Int J Pharm Bio Sci 2013 Oct; 4(4): (B) 923 - 931

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

ANNOTATION OF NON-CODING REGIONS OF *STREPTOCOCCI* FAMILY: A CASE STUDY

ONDARI NYAKUNDI ERICK*

Department of Biotechnology, Karpagam University, Coimbatore 641 021, India.

ABSTRACT

Non-coding functional elements recently have proved to play vital functions in eukaryotes worthy ignoring in the prokaryotic counter parts. We argue that elucidating these non-coding elements will lead to a better understanding of the differences co-existing between the *Streptococci* species, an important human pathogen responsible for vast diseases devastating the economy. The study analyzed 56 strains from NCBI, where 44,523 sequences were extracted by PERL script. Similarity search by Blast, Genemark and CPC revealed a total of 1443 potential sequences preceding functional studies by deploying Pfam, InterproScan and COG depicting 144 proteins allowing them to be designated as novel ones. The results potentially could be used as new vaccine targets, better understanding of the different biological niches; re-annotation also can be looked unto for potential genomic island identification and further wet-lab extensions research work for better understanding of the *Streptococci* family for further directions and incorporation of the identified proteins into database.

KEYWORDS: Comparative genomics, functional annotation, non-coding regions (junk DNA) and *Streptococci* spp.

ONDARI NYAKUNDI ERICK
Department of Biotechnology, Karpagam University, Coimbatore 641 021, India.

Corresponding author

INTRODUCTION

Genomics has played a vital role towards unveiling functional and structural studies. Annotation can be defined as the progression by which structural or functional information is deduced for genes or proteins regarding the similarity to previously characterized sequences in public databases for solving biological tribulations. The annotation process links genome sequences with functional information and guides experimentation by relating genotypes to phenotypic properties. Re-annotation is where the already annotated genome sequence is annotated again paving way for discovery of more genes and protein functions, testing and performance-comparison of existing or newly developed annotation methods, and assessment of annotation reproducibility and significantly re-annotation also provides up to-date information for end-users, using the latest resources such as new or improved algorithms and richer databases [1]. A lot of research is currently going on to shade light on the hidden functions of non-coding regions (ncRNA) exhibited by more than 98% of the total RNA [2,3,4] in eukaryotes as they were thought to have no known function and thus termed as 'junk DNA' until eukaryotic studies proved ncRNAs to play role in gene regulation and RNA processing. In the human genome, it has been discovered that many non-coding transcripts have roles in gene regulation and RNA processing. Cross-species sequence comparisons have identified conserved non-coding elements (NCEs) that are candidates for various functions. SNPs in many non-coding regions have been linked to disease in Genome-Wide Association (GWA) studies. Non-coding DNA also provides a historical record of genome evolution, as it contains 'fossils' of molecules that were historically active [5].

The various non-coding elements found within the genome are short repeats, regulatory factor binding regions, small RNAs, broad histone marks, transcripts, transposable elements, pseudogenes, regulatory elements, segmental duplications and structural variants. Comparative genomics has found non-coding elements (NCEs) that are conserved to varying degrees across mammalian or vertebrate genomes, which suggest some functions conserved by natural selection [6]. Regulatory elements to which transcription factors bind include promoters, enhancers, silencers, insulators and locus-control regions (LCRs). Non-coding RNAs (ncRNAs) such as snRNAs, snoRNAs and microRNAs play important roles in transcription and translation control and work by Jayavel [7] and Stortz et al., 2005 [8] pointed vital roles in transcriptional regulation, chromosome replication, RNA processing and modification, mRNA stability, protein degradation and translocation. The regulatory RNAs, whose size lies between the ranges of 50-550 nucleotides, interact with various specific proteins (Christophe and Brice 2007) [9]. They are known to have intrinsic dynamic structures upon binding with other proteins whereby a biological response is initiated. The binding may yield base pairing with the help of specific proteins throughout the entire interaction process i.e. before, during and after and also formation of ribonucleoprotein complexes which facilitate recognition between the two interacting RNAs. Recognition may result loosening of their structures, induction of specific hydrolysis and degradation of the non-coding RNAs, thereby blocking their regulatory functions. Quite a good number of small RNAs have been identified by the earlier researchers in *Escherichia coli, Stapylococcus aureus* [10]; *Listeria monocytogenes* [11] among others with their possible outcome e.g pairing enhances translational inhibition [5], stimulation [12], stabilization [13] of mRNA and degradation [7]. Ribonucleoprotein complexes may either be active or inactive due to formation of intricate scaffolds with specific binding sites for the associated proteins. For example 4.5 sRNA has been reported as active in absence of SmpB (small protein B), RNAse P is active in absence of C5 protein which is vital for *invivo* activity pointing that many sRNAs with their associated proteins are important for any biological

molecular activity. Re-annotation projects for individual species have been reported in the literature by a handful of groups and includes: *Haemophilus influenzae, Mycoplasma genitalium, Methanococcus jannaschii, Mycoplasma pneumoniae, Chlamydia trachomatis, Thermotoga maritima, Saccharomyces cerevisiae, Plasmodium falciparum* (chromosome II) [15], *Aeropyrum pernix*, various archaeal species and isolated cases of single genes. The present research work focus on the examination of the *Streptococci* family in order to identify any potential non-coding regions which perhaps could be responsible for disease progression, environmental adaptation, diagnostics and therapeutic targets by annotation studies.

MATERIALS AND METHODOLOGY

Figure 1 Overview of Methodology

Sequence retrieval from NCBI

To analyze the non-coding sequences from the *Streptococci* family, the .ptt (contains functional analysis information) and .fna (contains whole genome sequence nucleotides) files of all the 56 strains were downloaded from NCBI FTP site (http://www.ncbi.nlm.nih.gov/Ftp/).

Extraction of non-coding sequences

Perl script was deployed to extract the sequences for easier extraction of the individual non-coding nucleotides since its' difficult for one to manually count the individual nucleotides as analyzed from the functional notation .ptt file downloaded from NCBI FTP site. The file defines the starting digit of the coding sequences and end sequence making it simpler for non-coding sequence extraction. Sequences with more than 100 nucleotide bases were considered for further analysis to enhance high confidence of the findings.

Sequence similarity search-Blast, Genemark & Coding potential calculator

The extracted sequences were converted to fasta format for similarity search by (1) Blast version 2.2.8 (2), genemark 2.8 and (3) CPC. Blast enables identify the regions of local similarity between sequences by comparing either the nucleotide or proteins datasets and returns the statistical meaningful matches and as well guarantees categorization of genes into their respective families and deduce the functional and evolutionary relationships (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Genemark provides prediction models for a total of 313 both bacteria and archaea complete genomes. The parent organism was used for prediction and if it was found missing a family of *Streptococci* was considered with the largest genome size to allow whole genome coverage. The prediction modes were selected such that the genes present in both direct and complementary strands of the query sequence were predicted (http://exon.gatech.edu/gmhmm2_prok.cgi). To determine the protein coding potential of the extracted sequences, genemark which uses Markov model for both coding and non-coding parameters thereby allowing demarcate local variations and coding potential (http://exon.gatech.edu). CPC was used in order to discern the findings of Blast and genemark owing to its capacity of pointing protein coding RNAs from non-coding RNAs based on six biologically significant features [16] (http://cpc.cbi.pku.edu.cn). The CPC server also guarantees extraction of sequences features and additional annotations of the transcripts may facilitate further investigation. matching non-coding regions of the nucleotide sequences against protein databases were carried out routinely during all genome sequencing projects to eliminate inconsistencies and false positive (or negative) ORF as suggested earlier by Ouzonis *et al*. 2002 [17].

Functional analysis

In order to elucidate the protein functions of the selected sequences, we converted the DNA to protein sequences using StarORF package where the putative ORF proteins with 150bp length were considered [18]. StarORF just like the central dogma, transcribes DNA sequence into RNA which is translated into potential Open

Reading Frames (ORF) encoded within each of the six translation frames (3 in forward and 3 in reverse direction). Overlapping proteins were designated based on their domains or the functional regions using Pfam, InterProScan and COG tools. Pfam provides a large collection of protein families represented by multiple sequence alignment (MSA) and Hidden Markov Models (HMMs) http://web.mit.edu/star/orf/index.html [19], InterProScan (tool that combines different protein signature recognition methods onto one resource) and Clustered Orthologous Groups (COG) (presents individual or group protein families (paralogs) from major phylogenetic lineages comprised of whole genomes) [20]. The novel proteins identified were cross checked based on all the three programs with the .ptt files containing the coding information and the proteins which were found present were consequently selected. Based on Kong *et al.* 2007 [16] recommendations for non-coding protein coding capacity by their resultant score < -1, non-coding; from 1 to 0, weakly non-coding; from 0 to 1, weakly coding; and >1, coding, sequences which only appeared on Blast, Genemark and CPC with >1 were selected to precede functional analysis.

RESULTS AND DISCUSSIONS

Retrieval of non-coding regions

The study covered 56 *Streptococci* strains whose complete genomes thus far are publically available from NCBI FTP site (http://www.ncbi.nlm.nih.gov/Ftp/) as shown in table I (footnotes). A total of 44,523 sequences were extracted by PERL script to precede functional studies.

Similarity search

Out of the 44,523 sequences extracted, a total of 1443 sequences were found to code for potential coding from the three similarity searches i.e. Blast, Genemark and CPC and the same were employed for functional studies.

Functional studies

A total of 44,523 sequences were obtained from the sample set of 56 *Streptococci* strains for analysis. Blast 2.2.8, genemark 2.8 and CPC identified 1443 which were found to harbor coding regions showing significant similarities with proteins sequences from *Streptococci* and other bacterial species. High degree of similarities vindicated known and hypothetical proteins from respective and more closely related species. Only in very few cases, similarities were observed with proteins from less closely related genomes. *Streptococcus parauberis* and *Streptococci suis* are more prevalent as shown in table I and revealed important gene DapB coding for dihydrodipicolinate reductase responsible for L-lysine and meso-diaminopimelate which are crucial for protein and cell wall's peptidoglycan bacterial synthesis in both Gram-positive and Gram-negative and disruption of diaminopimelate biosynthesis in *mycobacteria* lead to cell death [21]. The choice of cell wall as the major target towards the rise of antimicrobial resistance has for long been studied and DapB has been indicated as a potential drug candidate. Other various important proteins include sortase together with fibronectin binding protein which aids pili attachment to the cell wall, additionally sortases act as housekeeping by playing both roles as a protease and transpeptidase. Though sortase are credited for new antibiotics, little interest for commercialization has been reinforced [14, 22, 23, 24, 25, 26]. LicD has been suggested to possess low choline uptake and its potency is low [27, 28], indicating that LicD manipulation could help manage the streptococcal spread. *Streptococci* family contains extracellular toxins which are commonly referred to as SPEs. TENA which enhance expression of extracellular enzymes was vitally observed and maybe this could be responsible for immune disguise of its antigenic properties as well as THI-4 and PQQC for thiamine and pyrroliquinoline synthesis were noted respectively. MATE (multiple antimicrobial extrusion protein) is responsible for metabolic and xeno-biotic translocation and is known for its resistance. As well, Fic\Doc which is implicated

in cell division and PAB or folate synthesis, Doc as a reverse growth inducer in *E.coli* studies indicated that C-terminal antitoxin partner Phd (prevents host death) through fold complementation in disordered solutions [29]. In the long run of identifying potential drugs, SMC which facilitate chromatin and DNA dynamics, RecF and RecN which play roles in DNA metabolism and recombination [30] should be researched as inhibition of DNA deprives microbes an opportunity to live. Branched chain amino acid which has been indicated to facilitate 4-azaleucine resistance by Belitskey *et al* 1997 [31] was well noted, however little information was available. As emergency of antibiotics resistance is on the rise worldwide, comprehensive research should be carried out for development of more effective therapeutics to contain this contagious *streptococci* family devastating the economy and rendering not only mankind weak immunity but also huge losses to swine industry and marine organisms. The analysis showed high similarity to other various molecular proteins such as Transposase, Integrase, Glycosyltransferase, ABC transporters, DNA methylases, oxidoreductases, short chain dehydrogenase, MeoR, Transketolase, UDP-glucose dehydrogenase with few as representatives. Most of the predicted novel coding regions in all the species in the present study showed significant similarity with transposase and integrase among others showing that these copies of genes would have been missed during annotation as evidenced by the findings than many non-coding regions possess higher probabilities of being coding segments. Further, it has been suggested that when disruptions occur in genes that are no longer required, the non-functional regions can be maintained in the genome, however, gradually they are eroded and thus eliminated. Arguably, if these non-coding regions predicted as novel are not functionally valuable, they would have equally been eroded for eventual elimination from the genome in due process of evolution e.g. 6S RNA which has been detected in many species suggest that the 6S RNA-RNAP interaction was maintained as the bacterial evolution was taking place for efficient gene promoter switches during growth [10, 32]. The fact that these sequences still remain conserved, they exhibit a higher probability of being potential coding segments. Hence, we cannot ignore revising the re-annotation by adopting high throughput technologies such as RNA-seq, next generation DNA sequencing and tilling arrays to conceal the functional elements identified for further directions especially towards drug development, diagnostics and better understanding of the different biological niches as synchronized by the molecular machines to curb enormous *streptococcal* infections.

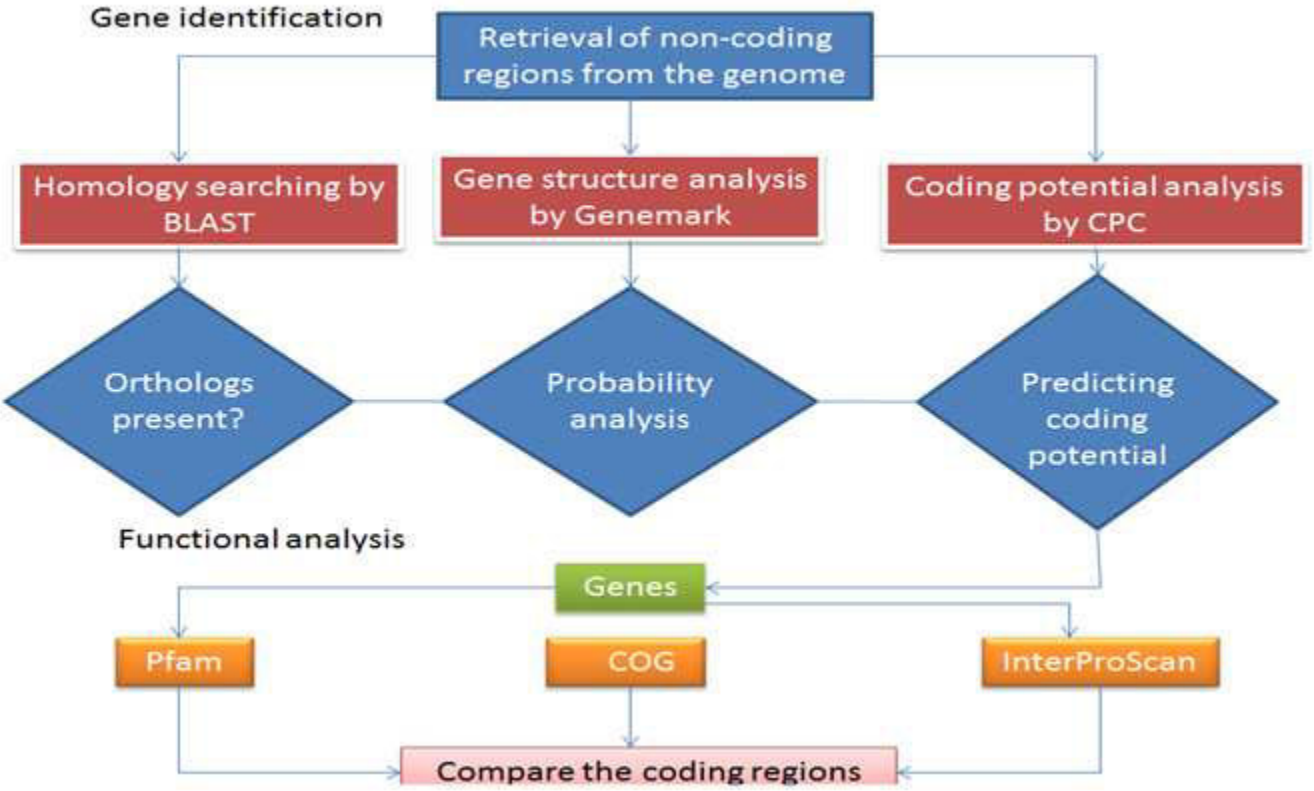
Table I
List of all the identified genes and proteins

Species *	Gene/ Protein	Function(s)
Streptococcus pyogenes (13)	AICARFT/IMPHase	Purine synthesis
	Sortase	Pili assembly and attachment
	BCCT	Protein transporter
	PBPs	Attraction of β-lactams
	Short chain dehydrogenase	Substrate specificity
	Aldo/keto reductase	Oxidoreductase activity
	Binding-protein-dependent	Transport system
	GHMP	ATP binding
	Fibronectin binding protein	Virulence factor attachment
	Phosphoribosyl transferase	Phosphoryl transfer
	Formyl transferase	N-formylmethionyl-tRNA identity
	Aldehyde dehydrogenase	Aliphatic and aromatic aldehydes oxidation
	Glycosyl transferase	Sugar biosynthesis
	Amino acid permease	Amino acid transportation
	Major facilitator family	Transport of small molecules
	Streptococcal protein	Human complement inhibition
Streptococcus pneumonia (14)	UDP-glucose dehydrogenase	NAD catalysis
	Alpha amylase	Glycosyl hydrolysis
	Peptidase M20	Glutamate carboxypeptidase
	Helicase	Molecular scissors

	Integrase	Viral DNA integration
	MerR	Transcription regulation
Streptococcus agalactiae (3)	LicD	Phosphorycholine metabolism
	TENA/THI-4/PQQC	Extracellular enzyme expression and thiamine synthesis
Streptococcus mutans (2)	ArgR	Arginine anti-regulator
Streptococcus equi (5)	Transketolase	Thiamine pyrophosphate co-factor
	Biotin carboxylase	Carboxyl catalysis and attached to lysine
Streptococcus suis (5)	MATE	Organic molecule transport
	Fic/Doc	Cell division and folate synthesis
	Oxidoreductase	NADP, NAD synthesis
	Transketolase	Thiamine pyrophosphate co-factor
	Haloacid dehalogenase	Removal of halides
	ABC transporters	Compound translocation
	DapB	Codes for dihydrodipicolinate reductase
Streptococcus parasanguinis (1)	-	-
Streptococcus salivarius (1)	-	-
Streptococcus pseudopneumoniae (1)	-	-
Streptococcus pasterrianus (1)	RecF/RecN/SMC	DNA metabolism and recombination, DNA dynamics
	Binding-protein-dependent	Transport system
	Histidine kinases	Histidine dimerization and phosphor-acceptor
	Replication protein	Plasmid replication
Streptococcus thermophiles (3)	Short chain dehydrogenase	Substrate specificity
	Phosphoribosyl transferase	Phosphoryl transfer
	Phosphotransferase system II	Catalyze phosphorylation reactions
Streptococcus sanguinis (1)	-	-
Streptococcus gallolyticus (2)	-	-
Streptococcus mitis (1)	CorA	Mg^{2+} influx, thought to splice protein
Streptococcus oralis (1)	-	-
Streptococcus gordonii (1)	-	-
Streptococcus parauberis (1)	CodY	Dipeptide transport operon repression
	Short chain dehydrogenase	Substrate specificity
	Binding-protein-dependent	Transport system
	DapB	Codes for dihydrodipicolinate reductase
	Integrase	Viral DNA integration
	DegV	Fatty acid transport and metabolism
Almost in all species	Transposase	Excision and insertion of mobile elements

Values within the parenthesis indicate the number of genomes (strains) for this study (56) in total. Out of the 144 genes and proteins, 37 (transposases) were present in almost all the species while no observation was made in some strains that could be attributed to absentia of source organism from genemark. S. pyogenes, S. pneumonia, S. Suis and S. parauberis recorded the highest score.

Figure 1 overview of methodology



CONCLUSION

The research purposed to find the functions concealed by the non-coding regions in *Streptococci* species which perhaps are involved in disease progression, diagnostics, therapeutics and biological niches. The findings revealed potential genes and proteins contained in these regions as listed in table I and its suggested that any research extensions will conceal the functions of the non-coding sequences as potential genomic islands and as well pave way for new therapeutic agents and explain the biological niches which could help in the management of Streptococcal infections.

ACKNOWLEDGEMENT

Thanks to Mr. Anand Kumar and Amith Patidar from the Department of Bioinformatics, Bharathiar Univeristy for vital suggestions and perl script language for sequence extraction.

REFERENCES

1. Christos A Ouzounis, Peter D Karp., The past, present and future of genome-wide re-annotation. Genome Biology. 3(2): comment 2001: 1-2002.6. (2002)
2. Venter, J et al., The sequence of human genome. Science. 291: 1304-1351, (2001)
3. Roger P. Alexander, Gang Fang, Joel Rozowsky et al., Annotating non-coding regions of the genome. Nature Reviews Genetics 11: 563, (2010)
4. International Human genome Sequencing Consortium. Initial sequencing and analysis

of the human genome. Nature. 409: 860-921, (2001)
5. Guttman. M., Amit. I., Garber. M. et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 458: 223-227, (2009)
6. King, M.C. and Wilson, A.C., Evolution at two levels in humans and chimpanzees. Science. 188: 107–116, (1975)
7. Jayavel Sridhar, Govindaraj Somiya, Kanagaraj Sekar et al., PsRNA: A comparison Engine for the comparative identification of putative Small RNA locations within intergenic regions. Geno. Proteo. Bioinfo. 8(2): 127-134, (2010)
8. Stortz G, Altuvia S, Wassarman K. M., An abundance of RNA regulators. Annu Rev Biochem 74: 199-217, (2005)
9. Christophe Pichon, Brice Felden., Proteins that interact with bacterial small RNA regulators. FMS Microbio Rev 31: 614-625, (2007)
10. Pichon C, Felden D., Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. Proc Natl Acad Sci USA 102: 14249-14254, (2005)
11. Ghildiyal, M. and Zamore, P.D., Small silencing RNAs: an expanding universe. Nature Rev. Genet., 10: 94–108, (2009)
12. Garcia-Pino A, Christensen-Dagnuson RD, Gerdes K et al., Dc of prophage P1 is inhibited by its antitoxin partner Phd through fold complementation. J Biol chem [Epub ahead of print], (2008)
13. Opdyke JA, Kang JG, Stortz G., GadY, a small-RNA regulator of acid response genes in *Escherichia coli.* J Bacteriol 186: 6698-6705, (2004)
14. Mazmanian SK, Liu G, Ton-That H et al., "*Staphylococcus aureus* sortase, an enzyme that anchors surface proteins to the cell wall". Science 285 (5428):760–3, (1999)
15. Moller T, Franch T, Udesen C et al., Spot 42 RNA mediates discoordinate expression of the *Escherichia coli* galactose operon. Genes Dev 16:1696-1706, (2005)
16. Kong L, Zhang Y, Ye ZQ et al., CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35: W345–W349, (2007)
17. Ouzounis, C.A., Karp, P.D., The past, present and future of genome-wide re-annotation. Genome Biology, 3: 1-6, (2002)
18. Pierre Mandin, Francis Repoila, Massimo Vergassola et al., Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA target. Nuc. Acids. vol 35, No 3: 962-974, (2007)
19. Bateman, A., Birney, E., Durbin, R et al., The Pfam protein families' database. Nucleic Acids Res., 28: 263-366, (2000)
20. The Interpro Consortium Apweiler. R, Attwood. T.K., Bairoch, A et al., The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29: 37-40, (2001)
21. M.S. Pavelka, W.R. Jacobs., Biosynthesis of diaminopimelate, the precursor of lysine and a component of peptidoglycan, is an essential function of *Mycobacterium smegmatis* J. Bacteriol. 178: 6496-6507, (1996)
22. LeMieux J, Woody S, Camilli A., "Roles of the sortases of *Streptococcus pneumoniae* in assembly of the RlrA pilus". J. Bacteriol. 190 (17): 6002–6013, (2008)
23. Schneewind O, Mazmanian SK, Ton-that H., "Sortase-catalysed anchoring of surface proteins to the cell wall of *Staphylococcus aureus*". Mol. Microbiol. 40 (5): 1049–1057, (2001)
24. Cossart P, Jonquières R., "Sortase, a universal target for therapeutic agents against gram-positive bacteria?" Proc. Natl. Acad. Sci. U.S.A. 97 (10): 5013–5, (2000)
25. Massé E, Escorcia FE, Gottesman S., Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. Genes Dev 17: 2374-2383, (2003)

26. Maresso AW, Schneewind O., "Sortase as a target of anti-infective therapy". Pharmacol. *Rev.* 60 (1): 128–141, (2008)
27. Zhang JR, Idanpaan-Heikkila, Fischer W et al., *Pneumonococcal* licD2 gene is involved in phosphorylcholiine metabolism. Mol Microbiol. 31:1477-1488, (1999)
28. Kuchta K, Knizewski L, Wywicz LS et al., Comprehensive classifications of nucleotidyltransferase fold representatives in human. Nucleic Acids Res. [Epub ahead of print], (2009)
29. Sledjeski DD, Whitman C, Zhang A., Hdf is necessary for regulation by the untranslated RNA DsrA. J Bacteriol 183: 1997-2005, (2001)
30. Strunnikov AV, Jessberger R., Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions. Eur J Biochem. 263:6-13, (1999)
31. Belitsky BR, Gustafsson MC, Sonenshein AL et al., An lrp-like gene of *Bacillus subtilis* involved in branched-chain amino acid transport. J. Bacteriol. 179: 5448-57, (1997)
32. Barrack J, Sudarsan N, Weinberg Z et al., 6S RNA is a wide regulator of eubacteria RNA polymerase that resembles an open promoter. RNA 11: 774-784, (2005).