BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(19)

Strategic genome-scale prioritization of unique drug targets: A case study of *Streptococcus gordonii*

Sandeep Telkar, Hulikal shivashankara Santosh Kumar, Nagaraja Deeplanaik & Riaz Mahmood*

Department of Biotechnology and Bioinformatics, Kuvempu University, Shankarghatta- 577451, Karnataka, India; Riaz Mahmood - Email: rmahmood@kuvempu.ac.in; *Corresponding author

Received October 24, 2013; Revised November 05, 2013; Accepted November 06, 2013; Published December 06, 2013

Abstract:
The current reach of genomics extends facilitated identification of microbial virulence factors, a primary objective for antimicrobial drug and vaccine design. Many putative proteins are yet to be identified which can act as potent drug targets. There is lack and limitation of methods which appropriately combine several omics ways for putative and new drug target identification. The study emphasizes a combined bioinformatic and theoretical method of screening unique and putative drug targets, lacking similarity with experimentally reported essential genes and drug targets. Synteny based comparison was carried out with 11 streptococci considering *S. gordonii* as reference genome. It revealed 534 non-homologous genes of which 334 were putative. Similarity search against host proteome, metabolic pathway annotation and subcellular localization predication identified 16 potent drug targets. This is a first attempt of several combinational approaches of similarity search with target protein structural features for screening drug targets, yielding a pipeline which can be substantiated to other human pathogens.

Keywords: *Streptococcus gordonii*, infective endocarditis, drug targets, tractability, druggibility, ensemble

Background:

Streptococcus gordonii is a primary colonist of multispecies that forms biofilm, dental plaques and a potential agent of infective endocarditis (IE) **[1].** IE occurs along the edges of the heart valves forming vegetations of commensal pathogens at the site of infection. As inflammation continues, ulceration may result into the host death by intractable heart failure. Mortality due to IE is high; more than one-third of patients die within first year of diagnosis **[2].** Despite improvements in the diagnostic and therapeutic strategies, the fatality rate due to IE has not significantly decreased **[3].** This strongly indicates the need to search new therapeutic targets for pathogens that would offer better protection against IE.

Currently, various post-genomic methods are being widely used to identify novel drug and vaccine targets. There is a requirement for identification of novel drug targets for drug resistant pathogens, and theoretical approaches cater the need through '*in silico*' methods **[4, 5].**
In recent years the number of whole genome sequenced species has increased rapidly with approximately 6862 bacterial genomes currently available (Genomes OnLine Database-GOLD - http://www.genomesonline.org). Among these, genomes of several streptococci including *S. gordonii* and *S. sanguinis* are also available which are predominantly reported as causative agents of IE **[6].** This study reports a strategic approach of prioritizing the identification of putative drug targets that are unique for *S. gordonii*.

Methodology:

Subject organism, genome comparison and synteny

An electronic taxonomy of bacteria, eMLSA.net (http://viridans.emlsa.net/) was considered for streptococci classification. Viridans group streptococci (VGS) database was focused which are reported for IE **[7].** All the strains within VGS available in eMLSA were cross-checked in the literature for IE infection. Organisms with complete genome sequence that were common in both eMLSA and SynteBase were subjected for comparative study. SynteView visualizer of SynteBase was used to select these common organisms **[8].** *S. gordonii* was considered as reference genome and compared with ten other strains namely *S. sanguinis* SK36, *S. agalactiae* A909, *S. agalactiae* NEM316, *S. agalactiae* 2603V/R, *S. pneumoniae* D39, *S. pneumoniae* R6, *S. pneumoniae* TIGR4, *S. suis* 98HAH33, *S. suis* 05ZYH33, *S. mutans* UA159, which are reported for IE. Gene order was visualized with SynteView as synteny blocks. All proteins of *S. gordonii* that were putative and non-homologous to any compared strains were obtained from SynteView. Homologous putative and non-homologous hypothetical proteins within the comparison were excluded from the study.

Identification and subcellular localization of unique non-host pathoproteins

To identify putative candidate drug targets, the foremost concern was to find proteins present in *S. gordonii* and absent in humans (non host). Unique non-host proteins of pathogen were identified by two consecutive steps, one by sequence similarity search, followed by metabolic pathway mapping and annotation. All the unique putative proteins of *S. gordonii* were submitted to DELTA-BLAST program **[9].** E-value was restricted to 1e-3 for similarity search against human proteome with taxonomic id 9606. Hits obtained were neglected for further study, and the proteins with no significant similarity were subjected to KEGG database analysis tools (Kyoto Encyclopedia of Genes and Genomes at http:// www.kegg.jp /kegg/). With 'sgo' as the KEGG organism code for *S. gordonii*, available pathways were mapped by using KEGG Mapper (ver. 1.6) **[10].** Unmapped proteins were annotated by KAAS (KEGG Automatic Annotation Server ver. 1.67x) for pathway reconstruction using bi-directional best hit method, where functional annotation of genes are assigned by BLAST comparison against the manually curated KEGG GENES database **[11].** Entire 526 organisms of phylum Firmicutes available in KEGG was considered for similarity search by KAAS. All the metabolic pathways of humans from KEGG were extracted with the code 'hsa'. Pathways common to human and *S. gordonii* were disused and only proteins involved in unique pathways and unannotated by KAAS were noted. These proteins were submitted to iLoc-Gpos server which predicts subcellular localization of gram-positive specific bacterial proteins with single and multiple sites **[12].** Proteins localized to cell membrane, cell wall and cytoplasm were considered for further study and extracellular proteins were neglected.

Prioritization of target proteins

Screening was performed based upon the proteins druggibility. DrugEBIlity service at EMBL-EBI (https: //www.ebi.ac.uk /chembl/drugebility) was used. It exploits the structural data to evaluate whether a protein can be targeted with small molecule either by using structure or sequence file as input. Tractability, druggibility and ensemble were predicted using protein BLAST. Search was set to PDB Domains, filtered by greater than 30% identity with cutoff E-value 0.0001. Proteins with that of tractability and druggibility valued as 1 and ensemble being positive was preceded. Ensemble score ranging from +1.0 (signifies as druggable) to -1.0 (as undruggable) were analyzed. Only positive ensemble score were considered for target prioritization in descending order. For these predictions, relevant literature was collected, emphasizing the role of each protein in pathogenicity.

Results & Discussion:

In reductive genomics, any pathoprotein can be a candidate drug target with at least 5 targetability characters. 1) It shouldn't represent similarity with host proteome. 2) It should not be a secreted protein. 3) Shows tendency towards forming a protein structure with adequate residues being accessible to drug-like molecule. 4) Should have a biological significance in pathogen, being alterable and disturb the system and 5). After disturbing the nativeness of pathoprotein, the pathogen should be unable to develop an alternate ways to compensate the changes in its system **[5, 13, 14, 15, 16].** First criteria were contented by DELTA BLAST and metabolic pathway deduction between host and pathogen. Second criteria were checked by iLoc-Gpos, third criteria were screened by EBIs DrugEBIlity portal. The fourth and fifth criteria were screened by literature search.

Unique putative genes of S. gordonii

A new and electronic taxonomy of VGS available through eMLSA.net was followed, as it has always been problematic to classify streptococci taxonomically. Currently VGS database of eMLSA constitutes of 19 strains, of which 11 strains were considered for study. Among these, *S. agalactiae* and *S. pneumonia* consists 3 serotypes each and *S. suis* consists 2. *S. gordonii*, the reference organism, consisting of 2051 genes was subjected to SynteView. SynteView has many prokaryotic genomes accessible for which synteny based orthology and neighborhood data have been computed and stored in SynteBase **[8].** Among the compared strains, a large part of 534 genes (26.04% of genome) account for non-homologous, probably contributing for its unique adaptability, pathogenicity and virulence. This may be due to horizontal gene transfer for either gene gain, loss, change and gene decay which are considered to be a hallmark in the evolution of any pathogenic bacteria **[17].** These might also be an essential gene set, evolved unique to S. gordonii and might be similar to other class of commensal species. Within 534 genes, 334 (62.55%) are putative and 200 (37.45%) are hypothetical.

Non-host proteins

Homology between pathogen and host were primarily identified by DELTA-BLAST, which is considered to be more sensitive than BLASTp, as it searches a database of pre-constructed PSSMs before searching a protein-sequence database to yield better homology detection **[9].** By this, unique putative proteins of *S. gordonii* were separated showing tendency towards drug targetability at sequence level. Targeting the non-homologous set would cause no harm to host **[14].** Of 334 proteins, 110 didn't explain any similarity with host; this may be due to involvement of Low-complexity regions (LCRs) in some protein sequences, constricting the similarity search algorithms. LCRs were unfiltered as these facilitate in adaptation to fast evolving environments hence contributing to virulence **[18].** These 110 proteins were separated, adding to the first step towards populating the candidate protein targets.

Metabolic pathway annotation and reconstruction

The non-host 110 proteins obtained after DELTA similarity search were subjected to KEGG tools, resulting into 28 pathoproteins involved in various pathways. Among 28 proteins, 22 were mapped in *S. gordonii* specific pathways in KEGG with 16 involved in pathways similar to host and remaining 6 were reconstructed and annotated from KAAS against all 526 Firmicutes available in KEGG. KAAS results also described KEGG and COG (Clusters of Orthologous Groups) definitions for 26 proteins. Pathoproteins involved in pathways similar to host (16 proteins) were neglected and remaining 94 were conceded for further analysis in an attempt to identify strain specific targets. Protein sequences without pathway similarity to host or unpredictable pathways were 82 (74.54% of 110), which is a significant number. This may be again either due to involvement of LCRs at the sequence similarity search performed by KAAS **[11, 18]** or involvement of reduced dataset in KEGG server or pathogen evolution.

Subcellular localization

Subcellular localization of 94 non-host proteins resulted into 73 targetable pathoproteins. Among 94 proteins, 17 were predicted as extracellular, 33 as cytoplasmic, 6 as cell wall bound, 34 as cell membrane proteins and 4 were not possible with any predictions due to inadequate sequence length. Those confined to extracellular and remained unpredictable were ignored, as the interest was to uncover drug targets, and not vaccine targets. Hereafter, dataset was reduced to 73, and carried to druggibility screening.

Druggable targets

Pathoproteins though unique to pathogen are not always druggable, as they may lack the ability to form the structure, where a ligand can bind. Seventy three pathoproteins were screened by searching against existing PDB domains by considering three components. Firstly, Tractability scoring to be 1, evaluates the chemical drug available to pocket of a prospective target, modulating its activity leading to desired biological effect which is biologically viable. Secondly, Druggibility valued to 1, signifies the ability of a protein to bind a drug-like molecule with a therapeutically useful level of affinity. Lastly, ensemble described to be positive and nearing to 1 was selected as it is calculated by averaging the different structure-based druggability scores. The Ensemble Score ranging from Druggable: +1.0 to Undruggable: -1.0 was analyzed. Only positive scores of ensemble were considered as they are valuable representations of conformational flexibility of protein structures.

Restricting the results of 'DrugEBIlity' scores to 73 proteins, 57 were excluded. Remaining 16 proteins were further explored with literature search ensuring 9 as potent drug targets, 3 associated with at least one of the significant biological function for the pathogen survival and 4 revealed nothing **Table 1 (see supplementary material).**

This approach can be used to answer the pathoprotein sequences persisting with two problems. 1) Sequences whose identification of gene essentiality using databases like DEG fails to identify any pre-deposited essential gene (as only 4 out of 31 bacteria are gram positive in DEG 10.0 at the time of manuscript preparation). 2) Sequence similarity search with conventional and experimentally reported drug targets using databases like DrugBank, Therapeutic Target Database, Potential Drug Target Database etc., where one fails to identify any similarity.

Conclusion:

With the availability of complete genome and proteome of human pathogens, omic tools and databases, it is able to identify and characterize new and likely drug targets. This work exemplifies a theoretical approach for screening drug targets for a given human pathogen which is rapid but meaningful and momentous. This approach is based on genomic syntenies of pathogenome, similarity search, tractability, druggibility and ensemble of a pathoprotein. The study reveals; 16 potent, unique and putative drug targets in *S. gordonii*. Molecular modeling followed by virtual screening of these drug targets might be useful in the discovery of potential therapeutic compounds against *S. gordonii*. This strategy can be followed for other human pathogens to prioritize potent drug targets.

References:

[1] Vickerman MM *et al. J Bacteriol.* 2007 **189**: 7799 [PMID: 17720781]
[2] Thuny F *et al. Circulation*. 2005 **112**: 69 [PMID: 15983252]
[3] Habib G, *Heart*. 2006 **92**: 1 [PMID: 16365367]
[4] Pucci MJ, *Biochem Pharmacol*. 2006 **71:** 1066 [PMID: 16412986]
[5] Vetrivel U *et al. Hugo J*. 2011 **5**: 25 [PMID: 23205162]
[6] Westling K *et al. J Infect*. 2008 **56**: 204 [PMID: 18255158]
[7] Bishop CJ *et al. BMC Biol*. 2009 **7:** 3 [PMID: 19171050]
[8] Lemoine F *et al. BMC Bioinformatics*. 2008 **9:** 536 [PMID: 19087285]
[9] Boratyn GM *et al. Biol Direct*. 2012 **7:** 12 [PMID: 22510480]
[10] Kanehisa M & Goto S, *Nucleic Acids Res.* 2000 **28:** 27 [PMID: 10592173]
[11] Moriya Y *et al. Nucleic Acids Res*. 2007 **35:** W182 [PMID: 17526522]
[12] Wu ZC *et al. Protein Pept Lett*. 2012 **19**: 4 [PMID: 21919865]
[13] Read TD *et al. Drug Discov Today*. 2001 **6:** 887 [PMID: 11522517]
[14] Rathi B *et al. Bioinformation*. 2009 **4:** 143 [PMID: 20198190]
[15] Barh D *et al. Drug Development Research*. 2011 **72**: 2 [DOI: 10.1002/ddr.20413]
[16] Gashaw I *et al. Drug Discov Today*. 2011 **16**: 1037 [PMID: 21945861]
[17] Pallen MJ & Wren BW, *Nature.* 2007 **449**: 835 [PMID: 17943120]
[18] Verstrepen KJ *et al. Nat Genet*. 2005 **37**: 9 [PMID: 16086015]

Edited by P Kangueane
Citation: **Telkar *et al.*** Bioinformation 9(19): 983-987 (2013)

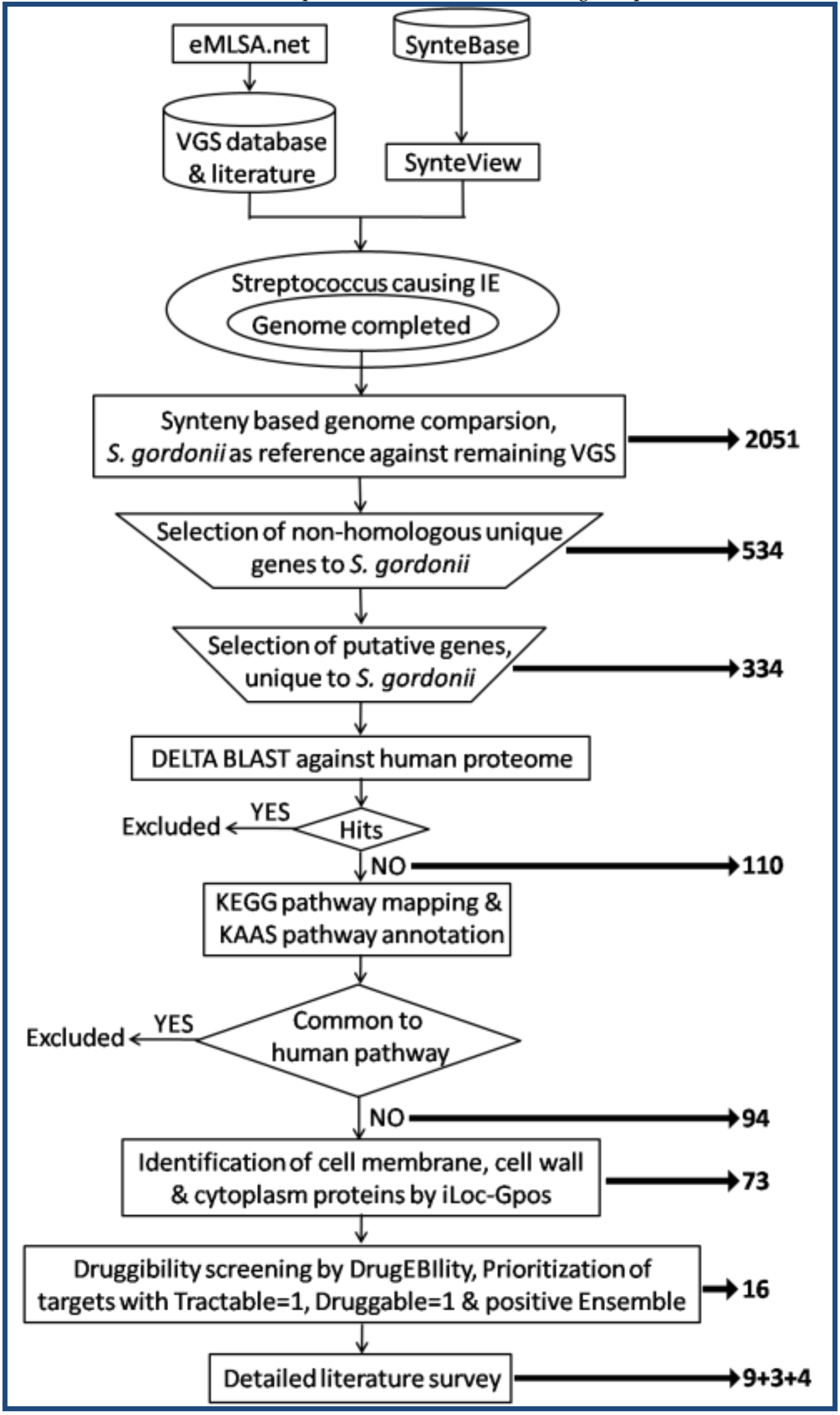
Supplementary material:

Table 1: 16 proteins screened after tractability, druggibility and ensemble scores assumed to be potent drug target. The definition column represents the protein name of *S. gordonii* str. Challis substr. CH1 as obtained from NCBI.

NCBI GI number	Definition	Sequence length	Cellular localization	PMID of the reference article	Inference
157150470	Universal stress protein	150 aa	cell membrane	12732303; 17081727	protecting the cell against DNA-damaging agents; *Salmonella* growth arrest, stress, and virulence
157150121	TetR-type transcriptional regulator	212 aa	cell membrane	15944459; 23602932	clearly mentioned as potent drug target
157150284	Ferric transport regulator protein	156 aa	cell membrane	11018148	clearly mentioned as potent drug target
157151402	TetR/AcrR family transcriptional regulator	211 aa	cell membrane	15944459; 23602932	clearly mentioned as potent drug target
157150255	PTS system cellobiose-specific transporter subunit IIB	106 aa	cytoplasm	7815935	clearly mentioned as potent drug target
157150241	chromosome segregation protein	254 aa	cytoplasm	-	-
157150641	SsrA-binding protein	155 aa	cytoplasm	16450010	The SmpB-SsrA system is significant in bacterial pathogenesis, survival under stress, and motility
157150139	oligopeptide-binding lipoprotein	660 aa	cell membrane	23734737	associated with anti-phagocytic activity
157149948	carbonic anhydrase	164 aa	cytoplasm	21779249	clearly mentioned as potent drug target
157151053	penicillin-binding protein 2A	740 aa	cell membrane	23873669	clearly mentioned as potent drug target
157150123	TetR/AcrR family transcriptional regulator	188 aa	cell membrane	15944459; 23602932	clearly mentioned as potent drug target
157150919	microcin immunity protein MccF	317 aa	cell wall	-	-
157150089	cell wall binding protein	290 aa	cytoplasm	-	-
157150576	multidrug ABC transporter	278 aa	cytoplasm	22312462	clearly mentioned as potent drug target
157149676	PTS cellobiose-specific enzyme IIA	107 aa	cell membrane	7815935	clearly mentioned as potent drug target
157150627	RRF2 family protein	149 aa	cell wall	-	-

Flow chart:

Schematic representation of steps followed for unique drug target identification in *S. gordonii*. Numbers just after right facing bold arrow marks indicates number of proteins obtained at each strategic step.


eMLSA.net
SynteBase
VGS database & literature
SynteView
Streptococcus causing IE
Genome completed
Synteny based genome comparsion, S. gordonii as reference against remaining VGS
2051
Selection of non-homologous unique genes to S. gordonii
534
Selection of putative genes, unique to S. gordonii
334
DELTA BLAST against human proteome
Excluded
YES
Hits
NO
110
KEGG pathway mapping & KAAS pathway annotation
Excluded
YES
Common to human pathway
NO
94
Identification of cell membrane, cell wall & cytoplasm proteins by iLoc-Gpos
73
Druggibility screening by DrugEBIlity, Prioritization of targets with Tractable=1, Druggable=1 & positive Ensemble
16
Detailed literature survey
9+3+4