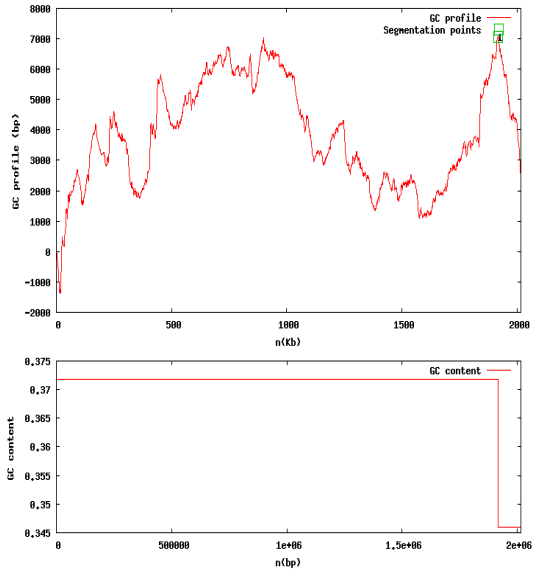


Figure 1: Graphical representation of the length of the predicted genes in *S. mutans* LJ23 by GLIMMER.

Evaluation of the GC content in the genome of the Streptococcus mutans LJ23:

The total GC content was found to be 37.1% by using GC profile. The remaining AT content was found to be 62.9%. There is an increase (slope=0.07) in the fraction of annotated genes that are predicted, with the growing GC content. There is an increase in the fraction of annotated genes that was predicted by GLIMMER lying in the range 1500-2000 with the growing GC content. By default GC- profile generates figure file for each job representing the distribution of GC content in the genome of *Streptococcus mutans.*

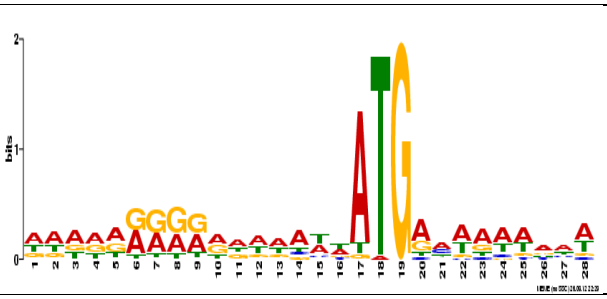
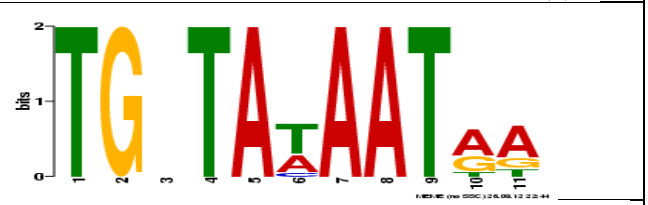
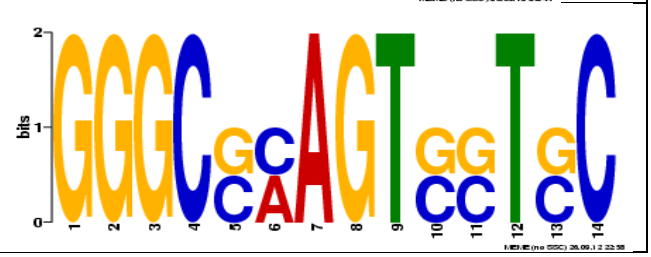
Figure 2: Graphical representation of the GC content in the complete genome of *S.mutans* LJ23 by GC profile:



Determination of putative motifs of the positive frame of the Streptococcus mutans LJ23 serotypes K using MEME Suite.

MEME output contains sequence LOGOS for each discovered motif, as well as buttons to allow motifs to be conveniently submitted to the sequence and motif database scanning algorithms (MAST and FIMO), for further analysis. The motif 1 consists of 1.2e-833, 28 width and 1002 sites of the conserved motifs. MEME's hypertext (HTML) output also contains buttons that allow for the convenient use of the motifs in other searches. The motif 2 consists of E-value 6.0e – 009 with width of 11 and its sites are 76. The MEME output is HTML and shows the motifs as local multiple alignments of (subsets of) the input sequences, as well as in several other format. The best match of the motifs are selected and finally analyzed according to their E- value. We found three best motifs along with their width and sites.

Table 2. MEME results of the positive set of sequences of the *Streptococcus mutans* LJ23.

Motifs	E value	Width	Sites	Sequence logos
Motif 1	1.2e-833	28	1002	
Motif 2	6.0e-009	11	76	
Motif3	6.9e+005	14	2	

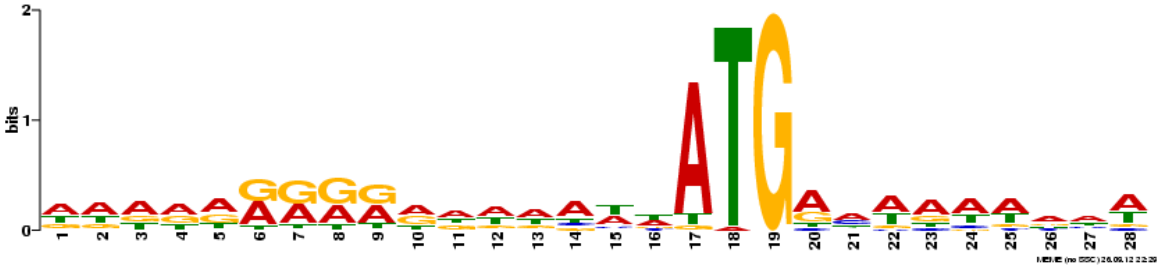


Figure 3: LOGO of conserved motif. LOGOS are a visualization tool for motifs. The height of a letter indicates its relative frequency at the given position(x-axis) in the motif

The sequence logos represents the conserved region was found to be between 16 to 20 bp. The most appropriate region of the conserved motif was found in the Motif 1 of the total motifs shown by the MEME tool.

Comparing DNA motifs with the known regulatory motifs using MAST tool:

The sequence that would achieve the best possible match score and its reverse complement for nucleotide motifs are considered best motif. So finally, motif1 is considered best motif with the 28 residue width. The search results showed top scoring sequences with tiling of all of the motifs matches shown for each of the sequences. MEME putative promoter (motifs) compared with a promoter database search using MAST showed three similarity searches the Motif 1 consist of 0.39 and 0.12. The Motif 2 showed the score 0.39 and 0.28 and the motif 3 represent the score 0.12 and 0.28

Table 3. Identification of the putative motif alignment of the DNA sequences of the *S.mutans* LJ23 serotype K with the MAST search tool

Motif	Width	Best possible match	Similarity	
		(+)	(-)	
Motif1	28	AAGGGGGGGGGAAATTATGGCAAAAGCA	TGCTTTTGCCATAATTTCCCCCCCCCTT	-
Motif 2	11	TGCTATAATGA	TCATTATAGCA	0.39
Motif 3	14	GGGCCCAGTCCTCC	GGAGGACTGGGCCC	0.12

Analysis of the non-coding regions in the genome of Streptococcus mutans LJ23:

The 5s rRNA sequences of the different species of *Streptococcus* as Query was identified using 5S rRNA database the respective location was found in the genome of interest using nBLAST and organisms NCBI TaxoID 1309 of *Streptococcus* was selected(Table 4)

Table 4. Identification of 5s non - coding regions of the *S. mutans* LJ23 using 5s ribosomal RNA Database.

Organisms	E value	Max /total score	Location beg	Location end
Streptococcus mutans LJ23 Query Sequence1.	1e-55	209/1048	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834670
Query Sequence2.	6e-54	204/1021	21910 231975 407759 437758 1834785	22022 232087 407872 437818 1834725
Query sequence3.	3e-57	215/1076	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834670
Streptococcus pneumonia Query sequence1	2e-43	169/845	21911 231976 407760 437759 1834784	22025 232090 407874 437873 1834670
Query sequence2	7e-44	171/855	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834725
Streptococcus pyrogenes Query sequence 1.	7e-44	171/855	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834670
Query sequence2.	7e-44	171/855	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834670
Query sequence 3.	1e-45	176/882	21910 231975 407759 437758 1834785	22025 232090 407874 437873 1834725

Identification of the location of 16s ribosomal RNA genes in the Streptococcus mutans LJ23:

We first examined the gene locations for ribosomal RNA such as 5s, 16s and 23s in *S.mutans*. The comparison was done amongst the closest Streptococci family and significant results were obtained in both 5s and 16s but no significant hits were found in 23s non-coding regions (rRNA). The location of 16srRNA was almost similar in *S. pyrogene* and *S. pneumonia* i.e. from 21910 to 22025. The results were obtained with *Streptococcus mutans*, *S. pneumonia* and *S. pyrogenes*. The E-value for the various query sequences was found to be less than 1e-55.

Table 5. Identification of the 16s non – coding regions of the *Streptococcus mutans* LJ23 using nBLAST:

Organisms	E value	Max score	Total score	Location beg	Location end
Streptococcus mutans LJ 23(a)	1e-55	209	1048	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437759	437873
	-	-	-	1834785	1834670
Streptococcus mutans LJ 23(b)	6e-54	204	1021	21910	22022
	-	-	-	231975	232087
	-	-	-	407759	407871
	-	-	-	437758	437870
	-	-	-	1834673	1834785
Streptococcus mutans LJ23(c)	3e-57	215	1076	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	407874
	-	-	-	437758	437873
	-	-	-	1834785	1834670
Streptococcus pneumonia	2e-43	169	845	21911	22025
	-	-	-	231976	232090
	-	-	-	407760	407874
	-	-	-	437759	437873
	-	-	-	1834670	1834784
S.pneumoniae (b)	7e-44	171	855	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	437873
	-	-	-	18347670	1834785
Streptococcus pyrogenes(a)	7e-44	171	855	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	437873
	-	-	-	1834670	1834785
S.pyrogenes (b)	7e-44	171	855	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	437873
	-	-	-	1834670	1834785
S.pyrogene(c)	1e-45	176	882	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	437873
	-	-	-	1834670	1834785
S. pyrogenes (d)	1-e-45	176	882	21910	22025
	-	-	-	231975	232090
	-	-	-	407759	407874
	-	-	-	437758	437873
	-	-	-	1834670	1834785

Identification of lactose operon in S. mutans LJ23:

By nBLAST, location of upstream and downstream regions was found in the LacR and LacA genes of S. mutans LJ23 as shown in Table 7. The lactose operon starts from Lac R and end to Lac E. No termination codon was found in the LacD gene. Various mutations were found in the complete genome sequence of S. mutans LJ23 when compared with S. mutans lactose operon (M80797.1) as mentioned in remarks.

Table 7: Identification of the loci of streptococcus mutans lac operon (M80797.1**) against the genome of *Streptococcus mutans* LJ23 Serotype K (AP012336.1):**

Genes	-35 region	-10 region	Ribosomal binding sites (RBS)	Coding sequence (CDS)	Strand	Remarks
Gene1 (lacR)	665094-665099	665119-665124	665142-665153	665157-665912	Plus	T instead of C at 666011 G instead of A at 665210 A instead of G at 665423 A instead of G at 665459 C instead of T at 665478 A instead of G at 665489 A instead of G at 665555 T instead of A at 665663 C instead of G at 665774 T instead of C at 665780 C instead of G at 665829
Gene2 (lacA)	666129-666134	666154-666159	666205-666210	666218-666646	Plus	G instead of A at 666153 C instead of T at 666289 A instead of G at 666367
Gene3 (lacB)	-	-	666655-666664	666672-667187	Plus	T instead of G at 666702 G instead of T at 666703 G instead of T at 666704 A instead of G at 666735 G instead of T at 666922
Gene4 (lacC)	-	-	667187-667195	667206-668138	Plus	T instead of C at 667250
Gene5 (lacD)	-	-	**668129**-668134	668143-669120 (**669116**)	Plus	No termination codon found in our sequence C instead of T at 668529 C instead of T at 66835 A instead of G at 668708 G instead of A at 668860 G instead of A at 668946 G instead of A at 669019 G instead of A at 669109
Gene6 (lacF)	-	-	671366-671383	671400-671714	Plus	C instead of T at 67160
Gene7 (lacE)	-	-	671704-671710	671721-672088	Plus	T instead of C at671790 C instead of T at 671807 C instead of T at 671834

Conclusion:

Interpretation of raw DNA and amino acid sequence data involves the identification and annotation of genes, proteins, and regulatory and/or metabolic pathways and hence by this method we can perform improved and better characterization and categorisation of domains and families of protein sequence towards an understanding the biology of S. mutans. This process is typically performed using sequence annotation pipelines (i.e. a variety of software modules) and, in some cases, human expertise to handle the annotations generated automatically. The reference databases, computational methods and knowledge that form the basis of these pipelines are constantly being developed. Manual analysis is incalculably time-consuming activity so here we focussed on in-silico approach with excellent result and in order to concentrate on potentially the most interesting domain families. In addition, the rapid increase in new sequence data has necessitated the evolution of software resources from functional annotation of a single genome towards simultaneous analysis of information from multiple genomes. Also in this paper, we present a procedure for annotating core promoter in S. mutans by two step process of ORFs selection and computational prediction. A first analysis of motifs present in putative promoter have a consensus TATA box so finally, we conclude that there are relatively few recognizable binding sites for known transcription factor in S.mutans putative promoter while our analysis showed previously underappreciated motifs that are distinct feature in *S. mutans* promoter regions.

References

1. Altschul, S; Gish, W; Miller, W; Myers, E; Lipman, D (October 1990). BLAST:"Basic local alignment search tool". Journal of Molecular Biology 215 (3): 403–410.
2. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, et al.(1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
3. Bailey L. T. and Gribskov M. 1998 "Combining evidence using p-values: application to sequence homology searches", Bioinformatics, 14(1):48-54.
4. Chihiro Aikawa, Nayuta Furukawa, Takayasu Watanabe,Kana Minegishi, Asuka Furukawa, Yoshinobu Eishi, Kenshiro Oshima, Ken Kurokawa, Masahira Hattori, Kazuhiko Nakano, Fumito Maruyama,Ichiro Nakagawa, and Takashi Ooshimae.2012. Complete Genome Sequence of the Serotype k Streptococcus mutans Strain LJ23, Journal of bacteriology.
5. Charles E. Grant, Timothy L. Bailey, and William Stafford Noble 2011 FIMO: Scanning for occurrences of a given motif, Bioinformatics 27(7):1017–1018
6. Clarke, J. Kilian (1924). "On the Bacterial Factor in the Ætiology of DentalCaries".British Journal of Experimental Pathology 5: 141–7. PMC 2047899.
7. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.1999. Improved microbial gene identification with GLIMMER, Nucleic Acids Research, 4636-41.
8. Feng Gao and Chun-Ting Zhang. 2006. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic acid, 686-691.
9. Hamada Shigeyuki, Michalek M. Suzanne, et al. 1986. "Molecular Microbiology and Immunobiology of Streptococcus mutans." New York: Elsevier Science.
10. Szymanski M., Barciszewska M. Z., Erdmann V. A., Barciszewski J. 2002 5S ribosomal RNA database. 30 (1): 176-178
11. Loesche WJ (1986) Role of Streptococcus mutans in human dental decay. Microbiol Rev 50: 353-380.
12. Salzberg L. S., Delcher L. A., Kasif S. and White O. 1998. Microbial gene identification using Interpolated Markov Models.26:544-548.
13. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME:Suite: tools for motif discovery and searching, Nucleic acid research.