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 infective endocarditis (IE) in humans. Genome sequence surfaces forming dental plaque and a potential agent of primary colonist of the multispecies biofilm on tooth genome sequenced including reported, several Streptococcus species have their whole Database (GOLD as on October 29 been sequenced as declared on Genomes OnLine understanding and targeting may help in designing potent such hypothetical entities whose identification, proteins. Thriftiness of the pathogen may be hidden in 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. 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.

MATERIALS AND METHODS

Genomes and Subject Organism

eMLSA.net (http://viridans.emlsa.net/), an electronic taxonomy of bacteria was considered for streptococcus classification. Database of 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.
Sandeep Telkar et al.,

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 gordoni 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 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 gordoni 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, sinrus 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.

Comparative Genomics

Synteny information of all the 11 organisms was available in SynteBase. Synteny arrangement of these organisms were visualized via SynteView, as obtained from SynteBase. In total 2051 protein coding genes of gordoni were compared with the protein coding genes of other IE causing streptococcus strains. Among compared strains, a lion’s share of 534 genes (26.04% of genome) accounts to be unique and strain specific to gordoni. Within these unique 534 genes, 334 (62.55% of 534) are putative and 200 (37.45% of 534) are hypothetical.

Domain Analysis

Domains of all 200 unique and hypothetical proteins of gordoni were screened. Among 200 proteins, 115 revealed none of any domains, 22 represented domain of unidentified function (DFU) and 63 showed various protein domains.

Non-Host and Virulent Domains

A sum of 85 proteins (including proteins with DFU domain) were similarity searched against human proteome database using BLASTp, as our aim was to identify virulent protein domains in the set of unique and hypothetical proteins. Among 85 proteins 70 showed some amount of homology with human proteome. Fifteen proteins revealed to be non-homologous to human host. Further, these 15 non-host proteins having identified domains were preceded for detailed literature search which has revealed the involvement of these identified protein domains, directly or indirectly in pathogenesis within the host organism, hence behaving to be virulent domains. Table 1 represents all the 15 virulent domains with a brief description of their significance, cited with literature.

Using a 4 set data, Venn diagram was drawn using VENNY (Oliveros, 2007) for the set of proteins at each level, and is represented in figure 1. Schematic representation of steps followed in the entire scheme has been represented in figure 2. In all, the number of unique hypothetical proteins of gordoni among compared strains has changed from 200 in 2007 to 115 till date, predicting a few novel virulent protein domains in gordoni.

DISCUSSION

IE is an inflammation that can lead to death, if the pathogens are un-treated at proper and specific targets. Several novel strategies have been proposed for potent drug target identification by applying prediction models like that flux balance analysis (Bharath and Manjunatha et al., 2013). Also several phytoconstituents have been analyzed for their antibacterial activity against S. pyogenes and S. aureus among streptococcus species (Prashith Kekuda et al., 2013). None have been reported for re-annotating the hypothetical proteins of IE pathogens like S. gordonii or their drug targets.

In the present investigation we have re-annotated all the uncommon and hypothetical proteins of gordoni at domain level, assuming these would be disguising considerable number of pathoproteins at domain level (Camus et al., 2002; Dandekar et al., 2000).
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.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Gene ID</th>
<th>Domain Identified</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGO_2077</td>
<td>157151091</td>
<td>ABC2_membrane</td>
<td>Bacterial ABC transporters are essential in cell viability, virulence, and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>pathogenicity. Other than functioning in transport, some bacterial ABC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>proteins are also involved in the regulation of several physiological</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>processes. (Davidson et al. 2008); In bacterial efflux systems, used to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>secret effector molecules (Davidson et al., 2004).</td>
</tr>
<tr>
<td>SGO_0469</td>
<td>157150991</td>
<td>ECF-type</td>
<td>Energy coupling factor (ECF) are transporters used for uptake of vitamins</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>in Prokarya (Dean, 2011); Found exclusively in archaea and bacteria,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>including the human pathogens Listeria, Streptococcus, and Staphylococcus,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ECF transporters are used for the uptake of vitamins in Prokarya (Erkens</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>et al., 2011).</td>
</tr>
<tr>
<td>SGO_0023</td>
<td>157150287</td>
<td>MarC; Signal-peptide; trans membrane</td>
<td>Integral membrane protein family that includes the antibiotic resistance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>region.</td>
<td>protein MarC. (from CDD); contributes for the full expression of multiple</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>antibiotic resistance phenotype (Manu et al., 2011).</td>
</tr>
<tr>
<td>SGO_0326</td>
<td>157151686</td>
<td>NodB-like catalytic domain; DUF2194;</td>
<td>Several microbial pathogens have developed sophisticated strategies to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DUF2334</td>
<td>evade or modulate the host response to their advantage including NodB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>proteins (Balomenou et al., 2013).</td>
</tr>
<tr>
<td>SGO_0989</td>
<td>157149852</td>
<td>NTF2;</td>
<td>Significance in type IV secretion (Chandran et al., 2013)</td>
</tr>
<tr>
<td>SGO_0725</td>
<td>157151537</td>
<td>PreW-pro tease.</td>
<td>PreW is an important regulator of antimicrobial resistance and may be</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>important for colonization and survival during an infection (Ho and Ellermeier).</td>
</tr>
<tr>
<td>SGO_1646</td>
<td>157151505</td>
<td>RDD.</td>
<td>This family of proteins contain three highly conserved amino acids: one</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>arginine and two aspartates, hence the name of RDD family. The</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>molecular function of this region is unknown. However this region may be</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>involved in the transport of an as yet unknown set of ligands (Bateman A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>pers. obs.).</td>
</tr>
<tr>
<td>SGO_1501</td>
<td>157150744</td>
<td>TraX.</td>
<td>TraX is responsible for the amino-terminal acetylation of F-pilin subunits</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Moore et al., 1993; Maneewannakul et al., 1995)</td>
</tr>
<tr>
<td>SGO_1286</td>
<td>157151565</td>
<td>TraX.</td>
<td>TraX is responsible for the amino-terminal acetylation of F-pilin subunits</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Moore et al., 1993; Maneewannakul et al., 1995)</td>
</tr>
<tr>
<td>SGO_0555</td>
<td>157150667</td>
<td>DUF1837</td>
<td>Domain of unknown function</td>
</tr>
<tr>
<td>SGO_1474</td>
<td>157151196</td>
<td>DUF2829</td>
<td>Domain of unknown function</td>
</tr>
<tr>
<td>SGO_1562</td>
<td>157149847</td>
<td>DUF3169</td>
<td>Domain of unknown function</td>
</tr>
<tr>
<td>SGO_0380</td>
<td>157151523</td>
<td>DUF3290</td>
<td>Domain of unknown function</td>
</tr>
<tr>
<td>SGO_0559</td>
<td>157150475</td>
<td>DUF4238</td>
<td>Domain of unknown function</td>
</tr>
<tr>
<td>SGO_2068</td>
<td>157150274</td>
<td>DUF990</td>
<td>Domain of unknown function</td>
</tr>
</tbody>
</table>

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.

Synteny based genome comparison, S. gordonii as reference against remaining VGS

Synteny based genome comparison, S. gordonii as reference against remaining VGS

Selection of non-homologous unique genes to S. gordonii

Selection of non-homologous unique genes to S. gordonii

Selection of proteins containing domains

BLASTp against human proteome

Excluded NO

YES

Detailed literature survey

2051

534

200

85

946

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 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 (Adhkari 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


