Synteny Approach of Drug Target Prediction among Unique Hypothetical Proteins of Streptococcus Gordonii Causing Infective Endocarditis

Sandeep Telkar, Hulikal Shivashankara Santosh Kumar and Riaz Mahmood*

Department of Biotechnology and Bioinformatics, Kuvempu University, Shankarghatta- 577451, Karnataka, India

Abstract
At the time of infection, many of the gene products of completely sequenced organisms yet remain ‘hypothetical’ meaning they remain unsimilar to any previously characterized and may disguise some true virulent factors. Domain scanning provides a means of understanding functional information in these cases, extending facilitated identification of their virulence factors, proceeding for antimicrobial drug and vaccine design. In developing countries, mortality rate due to Infective endocarditis is accelerating along with retardation in efficiency of pathogen specific drugs. We have re-annotated at domain level and predicted cellular localization of 200 unique and hypothetical proteins obtained by syntenic comparison of Streptococcus gordonii among other strains of similar species for similar infection. The study resulted into 200 unique and hypothetical proteins, of which, domains of 85 proteins are predictable, representing 15 with no similarity with human proteome. Later, 9 proteins with 8 domains predicted to be antimicrobial targets. Further, these can be experimentally validated for drug and vaccine target ability.

INTRODUCTION
Hypothetical proteins still remain to be a good resource of a number of virulent factors. Microbial genome sequencing has produced a surplus amount of new information for identifying numerous genes encoding virulent proteins, where reductive and comparative genomics along with similarity search plays a significant role. Human deaths are extrapolated as a result of microbial infections due to ineffective drug and vaccines that work apart from a bunch of unidentified target proteins which make a part of so called ‘hypothetical’ proteins. Thriftiness of the pathogen may be hidden in such hypothetical entities whose identification, understanding and targeting may help in designing potent antimicrobial drugs or vaccines (Eisenstein et al., 2000).

Recently, approximately 6858 bacterial genomes have been sequenced as declared on Genomes OnLine Database (GOLD as on October 29, 2013). Among those reported, several Streptococcus species have their whole genome sequenced including S. gordonii, which is a primary colonist of the multispecies biofilm on tooth surfaces forming dental plaque and a potential agent of infective endocarditis (IE) in humans. Genome sequence of gordonii was first published in August 2007 with 2,151 open reading frames reported and deposited in GenBank (Vickerman et al., 2007). Mortality due to IE is high since more than one-third of patients will die within the first year of diagnosis (Thuny et al., 2005).

Despite improvements in the diagnostic and therapeutic strategies, the fatality rate due to IE has not significantly decreased (Sy and Kritharides, 2010). Even after gordonii first sequencing report and annotation, it still remains with none of the re-annotated data. This strongly indicates the need to search for new therapeutic targets hidden within hypothetical genes. Integrating various ‘omic’ approaches, we represent synteny based identification of hypothetical genes unique to S. gordonii among several compared endocarditis pathogens of same species and re-annotating them at domain level for the better understanding of hidden virulence.

MATERIALS AND METHODS
Genomes and Subject Organism

Streptococci (viridans group streptococci [VGS]) were focused which are reported for IE (Bishop et al., 2009). All the strains within VGS available in eMLSA were cross-checked in the literature (with the key word search “S. <genus name> human infective endocarditis” in pubmed and google scholar) for their IE infectiveness.
Genome completed organisms that were common to both in eMLSA and SynteBase were subjected for comparative study. Such organisms in SynteBase were selected using its own JAVA based visualizer plugin SynteView (Lemoine et al., 2008). S. gordonii has been considered to be the subject and hence was selected 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, and S. mutans UA159, which are reported for IE.

Syntenic Comparison & Unique Hypothetical Proteins

Genome comparison was performed based upon synteny of compared pathogens. Syntenic gene orders of studied organisms were visualized with SynteView. Synteny blocks of all proteins of gordonii that were unique which is non-homologous and hypothetical to any compared strains were obtained from SynteView. All homologous proteins that were either putative or hypothetical within the comparison were excluded from the study subjecting exclusively towards unique and hypothetical for further analysis.

Domain Analysis

Manually curated hypothetical proteins were further considered for domain analysis. This was achieved using EMBL-EBIs InterProScan and NCBI’s Conserved Domain Database (CDD) in parallel for each protein query. InterPro uses different protein signature recognition methods from the InterPro consortium member databases into one resource (Quevillon et al., 2005). CDD is the protein classification component of NCBI which is an interactive tool to identify conserved domains in new protein sequences (Marchler-Bauer et al., 2005).

Non-Host Protein Prediction and Their Virulence

Proteins whose domains are identifiable were further searched for any similarity against human proteome, as the foremost concern was to make out proteins present in gordonii and absent in humans (non host). Similarity search was performed by BLASTp with default values at NCBI using BLOSUM62 matrix against human proteome with tid9606 (Altschul et al., 1990). The proteins which remained non-homologous to that of human proteome preceded for further analysis. A detailed literature survey for their virulence role in various pathogens was performed for the refined proteins that were obtained with definitive domain architecture from CDD or InterProScan. Literature search engine like PubMed, sirus and Google Scholar were used for the search of desired articles.

RESULTS

Organisms Studied

In total 11 genome completed strains of streptococcus species were studied. Streptococcus species database within the electronic taxonomy of eMLSA.net and literature search revealed 11 pathogens as causative factor for IE, of which S. gordonii is the organism of interest. This has been considered as reference organism and on contrary compared with remaining 10 strains viz., 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, and S. mutans UA159, which are reported for IE.
Table 1: 15 proteins were screened, representing 9 with virulent domain and 6 with DUF domain. All the proteins with identified domains are provided with an inference taken from literature and the reference article is cited. The 6 DUF domains are also represented as their function can be identified in near future.

| Gene Name | Gene ID | Domain Identified | Inference |
|-----------|---------|-------------------|-----------|
| SGO_2077  | 157151091 | ABC2_membrane      | Bacterial ABC transporters are essential in cell viability, virulence, and pathogenicity. Other than functioning in transport, some bacterial ABC proteins are also involved in the regulation of several physiological processes. (Davidson et al. 2008); In bacterial efflux systems, used to secret effector molecules (Davidson et al., 2004). |
| SGO_0469  | 157150991 | ECF-type           | Energy coupling factor (ECF) are transporters used for uptake of vitamins in Prokarya (Dean, 2011); Found exclusively in archaea and bacteria, including the human pathogens Listeria, Streptococcus, and Staphylococcus, ECF transporters are used for the uptake of vitamins in Prokarya (Erkens et al., 2011). |
| SGO_0023  | 157150287 | MarC; Signal-peptide; trans membrane region. | Integral membrane protein family that includes the antibiotic resistance protein MarC. (from CDD); contributes for the full expression of multiple antibiotic resistance phenotype (Manu et al., 2011). |
| SGO_0326  | 157151686 | NodB-like catalytic domain; DUF2194; DUF2334 | Several microbial pathogens have developed sophisticated strategies to evade or modulate the host response to their advantage including NodB proteins (Balomenou et al., 2013). |
| SGO_0989  | 157149852 | NTF2;              | Significance in type IV secretion (Chandran et al., 2013) |
| SGO_0725  | 157151537 | PreW-protease.     | PreW is an important regulator of antimicrobial resistance and may be important for colonization and survival during an infection (Ho and Ellermeier). |
| SGO_1646  | 157151505 | RDD.               | This family of proteins contain three highly conserved amino acids: one arginine and two aspartates, hence the name of RDD family. The molecular function of this region is unknown. However this region may be involved in the transport of an as yet unknown set of ligands (Bateman A pers. obs.). |
| SGO_1501  | 157150744 | TraX.              | TraX is responsible for the amino-terminal acetylation of F-pilin subunits (Moore et al., 1993; Maneewannakul et al., 1995). |
| SGO_1286  | 157151565 | TraX.              | TraX is responsible for the amino-terminal acetylation of F-pilin subunits (Moore et al., 1993; Maneewannakul et al., 1995). |
| SGO_0555  | 157150667 | DUF1837            | Domain of unknown function |
| SGO_1474  | 157151196 | DUF2829            | Domain of unknown function |
| SGO_1562  | 157149847 | DUF3169            | Domain of unknown function |
| SGO_0380  | 157151523 | DUF3290            | Domain of unknown function |
| SGO_0559  | 157150475 | DUF3320            | Domain of unknown function |
| SGO_2068  | 157150274 | DUF990             | Domain of unknown function |

Figure 1: Venn diagram of various set of proteins obtained at each level. Here PROTEOME represents complete proteome of S. gordonii, UH represents Unique or non-homologous and hypothetical, AD represents After Domain analysis and lastly NONHOST represents non-host proteins.
Sandeep Telkar et al.,

Figure 2: Represents overall workflow in a schematic pictorial form. Right facing bold arrow marks and numbers indicates number of proteins obtained at each strategic step. Here in the last step of literature survey, 9 represents proteins whose literature information regarding the significance was available, 6 represents proteins with DUF domains.

Streptococcal pathogens are very thrifty, and due to their significant quantity of GC% they would deserve a character of high recombination frequency. This builds a difficulty of classifying them taxonomically. For this purpose, we have used a new and electronic taxonomy of VGS available through eMLSA.net, which is based on multilocus sequence analysis using house-keeping genes allowing users to assign new pathogens of species via the internet (Bishop et al., 2009). Currently, VGS database of eMLSA constitutes 11 strains reported for IE and which were considered for this study. Among these, S. agalactiae and S. pneumoniae consists of 3 serotypes each and S. suis consists of 2.

Synteny based analysis not just greatly reduces the complexity of comparative genome sequence analysis but also extends its roots into evolutionary relation, leading to a more meaningful and significant biodata of the subject organism (Kemkemer et al., 2009; Seshadri et al., 2004; Engström et al., 2007). Identification of syntenic regions across the species of interest also informs rearrangements in gene order (Adhikari et al., 2013). S. gordonii, the reference organism, consisting of 2051 genes was visualized through SynteView in comparison with other 10 strains.

The study interest was narrowed down to unique and hypothetical proteins of gordonii among the compared. Re-annotation was performed as an attempt to identify domains contributing to high degree of virulence and hidden within hypothetical proteins. Protein domains are not only known as units of structure, function and evolution but they also have a direct or indirect contribution and regulation in the bacterial pathogenesis. They also cumulatively intensify the virulence (Richardson, 1981; Bork, 1991; Zhang et al., 2013; Ryan et al., 2006; Schmidt et al., 2005; Dow et al., 2006; Simm et al., 2004; Ryan et al., 2004). This approach of re-annotation revealed 15 proteins with 6 having DUF domain. In the entire work-plan, DUFs were not neglected, as these are new biological entities that are likely and waiting to be discovered. DUFs remain a treasure trove of novel biology waiting to be explored (Bateman et al., 2010; Jaroszewski et al., 2009). They may also likely to play a role in the lifestyle of pathogens and hence can be promising targets for further experimental validation (Seidl et al., 2011; Ohm et al., 2012). In our analysis, out of 200 hypothetical proteins, 85 showed various protein domains.

More the non-homology between a host protein and a pathoprotein increases the tendency of pathoprotein for being an ideal drug target. Later it depends on the essentiality of non-host pathoprotein in the pathogen making it a good candidate drug target, and causing no harm to the host. This is usually performed by BLASTp similarity search (Rathi et al., 2009). Our study showed 85 proteins which were non-host by means of similarity search against human proteome and showing tendency towards drug targetability at sequence level. While performing the similarity search, protein low-complexity regions (LCRs) which are defined by a compositional bias and might give high scores that confuse the search program to find the actual significant sequences in the database (Mount, 2004). On contrary, these have a role in virulence (Maria Velasco et al., 2013; Coletta et al., 2010) and facilitate pathogens in adaptation to fast evolving environments hence contributing to virulence (Verstrepen et al., 2005). In our study, among 85 proteins, 15 were conceded further for literature survey in detail.

CONCLUSION

With the availability of complete genome and proteome of some human endocarditis pathogens, omic tools and databases, it is possible to identify and characterize likely drug targets. Here we represent 15 proteins, which can be targeted as novel drug targets. Their domain level functional re-annotation is explored finding 9 proteins to be seriously participating in the pathogenesis directly or indirectly inferred from previous available literature. Further, experimentally understanding of functions of 6 proteins with DUF domain would also lead to novel therapeutic drug targets. Structural genomics studies followed by molecular modeling followed by virtual screening of these deduced candidate targets might be useful in the discovery of potential therapeutic compounds against S. gordonii.
REFERENCES

Adhikari, B.N., Hamilton, J.P., Zerillo, M.M., Tisserat, N., Lévesque, C.A., Buell C. Robin. (2013). Comparative genomics reveals insight virulence strategies of plant pathogenic oomyctes. PloS One 8: e75072.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

Balomenou, S., Fouet, A., Tzanodaskalaki, M., Couture-Tosi, E., Bobiota, V., Boneca, Ivo G. (2013). Distinct functions of polysaccharide deacetylases in cell shape, neutral polysaccharide synthesis and virulence of Bacillus anthracis. Molecular Microbiology 87: 867-883.

Bateman, A., Coggill, P. and Finn, R. D. (2010). DUFs: families in search of function. Acta Crystallographica. Section F, Structural Biology and Crystallization Communications 66: 1148-1152.

Bharath, B.R. and Manjunatha, H. (2013). Flux Balance Analysis Based Model for the Identification of Potent Drug Target: A Novel Strategy. Science Technology Arts Research Journal 7522: 50-55.

Bishop, C.J., Aanensen, D.M., Jordan, G.E., Kilian, M., Hanage, W.P., Spratt, Brian, G. (2009). Assigning strains to bacterial species via the internet. BMC Biology 7: 3.

Bork, P. (1991). Shuffled domains in extracellular proteins. FEBS Letters 28: 47-54.

Camus, J.C., Pryor, M.J., Medigue, C. and Cole, S.T. (2002). Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology 148: 2967-2973.

Chandran, V. (2013). Type IV secretion machinery: molecular architecture & function. Biochemical Society Transactions 41: 17-28.

Coletta, A., Pinney, J.W., Solis, D.Y.W., Marsh, J., Pettifer, S. R., Attwood, Teresa, K. (2010). Low-complexity regions within protein sequences have position-dependent roles. BMC Systems Biology 4: 43.

Dandekar, T. (2000). Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Research 28: 3279–3288.

Davidson, A.L. and Chen, J. (2004). ATP-binding cassette transporters in bacteria. Annual Review of Biochemistry 73: 241-68.

Davidson, A.L., Dassa, E., Orelle, C., & Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiology and Molecular Biology Reviews 72: 317-364.

Dean, P. (2011). Functional domains and motifs of bacterial type III effector proteins and their roles in infection. FEMS Microbiology Reviews 35: 1100-1125.

Dow, J. M., Fouhy, Y., Lukey, J. F., & Ryan, R. P. (2006). The HD-GYP domain, cyclic di-GMP signaling, and bacterial virulence to plants. Molecular Plant-Microbe Interactions 19: 1378-1394.

Eisenstein Edward., Gilliland Gary L., Herzberg Osnat., Moul John., Orban John., Poljak, Roberto., Banerjei Linda., Richardson Delwood., Howard Andrew. J. (2000). Biological function made crystal clear — annotation of hypothetical proteins via structural genomics. Current Opinion in Biotechnology 11: 25-30.

Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S. and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Research 17: 1898–908.

Ekersn Guus, B., Berntsson Ronnie, P.A., Fulyani Faizah., Majsnerowska Maria., Vujčić-Zagar Andreja., Ter Beek Josy., Poolman Bert., Slotboom Dirk Jan. (2011). The structural basis of modularity in ECF-type ABC transporters. Nature Structural and Molecular Biology 18: 755–760.

Hamelin Richard, C., Kema Gert H.J., Lawrence Christopher., Scott James, A., Spatafora Joseph W., Turgeon, B. Gillian., de Wit Pierre, J.G.M., Zhong Shaobin., Goodwin Stephen, B., Grigoriev Igor, V. (2012). Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. (A. Andrianopoulos, Ed.). PLoS Pathogens 8: e1003037.

Ho, T.D., and Ellermeier, C.D. (2011). PsaW is required for colonization, resistance to antimicrobial peptides, and expression of extracytoplasmic function σ factors in Clostridium difficile. Infection and Immunity 79: 3229-3238.

Jaroszewski Lukasz., Li Zhanwen., Krishna, S. Sri., Bakolitsa Constantina., Wooley John., Deacon Ashley., M., Wilson Ian., Godzik Adam (2009). Uncharted regions of the protein universe. PLoS Biology 7: e1002055.

Prashith Kekuda., Manasa, M., Poominna, G., Abhipsa, V., Rekha, C., Upashe, Shivaleela P., Raghavendra, H.L. (2013). Anti-bacterial, Cytotoxic and Antioxidant Potential of Vitex Negundo Var. Negundo and Vitex Negundo Var.Purpurascens. – A Comparative Study. Science Technology Arts Research Journal 2(3): 59–68.

Kemkemer Claus., Kohn Matthias., Cooper David, N., Fronenke Lutz., Högel Josef., Hameister Horst., Keher-Sawatzki Hildegard. (2009). Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. BMC Evolutionary Biology 9: 84.

Lemoine, F., Labeled, B. and Lespinet, O. (2008), SyntBase/SyneView: a tool to visualize gene order conservation in prokaryotic genomes. BMC Bioinformatics 9: 536.

Maneewannakul, K., Maneewannakul, S., & Ippen-Ihler, K. (1995). Characterization of traX, the F plasmid locus required for acetylation of F-pilin subunits. Journal of Bacteriology 177: 2957-2964.

Manu, D., Lupon, I. and Popescu, O. (2011). Mechanisms of pathogenesis and antibiotics resistance in Escherichia coli Enterotoxigenic. Annals of RSCB XVI(2): 7-19.

Marchler-Bauer Aron., Anderson John, B., Cherukuri Praveen F., DeWeese-Scott Carol., Geer Lewis., Y. Gwadz Marc., He Siqian., Hunwitz David I., Jackson John, D., Ke Zhaoxi., Lanzcyczyk Christopher., J., Liebert Cynthia A., Liu Chunlei., Lu Fu., Marchler Gabriele, H., Mullokandov Mikhail., Shoemaker Benjamin A., Simoyan Vahan., Song James, S., Thiessen Paul A., Yamashita Roxanne, A., Yin Jodie., J., Zhang Dachuan., Bryant Stephen, H. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Research 33: D192–6.
pathogen genomes reveals unique protein organization. Plant Physiology 155: 628–44.

Seshadri Rekha, Myers Garry, S.A., Tettelín Hervé, Eisen Jonathan, A., Heidelberg John, F., Dodson Robert, J., Davidsen Tanja M., DeBoy Robert, T., Fouts Derrick, E. Haft Dan, H., Selegut Jeremy., Ren Qinghu., Brinkac Lauren, M., Madupu Ramana., Kolonay Jamie., Durkin Scott, A., Daugherty Sean, C., Shetty Jyoti., Swartsbeyn Alla Gebregeorgis Elizabeth., Geer Keita., Tsegay Getahun., Malek Joel., Ayodeji Bola., Shatsman Sofiya., McLeod Michael, P., Smajs David., Howell Jerilyn, K., Pal Sangita., Amin Anita., Vashisth Pankaj., McNeill Thomas, Z., Xiang Qin., Sodergren Erica., Baca Ernesto., Weinstock George, M., Norris Steven, J., Fraser Claire M., Paulsen Ian, T. (2004). Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes. Proceedings of the National Academy of Sciences of the United States of America 101: 5646–5651.

Simm, R., Morr, M., Kader, A., Nimtz, M. and Römling, U. (2004). GDGDE and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. Molecular Microbiology 53: 1123-1134.

Sy, R.W. and Kritharides, L. (2010). Health care exposure and age in infective endocarditis: results of a contemporary population-based profile of 1536 patients in Australia. European Heart Journal 31: 1890–1897.

Thuny Franck, Di Salvo Giovanni, Disalvo Giovanni, Belliard Olivier, Avierinos Jean-François., Pergola Valeria., Rosenberg Valerie., Casalta Jean-Paul., Goumert Joanny., Derumeaux Geneviève., Iarussi Diana., Ambrosi Pierre., Calabró Raffaele., Calabrò Raffaello., Riberi Alberto., Collart Frédéric., Metras Dominique., Lepidi Hubert., Raoult Didier., Harle Raffaele., Caättel Fabian., Brinkac Կարողանել, Derrick Raffaele., Caättel Fabian., Weiler Pierre-Jean., Cohen Ariel., Habib Gilbert. (2005). Risk of embolism and death in infective endocarditis: prognostic value of echocardiography: a prospective multicenter study. Circulation 112: 69–75.

Verstrepen, K. J., Jansen, A., Lewitter, F. and Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. Nature Genetics 37: 986-990.

Vickerman, M.M., Iobst, S., Jesionowski, A.M. and Gill, S.R. (2007). Genome-wide transcriptional changes in Streptococcus gordonii in response to competence signaling peptide. Journal of Bacteriology 189: 7799-7807.

Zhang Yuanbao., Wei Chao., Jiang Wendi., Wang Lei., Li Churui., Wang Yunyue., Dow John Maxwell., Sun Wenxian. (2013). The HD-GYP domain protein RpfG of Xanthomonas oryzae pv. oryzicola regulates synthesis of extracellular polysaccharides that contribute to biofilm formation and virulence on rice. PloS one 8: e69428.