Discussion
Streptococcal endocarditis is increasingly dominating the world. The causative pathogens, VGS are turning more virulent and resistant. Subtractive genomics, metabolomics, protein-protein interaction studies and virtual screening studies performed with the reference strain S. gordonii have resulted into many interesting findings. Additionally, the presence of ankyrin repeats in the strain S. sanguinis is captivate and made to study further and extrapolate the findings up to the level of residue.

In the following section, each findings are reviewed with hypothesis followed by a discussion. Ultimately the analysis of the implications of these findings are also discussed for future studies.

4.1 Viridans group streptococci taxonomy and species selection

The entry of streptococci into history was in 1879, when Louis Pasteur was studying puerperal fever [Nobbs et al., 2009]. Popularly, the classification of streptococci is based on the scheme proposed by Rebecca C. Lancefield in 1933, which groups streptococcal strains according to the carbohydrate composition of cell wall antigens. This is commonly known as Lancefield scheme of classification. This approach has always proved its success for the more pathogenic streptococci, but its extensivity is constrained by the view that group-specific antigens for less or non-pathogenic species may be absent or shared between distinct taxa [Kaiser et al., 2014]. Narrowing down to viridans group streptococci (VGS), one of the first classifications was proposed in 1906 by Andrewes and Horder, who classified three species of non-hemolytic streptococcus as the
"Streptococcus mitis group". The taxonomy of this group is very controversial, and for many years, a standardized naming scheme or typing system for this group is lacking [Coykendall, 1989; Facklam, 2002; Doern and Burnham, 2010]. Identification of VGS to the species level is difficult because VGS exchange genetic material rapidly [Simmon et al., 2008]. Such 'fuzziness' is predominantly found among bacteria with great recombination frequency [Hanage et al., 2005; Feil and Spratt, 2001] and such bacteria have high and variable guanine-cytosine (GC) composition [Wu et al., 2012; Musto et al., 2006].

This ambiguity of streptococci in general and VGS in specific, which are the major cause for streptococcal infective endocarditis (IE) was difficult to move with taxonomical classification for further study. Thus it eventually led to the search and use of molecular and electronic taxonomy based on multilocus sequence analysis (MLSA) approach accessed at viridans.eMLSA.net [Bishop et al., 2009]. Unlike multilocus sequence typing (MLST) which is usually applied to strains that belong to a well-defined species; MLSA is more often used when species boundaries are not well known [Gevers et al., 2005; Hanage et al., 2006]. Viridans.eMLSA.net was designed by taking 7 house-keeping gene sequences concatenated as obtained from 420 streptococcal strains to produce a VGS database. Further a reference tree was obtained from these concatenated sequence (fig.-R1(A)). This MLSA approach positioned the subjected type strains within same sequence clusters to that of the positions of various well-characterized genus as classified on the basis of 16s rRNA by Kilian in 2005 (fig.-R1(B)).

Hence, understanding of well-resolved species clusters through MSA and 16S rRNA facilitated the selection of species for further study.
4.1.1 Selection of species

Further, intending to extend the idea of sample size which are well reported isolates of IE from VGS to non-VGS, led to the addition of three more species of streptococci. Based upon literature, *S. mutans* represented to be a dominating causative from non-VGS and non-pyogenic group, *S. agalactiae* from non-VGS but pyogenic group and *S. suis* from non-VGS and zoonotic streptococci reported for IE in humans.

Care was taken so that the species identified causing IE were available within SynteBase, which is a database of identified synteny blocks. Synteny which is often referred to as conservation of genomic organization, affects both function and evolution of a bacteria [Junier and Rivoire, 2013]. Comparative genomics upon homologous synteny blocks explains the evolutionary forces that have influenced the genome organization [Kooij et al., 2005; Eichler and Sankoff et al., 2003]. Later, conservation and rearrangements of certain important genes that play a crucial role in adaptive changes for pathogenicity can be tracked [Groenen et al., 2012].

The resolving power of whole-genome comparisons to a large extent depends upon the phylogenetic relationship between the species [Kooij et al., 2005]. Hence, the phylogenetic classification based upon sequences suggested by Bishop et al., and Kilian along with synteny based information from SynteBase helped to meet the requirements of organism selection.

The extension made from VGS to various representative non-VGS species resulted into 11 strains (including all serotypes). Such an extension was certainly useful due to the following reasons:

i. To increase the sample size of 5 strains of VGS (*S. gordonii*, *S. sanguinis* and the three serotypes of *S. pneumoniae*), which is too
less for comparison to identify very unique regions within the reference strain.

ii. Cross-group comparisons within the genus *Streptococcus* reveals a differential conservation among various species involved in comparison, hence more unique features of VGS causative for IE would be identified.

### 4.1.2 Selection of reference strain

To answer the selection of reference strain, whose genome could be taken as reference for further comparison and analysis, a small pilot experiment was performed. Using the synteny information of each selected strain (table-R1), number of unique genes of each strain against the set of pan genome was calculated (union of the gene sets of all the strains of a species) as represented in fig.R-4. During this each strain was considered as reference and compared against the pan genome consisting of remaining 10 strains. Among all the 11 strains, the best three strains with highest number of unique genes have been observed (*S. mutans* UA159 > by *S. sanguinis* SK36 > *S. gordonii* str. Challis substr. CH1).

Adding to the information of high number of unique genes, information of GC% of each strain (table-R1) was also analyzed which indicates that streptococci have great recombination frequency. Such similar observation was also made by Zhou *et al.* [2014] and Gill *et al.* [2005]. Along with this, the group (VGS/non-VGS) to which strain belongs was also observed. The table-D1 below describes the various considerations that were used to select the reference strain.

<table>
<thead>
<tr>
<th>Strain (representative)</th>
<th>No. of unique genes</th>
<th>GC%</th>
<th>VGS/non-VGS</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. gordonii</em> str. Challis substr. CH1</td>
<td>534</td>
<td>40.5</td>
<td>VGS</td>
</tr>
<tr>
<td><em>S. sanguinis</em> SK36</td>
<td>723</td>
<td>43.4</td>
<td>VGS</td>
</tr>
<tr>
<td><em>S. mutans</em> UA159</td>
<td>808</td>
<td>36.8</td>
<td>Non-VGS</td>
</tr>
<tr>
<td><em>S. suis</em> 05ZYH33</td>
<td>444</td>
<td>41.1</td>
<td>Non-VGS</td>
</tr>
</tbody>
</table>
The observation of data from various strains tabulated above, the strain *S. sanguinis* SK36 would be ranked 1st because it has the 2nd highest number of unique genes, then followed by highest GC% and more importantly belongs to VGS. Despite, *S. gordonii* str. Challis substr. CH1 was considered as reference for the reason, Xu et al., [2011] have already presented a clear perception of gene essentiality in *S. sanguinis* and identified 218 essential genes using single-gene knockout method. Their experimentally validated data is also deposited in database of essential genes (DEG) available at http://www.essentialgene.org/ [Luo et al., 2013]. Apart from this the two pathogens *S. gordonii* and *S. sanguinis* are very close relatives having 97.01% homology based upon the reports of 16s rRNA sequence as reported by Kawamura et al., [1995]. In addition, earlier literature curation in relation to the colonizing characteristics of these species have led to some more interesting information. Both *S. gordonii* and *S. sanguinis* are considered primary colonizers and occupy the same intraoral niche when compared to *S. mutans* [Xu et al., 2014; Evans et al., 2014; Ye et al., 2013; Kreth et al., 2008]. Becker et al., [2002] have also showed that the pioneer colonizers *S. gordonii* and *S. sanguinis* can antagonize *S. mutans* when present in the oral biofilm in high numbers. These are the few important reasons to neglect *S. mutans* as reference, and further it also does not belongs to VGS. Though it has highest number of unique genes, it has the lowest GC% too. Lastly, *S. suis* even though has good GC% (not highest), has lowest number of unique genes and also belongs to a non-VGS group. Adding to this, *S. suis* is more importantly it is a pig pathogen which is a zoonotic agent causing severe infections to people in close contact with infected pigs or pork-derived products [Goyette-Desjardins et al., 2014].

In view of the above, *S. gordonii* strain was confirmed as reference strain among the members of VGS group which are reported to be the major causes for streptococcal infective endocarditis [Naveen Kumar et al., 2014].
Finally, considering *S. gordonii* as reference strain, its genome was compared with the remaining 10 strains, aiming to understand unique genes upon the pan genome.

### 4.2 Pan-genome and unique protein coding genes

Doolittle [1999], when just more than 20 complete prokaryotic genome sequences were publicly available, made a keen observation among the available genomes and has discussed in his review ‘Lateral genomics’ the driving forces to acquire new genes by a bacteria, which are unique. Later, the term ‘pan-genome’ was first coined by Tettelin *et al.*, [2005] describing it as the full complement of genes in a species, which can have large variation in gene content among closely related strains, while he studied the genome sequence of six strains representing the five major disease-causing serotypes of *S. agalactiae*. Next to this, Medini *et al.*, [2005] described it more scientifically as the union of the gene sets of all the strains of a species.

There is always a significant relation between unique set of genes and the quantity of pan genome among the compared closely related species of a genus. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any compared single genome [Tettelin *et al.*, 2005]. In this study, the quantity of pan genome is 22,855 and the total set of unique genes among all the compared strains is 5,004, which is 18% of the pan-genome.

Considering *S. gordinii* as reference genome which consists of 2051 protein coding genes identified, 26% (534 genes) of its genome is unique (fig.-R-3). This unique set of genes which approximately acquires a quarter part of its genome represents itself worth to be analyzed. Together with this the unique genes of remaining 10 strains have also been identified individually. *S. mutans* showed 808 unique genes that is highest followed by *S. sanguinis* with 723 unique genes and *S. gordonii* with 534
unique genes. Though *S. gordonii* represented the third highest set of unique genes to its genome size, it was considered as reference based upon various criteria as discussed in section 4.1.2.

The unique protein coding genes of *S. gordonii*, includes several hypothetical and putative proteins. *S. gordonii* might have incorporated such unique set of genes as these may be an essential part for its existence within the community of the commensal species. This represents recent acquisitions from distant sources outside the species which may have occurred substantially by lateral gene transfer [Lerat et al., 2005]. Doolittle [1999] has also expressed in his review that a bacterial cell receives such genes laterally because of one or more reasons as follows:

i. The recipient cell is in need to develop a novel biosynthetic or degradative capacity for which previously it had no gene or

ii. To develop resistance to an antibiotic or other toxic agent that inactivates the resident copy of the gene or

iii. To synthesize a protein whose properties are better adapted than those of the resident gene product to a new organismal niche.

Considerably, these might be few more causes why *S. gordonii* has the lion’s share of unique genes to its genome.

As discussed earlier in section 4.1.2, even though there exists 97.01% homology between *S. gordonii* and *S. sanguinis*, the latter acquires 189 (723 unique genes of *S. sanguinis* difference with 534 unique genes of *S. gordonii*) more unique gene set. Such a difference in the unique genes of these closely related bacteria could be a partial answer of local adaptation to their particular life styles or niches.

Unique genes will continue to be identified even after sequencing hundreds of genomes [Tettelin et al., 2005]. Yang et al., [2009] estimated that there are more than 9,000,000 unique genes in the human gut
bacterial community of which 65% were Firmicutes including the genus *Streptococcus*.

Hence, exploring such unique genes, by their annotation and re-annotation would help to understand few distinctive protein coding regions. Hoping such exploration would help to identify some targetable proteins ultimately inhibiting the growth of such primary colonizers.

### 4.3 Annotation of unique protein coding genes

Annotation is the ultimate result of any genome project, where a hard attempt is made to determine the complete genome sequence of an organism and understand the protein-coding genes and other important genome-encoded features [Primrose and Twyman, 2009]. Even there are several servers that are purely dedicated for the purpose of prokaryotic annotation like, national center for biotechnology information (NCBI) Prokaryotic Genomes Automatic Annotation Pipeline (www.ncbi.nlm.nih.gov genome/annotation_prok) hosted at NCBI, IGS Prokaryotic Annotation Pipeline (www.igs.umaryland.edu research/bioinformatics/analysis/index.php) hosted at Institute for Genome Sciences, University of Maryland, Rapid Annotation using Subsystem Technology (http://rast.nmpdr.org) abbreviated as RAST which is a fully-automated service for annotating bacterial and archaeal genomes, etc. There are many more such servers dedicated for the purpose of annotation.

Vickerman *et al.*, [2007] first sequenced the genome of *S. gordonii*. The strain Challis CH1 was the first to be ever sequenced in this species of *gordonii* and it is the only representative strain whose genome sequence is complete till date. Apart from this they have also deposited the expression data that was analyzed. They have annotated various protein definitions out of 2051 genes of *S. gordonii* and this is deposited in NCBI database. This genome consists of 534 unique genes out of 2051, which is
not a small quantity to be neglected. Deepening the annotation and functional inference of such unique genes would help in identifying unique drug targets and even develop some specific biomarkers. Keeping this in view, 534 unique genes were annotated. Protein definitions of 334 unique putative genes (appendix-1) were obtained as from NCBI, which also describes the definitions deposited by Vickerman et al., [2007].

Many researchers since a long time have shared their opinions about the hypothetical genes among pathogens. Conclusions from their discussion suggest that the hypothetical genes should not be neglected as the most genes annotated as hypothetical genes are functional genes [Kolker et al., 2004; Jiang et al., 2009; Xu et al., 2014]. Such hypothetical regions contribute for high degree of virulence or cumulatively intensify the virulence of pathogen [Richardson, 1981; Bork, 1991; Zhang et al., 2013; Ryan et al., 2006; Schmidt et al., 2005; Dow et al., 2006; Simm et al., 2004]. Any DUFs that were predicted were not neglected as these were widely distributed in bacterial signalling proteins at the time of their identification [Schultz et al., 1998]. Adding to this Goodacre et al., [2013] recognized that domain of unknown functions (DUFs) are essential in bacteria, by analyzing their frequency and conservation and named them as essential DUFs (eDUFs).

Primarily domain level re-annotation was focused assuming to contain considerable number of pathogenic proteins which are apparently disguised as hypothetical. 200 unique and hypothetical proteins, of S. gordonii have represented 85 proteins with domains. They also include 22 DUFs (appendix-2). Remaining 115 were without any domains or DUFs, and these were not considered for analysis as they may overlap unknowingly with human proteome or somewhere in metabolome [Barh et al., 2011].

Annotation using ontological terms, has a great significance for bacterial genes [Arnaud et al., 2009]. Adding to the present work of
domain annotations of hypothetical genes, genome ontology (GO) terms add more functional attributes that could be associated with particular domains [Goodacre et al., 2013]. GO terms were mapped to both putative and hypothetical proteins. The file containing GO annotations for each of 534 proteins are tabulated in appendix-3 stored in the compact disc. The main categories-biological process, molecular function and cellular component have been depicted in the fig.-R6 in a plot form.

After exploring and distinguishing the set of unique genes into putative and hypothetical, the gene (protein) essentiality was to be analyzed and non-host property were to be explored. Such exploration adds to protein targeting for various existing or new ligands. Hence the work was continued to identify such essential and non-host proteins.

4.4 Essential and non-host proteins

_In silico_ comparative genomics based subtraction using host and pathogen genomes is a powerful approach for the identification of genus- or species-specific genes, or groups of genes that are responsible for a unique phenotype as well as the virulence factors of the pathogen [Huynen et al., 1997, 1998]. It is then necessary to determine whether these genes are essential for pathogen survival and are non-homologues to host. The knowledge of non-host proteins generated with the help of bioinformatics tool is more significant and specific. This helps in drug discovery which have high affinity towards newly identified target site [Ravindranath et al., 2013; Hassan et al., 2014].

In the present study, all the non-homologues genes among the compared strains were assumed to be unique to the reference strain. As in nature such unique genes are adapted with an intention of use. It was later intended to identify non-host proteins among the set of unique proteins of _S. gordonii_. For the purpose, DELTA-BLAST (Domain enhanced lookup time accelerated - Basic Local Alignment Search Tool)
and BLASTp were used for putative and hypothetical proteins respectively. Preferably DELTA-BLAST was used than BLASTp because of its sensitivity and being able to be used for the detection of remote protein homologs [Boratyn et al., 2012]. It was thought worthy giving preference to DELTA-BLAST as the comparison was made between two different kingdoms of tree of life. Among 334 proteins, 110 didn’t explain any similarity with host (appendix-3); this may be due to involvement of Low-complexity regions (LCRs) in some protein sequences. LCRs were included while constricting the similarity search algorithms. LCRs were unfiltered as these are entities of pathogenic bacteria [Nandi et al., 2003]. It is also reported that these regions facilitate the pathogens in adaptation to fast evolving environments hence contributing to virulence [Verstrepen et al., 2005]. Subsequent application of the same principles to hypothetical set of proteins fetched no significant hits. Hence, this prompted to subject the hypothetical proteins for similarity search against host proteome using BLASTp, which has resulted into 15 proteins of which 9 had a definitive domain architecture and 6 were DUFs (appendix-3). These could be the candidate antimicrobial drug targets as they may contribute for virulence. A detailed literature survey of 15 domains within non-host hypothetical proteins ended all the proteins with known virulent domains (table-R4). Among 15, 9 had definitive domain architecture and 6 with DUFs. The 9 proteins with definitive domain architecture add more weight for to be targeted candidates.

The Database of Essential Genes (DEG) is the main resource that lists experimentally validated essential genes in bacteria, fungi, plants, and animals [Zhang et al., 2004; Luo et al., 2013]. Many earlier researchers have readily used DEG to identify gene essentiality and for target identification through comparative and subtractive genomics approaches [Sakharkar et al., 2004; Galperin et al., 2004; Damte et al., 2013, Shanmugham and Pan., 2013; Silver. 2013; Aksoy et al., 2014]. However in the present work DEG was not preferred as it did not produce any protein
hit against the set of all 534 unique genes of S. gordonii. This might be because till date, it only has 4 gram positive species among 31 available bacteria. Another interesting observation seen was that though there exists 97.01% homology between S. gordonii and S. sanguinis, and even upon the deposition of essential gene set of S. sanguinis in DEG database, S. gordonii did not represented even a single gene similarity with the DEG database with the run of default values.

All unique set of genes can be defined as essential because genome comparison was performed based upon syntenic approach. Later, non-hostliness of any protein in this subtractive genomics is defined based upon more sensitive DELTA-BLAST unlike to that of the ideas of above mentioned references where they have used BLASTp.

A protein to be entirely non-host not only depends on the sequence similarity, but also depends upon similarity between host and pathogen pathways. After filtering the data at pathway level, one can identify the protein targets in a more specific way. Hence pathway annotation and re-annotation was performed.

### 4.5 Pathway annotation and re-annotation

Pathway annotation and re-annotation of pathogens is very significant in finding therapeutic targets. Vetrivel et al., [2011] identified 8 potent target genes from prominent human respiratory pathogens - 2 targets from S. pneumonia and 6 from H. influenzae. Here they immediately mapped and annotated each essential and significant gene available in KEGG to know its metabolic pathway information. Later, Keshri et al., [2013] used the KEGG GENES database by KO system, to classify all essential genes of an important pathogen of rice X. oryzae pv. oryzae into different categories according to their involvement in different metabolic pathways. This implies, the host system be either
an animal or a plant, pathway mapping of the genes of pathogen that are essential and unique is always required to find the therapeutic targets.

With the inferences drawn from earlier works, the pathway annotation via mapping by KEGG Mapper was performed, resulting into only 22 proteins out of 125 mapped in various pathways of *S. gordonii* (table-R5). Remaining 103 proteins as annotated by KAAS produced 6 annotated pathways, 24 KO definitions (table-R6). A large set of 73 proteins remained un-annotated by either KEGG Mapper or KAAS (appendix-3). These unannotated proteins at pathway level includes both putative set and hypothetical set of proteins. Annotation in KEGG is essentially cross-species annotation giving K numbers to orthologous genes in all available genomes, and is known as KO definitions (http://www.genome.jp/kegg/ko.html). Hence, the KO definitions define some function to the subjected protein but in a broad view. Keeping this in view, the proteins obtained with KO definitions were not neglected.

The un-annotated 73 proteins of pathogens poses the challenge of annotation towards KEGG database as genes designated unannotated have no pathway assignments in KEGG. These set being un-mapped remain out of the box of being candidate drug target. This may also be due to involvement of LCRs while constructing the sequence similarity search by KAAS. LCRs might have added few more to the list of un-annotated proteins in different pathways.

The unifying goal of pathway annotation by both KEGG Mapper and KAAS was to perform subtraction of similar pathways between the host proteome and subject proteins of *S. gordonii*. This would add more specificity to the drug target list of *S. gordonii*. Aiming this the work was continued with pathway subtraction.
4.6 Non-host protein pathway subtraction

Simultaneously, in subtractive genomics approaches metabolic pathway subtraction is required to identify metabolic pathways that host and pathogen have in common as reviewed by Barh et al., [2011]. This helps in the identification of pathogen-specific pathways. Once the essential non-host homologue survival genes of a pathogen are identified, they need to be allocated to known pathways. If such gene is found crucial in any of the pathogen's metabolic pathways, it is can be proceeded for target validation via experimental methods. Even though non-host proteins were deduced after identifying the similarity at sequence level, there are few chances that the newly selected/designed candidate ligand may bind to a host protein involved in a pathway which is similar to that of the pathway in the pathogen. Hence the pathways annotated for the proteins of S. gordonii were further checked for the existence of similar pathways in human. Proteins of S. gordonii involved in a pathway which is similar to any pathway in host was hence subtracted from further analysis.

Among 125 proteins, 16 proteins shared the similar pathways between the human host and S. gordonii (table-R7). Pathways dedicated to the biosynthesis of amino acids, starch and sucrose metabolism, etc. have been found in common. As such pathways are essential for an organism, it is obvious that they overlap and later proteins involved in such pathways are not suitable for targeting. Theoretically exponentiating the data of 16 proteins among 125 involved in similar pathways with host, there are possibilities that ± 131 out of S. gordonii proteome (i.e 12.8% of 1025 proteins) may be involved in similar pathways with that of host. Though this fig. is experimentally unevaluated or bioinformatically unpredicted, it provides an opportunity for the researchers to identify and understand few more drug targets in future.
Identifying the essentiality of proteins on sequence similarity or exploring the non-host proteins on other partial deciding (based on pathway annotation tools) aren't enough to conclude the protein as drug targets. Instead, additional characters like protein availability to the candidate ligand molecule also decides whether the protein can be therapeutic target or not. This can be achieved by identifying protein localization adds to the information of a protein being therapeutic target. Hence, to proceed further, analysis of protein sub-cellular localization was additionally needed.

4.7 Subcellular localization

Essential non-human proteins that are crucial in pathways are identified and subsequently analyzed to determine their localization at various sites like cytoplasmic, cell wall, cell membrane, or extracellular locations, using appropriate localization prediction tools. Damte et al., [2013], have identified putative drug and vaccine target proteins in *Mycoplasma hyopneumoniae*. Also drug and vaccine targets in *Vibrio cholera* were identified by Chawley et al., [2013], along with this, many subtractive genomics work published earlier have always made use of such localization prediction tools to categorize the proteins later to which their ability of being good therapeutic targets was understood. Always unique, essential and non-host proteins are subjected for such analysis. Through this, the selected candidate therapeutic target as drug or vaccine target can be distinguished [Barh et al., 2011]. In the present investigation, sub-cellular localization prediction of 36 proteins were performed.

Since *S. gordonii* is a gram positive bacteria, localization predictions server iLoc-Gpos designed and dedicated specifically to gram positive bacteria was used. As per the developers claim [Wu et al., 2012] the overall success rate is up to 93% and explains that the predictor is trained to identify both single-location and multiple-location sites. This resulted into 10 cell membrane proteins, 2 cell wall proteins, 24 cytoplasmic
proteins and no extracellular proteins (table-R8). In actual, this prediction was performed in order to identify extracellular proteins in the set of 36 proteins obtained from earlier analysis. But the prediction resulted into zero extracellular proteins. This prediction was performed with an idea to omit the available extracellular proteins from further analysis as such proteins can be vaccine targets and not drug targets [Chawley et al., 2013; Telkar et al., 2013].

4.8 Protein druggability based target prioritization

Druggability of a protein represents its potentiality to be modulated in vivo by drug-like molecules to find the desired affects [Liu and Altman, 2014]. The selected 36 proteins were analyzed for druggability using ChEMBL’s DrugEBlility predictor. There are several commercial tools and databases for structure-based druggability assessment. But a publicly available database of pre-calculated druggability assessments for all structural domains within the Protein Data Bank (PDB) is provided through the ChEMBL’s as ‘DrugEBlility’ (https://www.ebi.ac.uk/chembl/drugebility). DrugEBlility can accept the inputs in the form of protein sequence and search them against the domains within PDB.

There are very few reports available, where the application of DrugEBlility has been implemented. This is not because it produces unreliable results, but because this feature was hosted by ChEMBL very recently. Horst et al., [2012] implemented the same server to build a strategic way to identify protein targets and then develop drugs to avoid dental caries. Later, Mpangase et al., [2013] developed an updated Discovery-2 resource (http://discovery.bi.up.ac.za/) where researchers are able to mine information on malaria proteins and predicted ligands. As a part of their update, implementation of druggability predictions from DrugEBlility are made available. Subsequently, in the same year
[2013], Trevarton developed MelanomaDB (http://genesetdb.auckland.ac.nz/melanomadb/chooseSets.php) which is a web tool for integrative analysis of melanoma genomic information to identify disease-associated molecular pathways implementing DrugEBIIlity as one of the criteria. Their work add to the significance of tool DrugEBIIlity. It is presumed that this is the only server which accepts the protein sequence input and calculates the druggability unlike to that of other structure based druggability prediction servers like DoGSiteScorer, DrugPred, fPocket, etc. which accept only the 3D structure of the subject protein.

A similar approach for structure-based druggability analysis for 36 non-host unique proteins obtained out of earlier analysis were predicted. This resulted into 10 candidate drug targets (table-R9). It is believed that, the present study is the first from any of the reports where subtractive genomics between a pathogen and host is combined with structure-based druggability prediction, to identify unique candidate drug targets. To understand the significance of all the 10 proteins, the earlier reports based upon experimental validations were reviewed and inferred (table-R10).

Among the 10 proteins which resulted from druggability analysis, only 4 proteins were selected. These selections were based upon the criteria discussed below:

i. In DrugEBIIlity, as ensemble score ranges from -1.0 to +1.0 it represents undruggable to druggable respectively. This shows that the ensemble predicted should be positive. Later, greater the value of ensemble is, better druggable the subject protein is (https://www.ebi.ac.uk/chembl/drugebility/faq). Grounded upon this the least ensemble score was set to 0.75.

ii. The volume of a druggable region of a protein also plays a critical role for being able to be selected as a druggable target. Larger pocket volumes increase the chance of finding the ligand in the predicted pocket [Laurie and Jackson, 2006; Volkamer et al., 2010].
Hence, the druggable proteins with minimum drug pocket volume of 7000.00 Å³ were considered for further analysis.

In view of the above constraints, only 4 proteins showed ensemble score greater than 0.75 along with drug volume above 7000.00 Å³. Both of these considerations are very important for a ligand to fit in the protein active pocket.

Among the 4 proteins which were eligible after druggability analysis, 3 were directly or indirectly related to sugar metabolism of *S. gordonii*. They also had a better druggability results with good ensemble score and drug volume. Looking at this, the PEP:PTS pathway was selected for targeting.

4.9 Phosphoenolpyruvate-dependent phosphotransferase system pathway analysis

There are three key energy-dependent solute transport systems in bacteria [Varela *et al.*, 2013]. The first is primary active transport, in which ATP hydrolysis is the mode of energy for the entry of molecules into, or efflux from, cells [Van and Hendrik, 1998]. Another system is the PEP:PTS in which a sugar solute is phosphorylated as it is transported across the membrane [Kumar *et al.*, 2011; Boos and Shuman, 1998]. Lastly, secondary active transport systems use ion gradients as the energy-mode for transport of nutrients into cells [Varela and Wilson, 1996] or efflux of molecules from cells [Mitchell, 1991].

After performing the analysis of protein druggability, the study was confined to PEP:PTS pathway including the reasons as described in section 4.8, which is also an important energy-dependent solute transport systems in bacteria [Varela *et al.*, 2013; Kumar *et al.*, 2011; Boos and Shuman, 1998]. Further, PEP:PTS is ubiquitous in bacteria and absent in eukaryotes [Postma *et al.*, 1993; Siebold *et al.*, 2001]. Hence, this entire system of protein can be a good target set.
The system consists of two phosphoryl carriers, enzyme I and HPr, and several PTS transporters, catalyzing the concurrent uptake and phosphorylation of several carbohydrates. Among these, enzyme I and HPr are common and proteins downstream of HPr vary accordingly with the different sugars [Kundig et al., 1964; Lux et al., 1995; Geese et al., 1989; Postma 1993].

Several earlier reports have been published proposing PEP:PTS as one of the potent drug target system. The subtractive genomics approach among the genus *Streptococcus* by Georrg and Umrana in 2012, have shown 16 proteins uniquely involved in the pathogen specific 6 unique pathways. Their prediction also includes the PEP:PTS. Recently, Huang et al., [2013] have also performed virtual screening against enzyme I of the PEP:PTS and has claimed xanthone derivatives could be potential antibiotics. Very recently, Uddin et al., [2014] have identified unique and essential proteins by metabolic pathway analysis approach that could act as drug targets against methicillin resistant *Staphylococcus aureus*. They also have concluded that the PEP:PTS could be as one of the potent systems to be targeted. Interestingly, two patents with patent numbers EP0866075 A2 and US6777176 B1 in 1998 and 2004 respectively by Bernhard Erni have also mentioned the bacterial PTS as a drug target system that catalyses the uptake and phosphorylation of carbohydrates.

Supported by the above mentioned reports, a keen interest was sought on the PEP:PTS pathway. This system was comprehensively understood based upon KEGG knowledgebase and earlier literature. The KEGG pathway map and associated KEGG pathway entries of this pathway was studied, constituting of KEGG pathway modules and 33 proteins involved as seen in table-R11 and table-R12 respectively. This was more understandable with the simple, clear and diagrammatic representation of KEGG pathway map with the entry number sgo02060 (fig.-R18).

This bacterial system involved several member proteins, and it is very important to understand the interaction among these members. Such
understanding of interaction helps to identify the main hub protein of the system which can be tested as a target for small molecules. After committing to PEP:PTS as an significant pathway that can be embattled, the PPI studies of its members was focused in S. gordonii.

4.10 Protein-protein interaction and identification of hub protein

It has been claimed earlier by experiments that proteins with more interacting partners are physiologically more important to the cell and are known as ‘hub proteins’ [Batada et al., 2006; Ekman et al., 2006]. Such proteins are less dispensable [Coulomb et al., 2005] and evolve slowly [Fraser et al., 2003]. Drawing conclusions from these works, it is clear that targeting such hub proteins, could be a vital action towards blocking the concerned metabolic pathway which forcefully disturbs the entire cell.

STRING PPI analysis of all the proteins involved in PEP:PTS metabolic pathway as observed from KEGG which, showed HPr with maximum number of interactions describing it as a hub protein (fig.-R19). HPr was concluded as hub protein because the interaction results of all the sources based upon genomic context, co-expression, experiments and data mining pointed HPr with the best STRING score (table-R13). Later, in the interaction it was the only protein which had an evidence at experimental level.

HPr have been extensively characterized at the structural level by many researchers to identify its functional significance. Almost 2 decades back Jia et al., in 1994 had first solved the experimental structure of HPr at 1.6 Å isolated from Streptococcus faecalis and described its function.

As the system consists of two common phosphoryl carriers, enzyme I and HPr, there is a maximum probability of structural consistency by targeting either of them using small molecules. It was a reliable idea to obstruct the interaction between these two members.
Hence disturbing the system of sugar uptake. In the support of PPI analysis performed, Napper et al., [1996] experimentally determined the mutational effects of Ser-46 (to Aspartate) in the member protein HPr. They ultimately hypothesized that in the HPr of gram-positive bacteria, when mutated to Asp46, the His15 phosphoacceptor activity of HPr decreases by about 2000-fold. His15 is the region where the active pocket of HPr lies. They concluded that, the phosphorylation or acidic replacement of Ser46 in HPr yields modest conformational changes that do not affect the active site at His15. But the principal effect of these alterations to Ser46 leads to inhibition of the binding of HPr to enzyme I. Their work provided enough support to target the member protein HPr at the residue Ser46, which avoids the interaction between enzyme I and HPr.

It has been observed that the experimentally resolved structure of HPr was absent in the structural database PDB. Hence, this forced to model the protein structure by taking the sequence of protein HPr from S. gordonii. Homology modelling is the bridge between genes and protein structures named as structural genomics.

4.11 Homology modeling of protein HPr

As no experimentally resolved structure of HPr isolated from S. gordonii was available, a computer generated model of the same was built. Hillisch et al., [2004] and later Cavasotto et al., [2009] have described in their reviews regarding percentage sequence identity between target and template. They have theorized that if the percentage sequence identity between target and template is greater than 50%, performing a pairwise sequence alignment is sufficient to build a homology model and such models can frequently be used for drug design purposes. Even annotation of new genes can be inferred [Gopal et al., 2001].
The model built was based upon the template 1PTF. The web based SWISS-MODEL protein modelling suite chose 1PTF as the best suitable template. This choice was more senseful and applicable both computationally and biologically because of the following reasons. Firstly, the template protein is isolated from *Streptococcus faecalis* (now called *Enterococcus faecalis* [Köhler, 2007]) which is also a gram-positive, commensal bacterium inhabiting teeth and gastrointestinal tracts [Rôças *et al.*, 2004; Molander *et al.*, 1998]. Secondly, the template protein is resolved at 1.6 Å, which is considerably a very good resolution depicting the structural information of protein in detail and clear. Lastly, the pairwise sequence alignment and its percentage identity is 79.31 between the target and the template, which is a good similarity. Because higher the similarity between the target and template, better the accuracy of the modelled structure will be.

Estimating the quality of protein structure models is a vital step in protein structure prediction. The model that was built was validated by three different methods like QMEAN-Z score, superimposition between template and modelled structure and by Ramachandran plot. All the three kinds of validations have reported the model as a good structure.

The QMEAN Z-score which is an estimate of the "degree of nativeness" by calculating the standard deviations of model with that of the experimental structures [Benkert *et al.*, 2008]. This represents the composite QMEAN of 4 factors (fig.-R21). Apparently, the deviation was near to zero, which made the position of model protein to lie within the dark shade built by the circles colored in different shades [Benkert *et al.*, 2011]. The superimposition between template and modelled structure obtained from target sequence showed visually almost no difference (fig.-R22).

The Ramachandran plot generated using PROCHECK (fig.-R23) inferred that 94.7% residues were positioned in favoured region, 3.9%
residues in additional allowed regions, hence explaining the exactness of the model built. Residues in generously allowed regions is 1.3%, constituting of 1 amino acid that is Ala16. The results of Ramachandran plot hence obtained for the model protein speak in correlation with that of the results of Jia et al., [1994], who had first solved the experimental structure of HPr. Both the results overlap and explain the remarkable disallowed Ramachandran torsion angles of Ala16 at the active center.

The ultimate results of protein structure determination has already proven its usefulness for lead optimization and direct drug design [Lundstrom, 2007]. Applying this thought, the predicted target by earlier steps led to the protein HPr. The experimentally unresolved structure of protein HPr from *S. gordonii* ultimately made to build a model using its sequence. Further, this structure was used for virtual screening against a large ligand library.

### 4.12 *In silico* molecular virtual screening of HPr

As discussed in section 4.10, protein HPr was considered as hub protein to be targeted. The modelled protein structure of *S. gordonii* was used for virtual screening against 9,41,971 ligands available with TACC server at the time of work.

The earlier literature survey suggest that there exists only one report by Huang et al., [2013], where the virtual screening has been performed against the targeted bacterial PEP:PTS. They have conducted screening using inhibitors against the enzyme I. Considering it as a potent drug target, they have presented that xanthone derivatives could be potent inhibitors, through the blockade of energy transport for the bacterial system.

The TACC server has no independent docking software or algorithm that could perform virtual screening. It uses AutoDock Vina, written by Trott and Olson at the Scripps Research Institute in 2010 to
perform the actual docking. The number of citations to all the versions of docking software designed by the Scripps Research Institute shows the remarkability and accuracy of their docking algorithm. AutoDock Vina is yet another new and very popular open-source program for doing molecular docking. The authors Trott and Olson also claim that it is more accurate and faster than the earlier versions. Along with this the TACC server uses AutoDock Tools, written by Morris and Olson by the same Scripps Institute to convert proteins and ligands to the format required by AutoDock Vina. The reliability of AutoDock has been proven by a significant observation that AutoDock even played an early role in developing the first clinically approved inhibitor for HIV integrase [Schames et al., 2004].

The search space i.e. the grid box should be as small as possible, but not smaller. The smaller the search space, the easier it is for the docking algorithm to explore the best ligand conformations. The grid box was optimally formed by 1 Å spacing (as suggested by the program designers) and set to 12 Å cube. The grid center was set on the significant residue Ser46 in the modelled HPr. This is because Ser46, a highly conserved residue behaves as a regulatory site in HPr for gram-positive bacteria [Jia et al., 1994; Napper et al., 1996].

The more samples the ligand library has, more the possibilities of getting best candidate ligand. Therefore a large dataset of approximately 9.5 lakh ligand molecules were used to screen against the modelled HPr. Flexible docking always has more biological relevance than rigid docking because in nature both proteins and ligands fluctuate between their alternative conformations [Fischer et al., 2014; Totrov and Abagyan, 1997; Lorber and Shoichet, 1998; Rosenfeld et al., 1995]. AutoDock Vina allows such flexible protein and flexible ligand docking and was used successfully.
In late 1990s, poor pharmacokinetics and toxicity were important causes of costly late-stage failures in drug development. It has become valued that these areas should be considered at the early stages of drug discovery process [Hodgson, 2001; Waterbeemd and Gifford, 2003]. Absorption, distribution, metabolism, excretion and toxicity (ADMET) based screening was hence performed resulting into best docked ligands which were filtered and prioritized.

Strong and weak H-bonds also are ubiquitous in protein-ligand interaction and play a crucial role. Suitable computational tools can be used to interpret such interactions in the protein-ligand interface [Panigrahi and Desiraju, 2007; Haider and Kamran, 2010]. In the present study, 5000 (appendix-4) best ligands were obtained after molecular protein ligand docking with their respective binding energy. Synthetic ligands with drug score ≥ 0.9 resulted into 6 molecules. Natural ligands with drug score ≥ 0.8 resulted into 2 molecules. Later, the interaction between the structure files of ligands and protein was answered by an automatic generation graphical system called LigPlot* [Laskowski et al., 2011]. This was answered in terms of number of H-bonds and their distance along with binding energy and ADMET results.

Based upon all these observations, out of 9, 41,971 ligands 8 molecules were preferred (table-R15). All the 8 molecules obtained were from ZINC database, which is a curated collection of commercially available chemical compounds prepared especially for virtual screening [Irwin and Shoichet, 2005]. Among the 8 ligands (ZINC01034243, ZINC01034351, ZINC03789766, ZINC00002070, ZINC03814354, ZINC00205423, ZINC15700003 and ZINC15672343) that were selected, the best 2 molecules were selected each from synthetic and natural group. The selected ligands are ZINC03789766 and ZINC15672343 from synthetic and natural molecules respectively. These were verified with
the existence of any earlier and known targets as available in ZINC database.

Among these the ligands, ZINC15672343 under natural category had none of any earlier targets. It also participated in molecular docking by forming 4 H-bonds with the protein HPr and also had a better ADMET score and binding energy. But did not formed any H-bond with Ser46.

Later in the case of synthetic category of molecules, the molecule ZINC03789766 showed the best results than any other. But when they were verified with the availability of earlier targets using information in ZINC database, it was already known to be potent inhibitor of Aldose reductase. Aldose reductase is active only at higher glucose concentration in sorbitol-aldose reductase pathway (popularly known as polyl pathway) and a root cause of diabetic neuropathy [Brownlee, 2001]. It is also a convincing target for the treatment of cardiovascular complications, cancer and also a [Reddy and Ramana, 2010; Tammali et al., 2011; Oates, 2008]. The popular name of ZINC03789766 is Fidarestat (http://zinc.docking.org/substance/3789766). El-Kabbani et al., [2004] has experimentally proved Fidarestat as a potent inhibitor of Aldose reductase. This ligand being a potent inhibitor for Aldose reductase, can even inhibit HPr in S. gordonii. Thus, this can be a better drug for diabetic patients with infections caused by streptococci. Fidarestat participated in molecular docking with Ser46 in the protein HPr by forming 2 H-bonds. It also represented the best ADMET score among all the 8 ligands with a better binding energy (table-R15).

4.12.1 Comparision of Fidarestat with ZINC15672343

The two best molecules obtained after virtual screening one each from synthetic and natural i.e. Fidarestat (ZINC03789766) and ZINC15672343 respectively were further explored and understood with the following observations:
1. The synthetic molecule **Fidarestat** mainly encompasses a coumarin heterocyclic nucleus whereas the structure of natural molecule **ZINC15672343** is mainly composed of 6 membered pyridine nucleus.  
**Inference:** Coumarins are the simplest heterocyclic structures having lowest toxicities [Arora et al., 1963; Wu et al., 2009]. The antibacterial properties of coumarins were first recognised in 1945 [Goth et al.,] A few antibiotics with the coumarin skeleton like novobiocin isolated from *Streptomyces niveus*, is mainly active against Gram-positive bacteria [Kawase et al., 2001].

2. **Fidarestat** contains oxygen as a hetero atom but **ZINC15672343** contains nitrogen as hetero atom. Generally, it is presumed that the presence of oxygen as a hetero-atom may strong thus the potency of the drug concern.

3. **Fidarestat** forms 2 hydrogen bonds between the oxygen and desired aminoacid Ser46. On the contrary, **ZINC15672343** forms 4 hydrogen bonds with aminoacids other than the desired Ser46.  
**Inference:** The rigidity of protein-ligand complex depends upon the number of hydrogen bonds [Panigrahi et al., 2007; Bissantz et al., 2010]. Certainly, **ZINC15672343** shows better bonding with protein having 4 H-bonds, but it failed to form even a single bond with the desired amino acid Ser46, whereas, **Fidarestat** has formed 2 H-bonds with Ser46.

4. Molecular weight of **Fidarestat** is 279.227 g/mol. whereas the molecular weight of **ZINC15672343** is 420.469 g/mol.  
**Inference:** Compounds with higher molecular weights are less likely to be absorbed and therefore hinders to reach the place of action. Thus, trying to keep molecular weights as low as possible is useful [Lipinski, 2000; Lipinski et al., 2012].
Moreover the compounds with low molecular weight get better renal elimination [Czock et al., 2012; Palmer, 2012].

Apart from these structural observations, the ADMET based comprehensive drug score as predicted by OSIRIS property explorer of Fidarestat and ZINC15672343 are 0.92 and 0.83 respectively. This indicates that Fidarestat may be less hazardous when compared to the latter. No doubt the binding energy of the latter is reasonably better than Fidarestat, however, other properties of, Fidarestat may promote it as a better promising drug.

The structure of Fidarestat can be further analyzed experimentally. If the results are unsatisfactory, these can be re-structured with a maintained ADMET property and can be re-evaluated.

In addition, although the focus of the current study was to identify the unique set of drug targets in S. gordonii, the study has also led to the identification of a distinctive protein repeat that was exclusively found in S. sanguinis among the compared 11 strains. The unique protein repeat was ankyrin (ANK) repeat-containing protein. A phylogeny and MSA based study was planned, which resulted into understanding of few biological phenomenon.

### 4.13 Ankyrin repeats

During the progression of syntenb based comparative genome analysis, the ANK repeat was unique to S. sanguinis among the 11 compared strains causative for IE. In actual, its ANK is the most common protein-protein interaction motif in nature and predominantly found in eukaryotic proteins [Bork, 1993; Al-Khodor et al., 2010]. The ankyrin repeat is a 33-residue motif in proteins consisting of two alpha helices separated by loops. The first characterized ANK-containing proteins were the yeast cell cycle regulator Swi6/Cdc10 and the Drosophila cell signaling Notch protein [Foord et al., 1999; Breeden and Nasmyth, 1987]. A similar
observation was made by Gomez-Valero et al., [2011] where comparative and functional genomics studies of two species of legionella (Legionella pneumophila and Legionella longbeachae) was conducted. They reported with an identification of eukaryotic like proteins as key players in host-pathogen interactions which even included ANK repeat.

To understand the cause for existence of this repeat in S. sanguinis, co-occurrence of this protein was sought from STRING database. STRING retrieved 30 different species of bacteria having ANK repeat (table R16). It included 23 gram negative bacteria, and 7 gram positive bacteria.

S. sanguinis is a well-known primary colonizer causing IE and uniquely has ANK repeat. But the co-occurrence of ANK repeats in 30 more pathogenic bacteria created an inquisitiveness of understanding the relation between ANK repeats with IE, hence phylogenic analysis was influenced. Molecular phylogenies provides comprehensive and well-supported classification that reflects species relationships [Ting and Sterner, 2013]. In the present study, phylogenetic relationship was constructed for 30 ANK repeats from different bacterial species. S. sanguinis fell into a separate cluster even after bootstrap was set for 1,000 replicates, this might be due to its very unique combination of amino acid composition. No earlier reports exists where a phylogenetic analysis of ANK within the genus streptococcus was studied. Very few reports exists where the studies were conducted on pathogens. It was seen that the phylogenetic analysis of the large family of poxvirus ankyrin-repeat proteins was performed within and across chordopoxvirus genera that ultimately revealed orthologous groups [Sonnberg et al., 2011].

Orientia tsutsugamushi is the causative obligate intracellular bacterium of scrub typhus, and were studied in regard to the ANK repeat containing protein [Vie Brock et al., 2014]. It is evident from their studies that many Type 1 secretion system (T1SS) substrates display a tropism for the host cell secretory pathway. Yet another significant report of Legionella
pneumophila, which is an aerobic, pleomorphic, flagellated, non-spore forming, gram-negative bacterium has been seen with ankyrin-repeat-containing proteins that are involved in the pathogen-host interplay during intracellular replication [Hubber and Roy, 2010].

In the present study, when the reports from earlier works were inferred for identification of each 30 species of bacteria as causative isolate of IE, it is identified that 5 bacteria other than S. sanguinis are also involved in inducing IE. They are Borrelia afzelii Tom3107, Fusobacterium nucleatum, Lactobacillus plantarum WCFS1, Legionella pneumophila str. Paris and Leptotrichia sp. oral taxon 215. Among these Borrelia afzelii Tom3107, Fusobacterium nucleatum, Legionella pneumophila str. Paris and Leptotrichia sp. oral taxon 215 are gram negative bacteria except Lactobacillus plantarum WCFS1 which is a gram positive. The un-rooted tree generated using NJ-method (fig.-R28) showed Borrelia afzelii Tom3107 and Lactobacillus plantarum WCFS1 species under one cluster even though they are different with their cell wall composition as evidenced by gram stain technique. The earlier is gram negative and the latter is gram positive. Therefore, it is clear from these observations that the ankyrin repeats are independent of cell wall composition of the pathogen as they as they are observed both in gram positive as well as gram negative bacteria.

Only 2 species S. sanguinis and Lactobacillus plantarum WCFS1 are gram positive and remaining are gram negative which indicates that IE is just not restricted to be caused by gram positive bacteria of either genus Streptococcus or Staphylococcus [Wilson et al., 1995].

The residual conservation of 6 IE causative bacteria showed that they are rich with Leu residue (fig.-R29). Following the observations of Kędzierski et al. [2004], it is clear that the repeats that are rich with Leu i.e. Leucine-rich repeats (LRRs) are versatile binding motifs found in a variety of proteins and are involved in protein-protein interactions especially in host-pathogen interactions. Moreover, it is also evident that they play a
critical role even in protein recognition [Kobe and Kajava, 2001; Bella et al., 2008].

The genome sequencing of various pathogenic or symbiotic bacteria and eukaryotic viruses identified numerous genes encoding ANK-containing proteins that were proposed to have been acquired from eukaryotes [Al-Khodor et al., 2011]. It is believed to have been acquired through horizontal gene transfer (HGT) [de Felipe et al., 2005]. Some of these proteins play important roles in microbial pathogenesis by mimicking or interfering with the host function [Bork, 1993; Walker et al., 2007].

The ANK were repeats prevailed in a pioneer primary colonizer S. sanguinis. Pre-clinical efforts in identifying drug targets and providing disease-modifying treatment for IE are desperately needed. The current deficiency of effective treatments for IE is partially due to the lack of appropriate biomarkers that aide in identifying specific strain type during diagnosis. In this study, while a candidate drug is proposed for further investigation and development for treating IE, it will be crucial to provide a reliable region within the genome of S. sanguinis to which a biomarker can be accompanied. Further, when it is used as a measurable indicator, the cause and severity of IE disease state can be known. The development of such novel diagnosis would help to administrate appropriate therapeutic interventions to maximize the chances of a successful treatment.

Thus, the overall observation of the present study has contributed in the understanding of the taxonomy of VGS. Subsequently, the genomics metabolomics and protein-interaction studies by employing available advanced bioinformatics resources using available experimental data from the public domain has thrown an insight into the unique genes and non-host protein of S. gordonii. Further, it has also led to identify the putative druggable targets. Among them PTS system
found to be more susceptible in the pathogen which further comprise HPr protein as a strong candidate drug target. Thus, this study has ultimately resulted in the prediction of two virtual drugs having different structures from two different sources. Further chemical synthesis and thorough pharmacological evaluation, may help to determine the efficacy of these drugs and certainly contributes in the development, of a novel antibacterial drug against IE.