Global Journal of
Technology & Optimization

Wadhwani, Global J Technol Optim 2016, 7:6
DOI: 10.4172/2229-8711.S1109

Research Open Access

In Silico Identification of Novel Potential Vaccine Candidates in *Streptococcus pneumoniae*

Anjali Wadhwani and Varun Khanna*

Institute of Life Sciences, Ahmedabad University, University Road, Navrangpura, Ahmedabad - 380009, Gujarat, India

***Corresponding author:** Varun Khanna, Institute of Life Sciences, Ahmedabad University, University Road, Navrangpura, Ahmedabad - 380009, Gujarat, India, Tel: +91-79-26302414-18; E-mail: varun.khanna@ahduni.edu.in

Received date: December 15, 2015; **Accepted date:** March 08, 2016; **Published date:** March 11, 2016

Abstract

Currently, most reverse vaccinology studies aim to identify novel proteins with signature motifs commonly found in surface exposed proteins. In the current manuscript, our objective was to computationally identify conserved, antigenic, classically or non-classically secreted proteins in pathogenic strains of *Streptococcus pneumoniae*. The pathogenic strains used in our analysis were TIGR4, D39, CGSP14, 19A-6, JJA, 70585, AP200, 6706B and TCH8431. PSORTb 3.0.2 was used to infer subcellular locations while SecretomeP 2.0 server was run to predict non-classically secreted proteins. Virulence was predicted using MP3 and VirulentPred webservers. A systematic workflow designed for reverse vaccinology identified 83 (45 classically secreted and 38 non-classically secreted) potential virulence factors. However, many proteins were uncharacterized. Therefore, InterProScan was run for functional annotation. Proteins failing to be annotated were filtered out leaving a set of 24 proteins (9 classically secreted and 15 non-classically secreted) as our final prediction for potential vaccine candidates. Nevertheless, predicted proteins needs to be validated in biological assays before their use as vaccines.

Keywords *Streptococcus pneumonia*; Vaccines; Reverse vaccinology; Pneumonia; Non-classically secreted proteins; Novel virulence factors

Introduction

Streptococcus pneumoniae, a gram-positive, alpha hemolytic, encapsulated, diplococcus human pathogen is a causative agent for sepsis, meningitis and pneumonia [1]. According to the UNICEF report in 2012, pneumonia was the leading killer of young children, accounting for 18% of the deaths among children (under age five) worldwide [2]. An estimated 120 million new cases of pneumonia occur each year, 97% of them in the developing world and 12% of them severe enough to require hospitalization [3]. 15 countries, mostly in Asia and sub-Saharan Africa accounts for 74% of the cases in the developing world with 43 million cases in India alone. Additionally, India tops in global pneumonia deaths of children under age five with nearly 400,000 cases reported in 2010 [2].

Antibiotics are often prescribed for treating pneumonia. Nevertheless, resistance to various classes of antibiotics, for example, β-lactams, macrolides, tetracycline and folate inhibitors is rapidly increasing [4, 5] which complicates the treatment and burdens the public health systems. Pneumococcal diseases are vaccine-preventable diseases and the preventive strategies available include 23-valent pneumococcal polysaccharide vaccine (PPSV) for two years and above individuals [6]. Children below two years fail to mount an adequate response to 23-valent adult vaccine and instead 13-valent pneumococcal conjugated vaccine (PCV) is used. However, none of the existing vaccines are effective for all the 105 different serotypes causing re-occurrence of pneumonia due to serotype replacement [7]. The other major drawbacks of current vaccines include unaffordable prices and shortage of supply to the poor and most affected countries [8]. Therefore, new affordable vaccines are needed to control pneumonia.

Identification and development of vaccines by conventional and traditional methods rely on empirical screening of few candidates based on the known features of the pathogen. The process is time consuming, expensive with a high failure rate and difficult for organisms that cannot be cultured in the lab. However, with the rapid accumulation of whole genome sequencing data in numerous online databases have tremendously increased the possibilities of selecting novel vaccine candidates using computational approaches [9]. Reverse vaccinology (RV) is one such approach, which involves the mining of genomic information for potential vaccine candidates using bioinformatics and sequence analysis tools [10]. This strategy depends upon identifying genes or gene products, which serve as critical components in metabolic pathways thus are essential for survival of pathogen but absent in the host. The notable advantage of computational screening is that it opens up vast repertoire of possible candidates and serves as an initial move to fish out proteins from the genome of a pathogen previously not accessible to researchers for vaccine development. A recent successful example of the RV to identify a potential vaccine candidate is Novartis's 'Bexsero'. In January 2013, an RV-derived vaccine (Bexsero) was approved by European commission for use in individuals from two months of age or older, making it first vaccine against meningococcus B (MenB) to help protect against meningitis B [11]. Sanofi has also used RV technique to develop a peptide-based vaccine for S. pneumoniae that is under Phase I/II trials, as well as other earlier-stage projects [12]. Besides bacteria, this approach has been tested against a variety of pathogenic organisms. Martiz-Oliver et al. [13] demonstrated the applicability of this approach against cattle tick Rhipicephalus microplus using a combination of functional genomics (DNA microarrays) and pipeline for in silico prediction of subcellular location and protective antigenicity. John et al. [14] employed RV technique to identify new vaccine candidates against a protozoan Leishmania through proteome screening by applying various filters such as non-homology to human

proteins, number of transmembrane helices subcellular localization and binding affinity to both MHC class I and class II alleles. However, no experimental validation was reported for the predicted candidates in both the studies mentioned above.

Though reverse vaccinology has many advantages, it has its own share of disadvantages. Primarily, because immunogenicity (immunogenic potential of antigens) has been difficult to predict [15], and more importantly, whole genomic information related to the pathogen must be available to initiate an RV-driven project. Generally, in RV project, proteins with the particular motifs that are found in secreted or surface exposed proteins are considered ideal while the cytoplasmic proteins are discarded [16, 17]. However, there is increasing evidence that cytoplasmic proteins could appear on the surface of bacteria and act as adhesins, invasins or provide drug resistance and modulate host immune response [18]. The term "moon-lighting protein" is now widely used to describe such cytoplasmic proteins that do not have classical features of bacterial surface proteins yet appear on the bacterial surface and can perform a secondary function. Interestingly, these proteins have been shown as immunogenic and even protective in several model organisms, including mice [19, 20]. Hence in the current study, RV strategy was applied to predict conserved, classically and non-classically secreted virulence factors of *S. pneumoniae*, which is an interesting organism due to the high degree of genomic variability among the pathogenic serotypes. It is expected that the identified potential vaccine candidates will not only expand our understanding of the molecular mechanisms of *S. pneumoniae* pathogenesis but also facilitate the production of novel therapeutics.

Materials and Methods

Sequence retrieval

A systematic workflow was designed with the goal of identifying potential vaccine candidates in *S. pneumoniae* **(Figure 1).** The protein sequences from one non-pathogenic and nine pathogenic and strains of *S. pneumoniae* **(Table 1)** were retrieved from UniProt database [21]. The non-pathogenic strain R6 [22] used in our analysis is a derivate of the serotype 2 clinical isolate D39. The gene encoding many virulence factors are present in R6 genome in addition to the genes of capsular biosynthesis [22]. It is a well-studied reference strain and researchers around the word use it to study *S. pneumoniae* infections because it is harmless and its genome can be easily manipulated. The pathogenic strains used in the current study are described below. TIGR4, a clinical isolate, encapsulated and highly virulent strain [23]. D39, is the encapsulated and virulent strain which was used in the landmark study on the role of DNA as the genetic material [24]. CGSP14 strain was a clinically isolated from a child in Taiwan who had necrotizing pneumoniae with complicating hemolytic uremic syndrome [25]. JJA is virulent serotype 14 strain and contribute significantly to pan-genome diversity. AP200 is a clinical strain isolated from Italy in 2003 which is resistant to erythromycin [26]. TCH8431 is an extremely virulent strain of serotype A19 which was isolated from human respiratory tract.

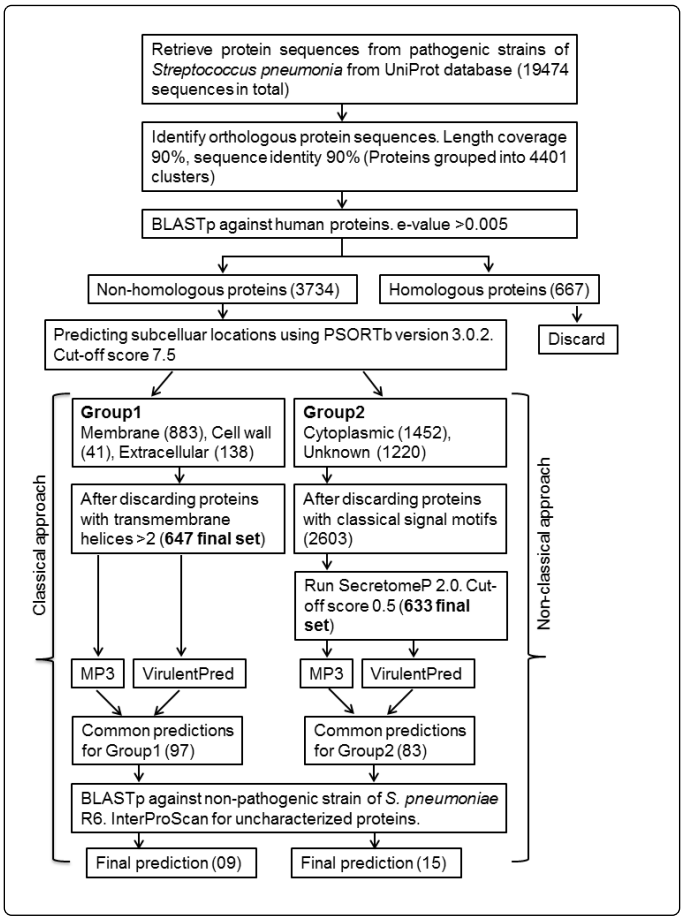

Figure 1: Schematics of the methodology adopted for identification of novel vaccine candidates. 83 proteins from pathogenic strains of *Streptococcus pneumoniae* under study were shortlisted as potential vaccine candidates using the principles of reverse vaccinology.

Proteome ID	Strain	Serotype	Number of proteins	Reference
UP000000585	TIGR4	4	2109	[23]
UP000001452	D39	2	1918	[24]
UP000001682	CGSP14	14	2193	[25]
UP000002163	19A-6	19A	2152	_
UP000002206	JJA	14	2120	_
UP000002211	70585	5	2178	_
UP000002229	AP200	11A	2205	[26]
UP000002232	6706B	6B	2336	[49]
UP000009083	TCH8431	19A	2263	[50]

UP000000586	R6	_	2030	[22]

Table 1: UniProt IDs and characteristics of the *Streptococcus pneumoniae* proteomes used in the study

Identification of orthologous proteins

Orthologs are homologs separated by speciation event. BLASTClust program within the standalone BLAST package [27] was used for clustering orthologous protein sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. The length coverage, sequence identity and e-value were set at 90%, 90% and 1e-6 respectively.

Identification of non-homologous proteins

Following on from the identification of orthologs, protein sequences of pathogenic strains were subjected to BLASTp against the human proteome. Proteins with an e-value greater than 0.005 were classified as non-homologous (not similar to human proteins) and were retained whereas others were discarded because they are likely to cause problems of autoimmunity.

Prediction of subcellular locations of proteins

All the non-homologous proteins identified were subjected to subcellular localization prediction using PSORTb version 3.0.2 [28]. PSORTb is the most commonly used software to predict the localization of proteins in prokaryotes. It uses a combination of six modules, each of which analyze factors influencing subcellular location of a protein such as the number of transmembrane helices, signal peptides, motifs known to be responsible for particular function and others. Each module outputs a score (between 1 and 10) for the probability of a protein being at a specific location. Cut of score of 7.5 is considered reliable for the predictions of subcellular locations. The PSORTb output was separated into different files according to the predicted locations. Proteins with more than two transmembrane helices were eliminated from further analysis not only because they are difficult to express, but also because they are likely to be embedded in the cell membrane and therefore, inaccessible to antibodies.

Non-classically secreted proteins

SecretomeP 2.0 server [29] was used for the prediction of non-classical, i.e., not signal peptide triggered secreted proteins. The method assigns a score between 0 and 1 to each protein, where a score above 0.5 indicates a possible secretion. Here we used a list of proteins classified as "Cytoplasmic" or "Unknown" which simultaneously did not get a prediction of containing a signal peptide by PSORTb as input to the SecretomeP server.

Prediction of proteins contributing to virulence of pathogen

Virulence factors are the disease-causing molecules usually associated with the pathogenesis. Immunization of the host with the combination of the virulence factors from a microbe would elicit enhanced protection when exposed to the microbial challenge. In the current study, two different methods (VirulentPred and MP3) were used, and the results were combined to obtain a consensus prediction. VirulentPred [30] is bacterial virulence factor (proteins) prediction software based on machine learning classification method called bilayer cascade Support Vector Machine (SVM). The first layer of SVM classifiers is trained using different individual protein sequence features. The results from the initial layer are then cascaded into the second layer SVM classifier to generate the final classifier. Similarly, MP3 webserver was developed by integrating SVM and Hidden Markov Model (HMM) approach to carry out fast, sensitive and accurate prediction of pathogenic proteins.

In this work, two different inputs were used to search for the possible virulence factors in VirulentPred and MP3 server with the SVM threshold set at > 0.7 to minimize the occurrence of false positives. The first input was the list of proteins classified as "Cytoplasmic Membrane" or 'Cell wall' from PSORTb, and the second set was composed of the non-classically secreted proteins from the secretomeP output. Subsequently, the consensus was obtained for the prediction of the virulence contributing proteins as potentially novel vaccine candidates. Finally, for all the identified virulence contributing proteins, a BLAST search was performed against the proteome of non-pathogenic strain (R6) of *S. pneumoniae*. Virulence contributing proteins with significant similarity to the proteins of R6 strain were discarded from the final prediction.

Results

Identification of orthologous proteins

Protein sequences from the nine pathogenic strains of *S. pneumoniae* were grouped into 4401 clusters, of which; 1302 were conserved in all the nine strains whereas, 1676 or 38% of proteins were present only in one of the strains. Our results closely match with a previous study on the 17 genomes of *S. pneumoniae* [31] where authors reported that 1454 (46%) of the total coding genes were conserved among all strains.

Identification of non-homologous proteins

In this section, we report the computational identification of non-homologous proteins of *S. pneumoniae*. First, the representative protein sequences from 4401 clusters were subjected to BLASTp against the human proteome to exclude the possibility of autoimmunity. We obtained 3734 proteins with no significant sequence similarity to human proteins, which were retained for further analysis while the others were discarded.

Prediction of sub cellular locations

All identified non-homologous proteins were subjected to subcellular localization prediction using PSORTb. The output of PSORTb was divided into two groups based on the predicted location. The first group comprised of 1062 proteins partitioned into 883 membranes located, 41 cell wall while 138 extracellular proteins. The second group comprised of 2672 proteins partitioned into 1452 cytoplasmic and 1220 proteins of unknown location. Proteins with more than two transmembrane helices (TMH) in group one and with the signal peptide motifs in group two were eliminated resulting in the selection of 647 and 2603 proteins from the first and second group, respectively. The proteins grouped under the *cytoplasmic* category are usually discarded from RV studies as they are not expected to interact with host immune system. However, number of recent studies have found some cytoplasmic proteins on the surface of the microbes even when they did not possess the classical peptide signals or surface exposure motifs. Therefore, instead of discarding these proteins we

searched for non-classically secreted proteins using SecretomeP 2.0 webserver. 633 proteins were found to be above the threshold score of 0.5 and hence were considered as secreted from the second group.

Prediction of vaccine targets

For further investigation, proteins in both the groups were screened to identity a possible role in virulence using VirulentPred and MP3 webservers. According to the criteria specified in the material and method section, we identified 376 and 153 virulence factors (proteins) from VirulentPred and MP3 webserver respectively for the first group. Similarly, we identified 409 and 126 virulence factors (proteins) from VirulentPred and MP3 webserver for second group. Subsequently, we took the common predictions for each group (97 and 83) and did a BLAST search against the proteome of a non-pathogenic strain of *Streptococcus pneumoniae* R6 to find virulence factors exclusively present in pathogenic strains. We found that there were 83 (45 from group 1 and 38 from group 2) potentially virulence factors present in all the nine pathogenic strains under investigation **(Supplementary File 1 and Supplementary File 2).** Further, we noticed that there were numerous uncharacterized proteins in both the groups. Hence, for functional classification and characterization of the proteins, we ran InterProScan tool [32] on proteins with unknown function. Proteins which could not be functionally annotated by InterProScan were filtered out while the remaining proteins are shown in **Tables 2 and 3** for group1 and group 2, respectively.

S. No.	Uniprot ID	Protein name	Pfam domains	Functional description	References
1.	B1IA00	Surface protein PspC	PF05062, PF00746	Choline binding protein	[51, 52]
2.	B1ICR7	Mucin binding protein (MucBP)	PF00746	LPXTG-motif containing cell wall protein	
3.	B1IB03	Uncharacterized	PF01289	Cholesterol binding cytolysins	
4.	D6ZQ98	Iron ABC transporter, iron-binding protein	–	ABC transporter	
5.	A0A0H2URA1	Uncharacterized	PF17000	Accessory secretory protein Sec, Asp5	
6.	C1CA12	Uncharacterized	PF06612	Protein of unknown function	
7.	E0SVH7	Uncharacterized	PF11683, PF01595		
8.	E0TRL3	Uncharacterized			
9.	C1CAD7	Uncharacterized	PF07760		

Table 2: List of *Streptococcus pneumoniae* proteins located on membrane, cell wall or extracellular and identified as potential vaccine candidates in our analysis. Any previous studies which list these proteins as antigenic are also referred in the table. Pfam domains were identified with e-values threshold of 0.01.

S.No.	Uniprot ID	Protein name	Pfam domains	Functional description	References
1.	A0A0H2UP08	BlpN protein	PF10439	Bacteriocins	[39]
2.	A0A0H2UNS9	Bacteriocin BlpJ	PF03047		
3.	A0A0H2UNT2	Bacteriocin BlpI			
4.	A0A0H2UNH0	Choline binding protein J	PF01473	Choline Binding Proteins	[38]
5.	A0A0H2UMY8	Choline binding protein I	PF13340		
6.	B1I9H9	Choline-binding protein F			
7.	D6ZNQ4	Cell wall-binding repeat protein			
8.	C1C9R1	PblB	PF13884	Phage encoded virulence factor	[53]
9.	A0A0H2UR83	Conserved domain protein	PF16996	Accessory secretory protein Sec, Asp4	
10.	A0A0H2URJ0	Conserved domain protein	PF15432	Accessory secretory protein Sec, Asp3	

11.	B1IBV4	Chloramphenicol acetyltransferase	PF00302	Chloramphenicol acetyltransferase	
12.	A0A0H2UNM0	Cell wall surface anchor family protein	PF05738	Component of collagen-binding surface protein	
			PF13620		
13.	B2IQ89	Uncharacterized	PF02534	Type IV secretory system Conjugative DNA transfer	
			PF12696		
			PF10412		
			PF12846		
14.	D6ZLA2	Uncharacterized	PF12687	Protein of unknown function	
15.	E0SYC8	Uncharacterized			

Table 3: List of *Streptococcus pneumoniae* proteins classified as proteins of unknown location or cytoplasmic and identified as potential vaccine candidates in our analysis. Any previous studies which list these proteins as antigenic are also referred in the table.

Discussion

The identification of novel vaccines in a timely fashion is crucial for protecting the human population from the ever rising burden of the fatal infections. RV has already been applied to *S. pneumoniae* for identifying virulence factors *in silico*. However, the majority of the studies focus on proteins with signature motifs commonly found in surface exposed proteins [33–35]. Therefore, in the current study, we used nine virulent strains of *S. pneumoniae* and incorporated non-classically in addition to classically secreted proteins to identify potential antigens in *S. pneumoniae*. Using RV principles 83 surface exposed proteins in virulent and pathogenic strains of *S. pneumoniae* were identified. Among them, the notable ones were Bacteriocins, Choline binding proteins (Cbp) and Cytolysins. Bacteriocins are proteinaceous toxins produced by bacteria to kill or inhibit the growth of other bacteria [36]. Bacteriocins such as BlpN and BlpM have been found to be involved in interspecies competition between pneumococci during nasopharyngeal colonization allowing one strain to predominate others [37]. *S. pneumoniae* expresses many Cbp, and most of these proteins have repeats that help the attachment of the protein to the cell wall of the bacteria. Significantly reduced colonization of the nasopharynx has been reported due to mutations in Cbp like PspC, CbpD, CbpE, CbpG, LytB and LytC [38]. CbpG, a serine protease has also been reported to play a significant role in sepsis. Cholesterol-binding cytolysins are a large family of pore-forming toxins produced by many species of bacteria, including *S. pneumoniae*. Cholesterol is necessary for the cytolytic activity of this toxin hence; they are also called cholesterol-dependent cytolysin.

However, a significant proportion of the predicted virulence factors were uncharacterized. Therefore, InterProScan tool was run for the functional annotation and we could annotate nine uncharacterized and reannotate two conserved domain proteins. We describe below proteins, which were predicted as the potential vaccine candidates in our analysis.

Bacteriocins

Bacteriocins are small heat-stable, antimicrobial peptides produced by many gram-positive bacteria to colonize the host more efficiently by eliminating intra- or inter-species competition to the producer strain [36]. The human pathogen *S. pneumoniae* frequently colonizes nasopharynx and produces a large number of bacteriocins to eliminate the commensal flora, and therefore, these peptides are indirectly responsible for the pathogenic potential of the strains. A dedicated ABC transporter is thought to recognize these peptides and transport it across the cytoplasmic membranes. A bacteriocin known as pneumocin is a well-known pneumococcal virulence factor [39]. Inhibiting the virulence factors which render the pathogen harmless instead of killing it is potentially attractive treatment in wake of increased resistance to traditional antibiotics [40]. For example blocking the expression of cholera toxin by the virstatin markedly decreases the colonization of *Vibrio cholerae* in mouse models [41]. Therefore, we propose that bacteriocins like BlpJ, and BlpI with secretomeP score of 0.92 and 0.84 respectively, discovered in our analysis could serve as potential vaccine targets.

Choline binding proteins

The Cbp are a family of surface proteins bound to the cell wall of *Streptococcus pneumoniae* by phosphorylcholine moiety. Most of these proteins have repeats of up to 11 highly conserved 20 choline binding amino acid residues. Nearly 10 – 15 members of the family have been identified and characterized for their roles in virulence [38]. To best of our knowledge following choline binding proteins 'A0A0H2UMY8', 'B1I9H9', 'D6ZNQ4' with the secretomeP score of 0.88, 0.70 and 0.95 respectively, have not been characterized for virulence earlier in *S. pneumoniae*. Therefore, these proteins could be tested for their antigenic properties.

Cytolysins

Cytolysins are the substances secreted by microorganisms, plants or animals that are specifically toxic to individual cells causing their dissolution through lysis. One of the best characterized cytolysin is pneumolysin (Ply), a member of cholesterol-dependent cytolysin family produced by several gram-positive bacteria, including *S. pneumoniae* [42]. Pneumolysin allows bacterial invasion of tissues and mediating inflammation and activation of complement cascade. In our analysis, we could annotate an uncharacterized protein 'B1IB03' as the member of cholesterol-dependent cytolysin family. The protein has the PSORTb localization score of 9.67 and could be used as a new vaccine target

Mucin-binding proteins

Mucin-binding proteins or MucBP are surface proteins that are involved in adherence and colonization of human lungs and respiratory tracts during pneumococcal infections. MucBP have previously been characterized as adhesins in a number of pathogens, including *S. pneumoniae* [43, 44]. The MucBP 'B1ICR7' with the PSORTb localization score of 9.97 identified in our analysis could serve as a novel vaccine target.

Accessory secretion system proteins

The bulk of the secretions in bacteria occur via the general secretory (Sec) pathway. However, in gram-positive bacteria proteins which lack N-terminal signal peptide have been proposed to be secreted through an alternate system, known as the accessory secreted (secA2) system [45]. In our analysis, we found two conserved domain proteins 'A0A0H2URJ0' (Asp3), 'A0A0H2UR83' (Asp4) with the secretomeP score of 0.76, 0.90 respectively, and one uncharacterized protein 'A0A0H2URA1' (Asp5) as a part of accessary protein secretion machinery. Asp4 and Asp5 have been reported to share a high sequence similarity to secE (52%) and secG (55%) proteins of Sec system [45]. Thus, Asp4 and Asp5 might function as components of membrane translocase as in case of SecE and SecG. Although, we could not find any reports that link components of accessory secretion system proteins with the virulence in *S. pneumoniae*. Nonetheless, their presence in all the pathogenic strains and surface localization makes them worthy of further investigation.

Component of the collagen-binding surface protein

Surface proteins in bacteria are important virulence factors. Our analysis identified a protein 'A0A0H2UNM0' with the secretomeP score of 0.77 having the collagen-binding surface protein, B-type domain. This domain is thought to form a stalk of the collagen-binding surface protein that presents the ligand binding domain away from bacterial cell-surface [46]. We propose them to be novel vaccine targets in *S. pneumoniae* because of their membrane localization. Moreover, the members of the collagen binding surface proteins have been established as virulence factors in several Gram-positive bacteria [47].

Type IV secretory system conjugative DNA transfer

The success of *S. pneumoniae* as a major human pathogen has been attributed to the genome plasticity and its remarkable ability to escape antimicrobials and host immune response. Type IV secretory systems (T4SS) are multisubunit protein complexes traversing the cell envelop in many bacteria that mediate the transfer of proteins and nucleoprotein complexes across membranes, thus contributing to genome plasticity through dissemination of antibiotic resistance and virulence factors [48]. In our analysis, we found a protein 'B2IQ89' with the secretomeP score of 0.92 is a member of TraG protein family. TraG is essential for DNA transfer. We hypothesize that whole T4SS machinery or few individual proteins would have a role in virulence and thus can be used a novel vaccine target.

Conclusion

Reverse vaccinology technique has been successfully employed by many researchers for determining likely vaccine candidates in different pathogens. Our analysis of the nine virulent *S. pneumoniae* strains has resulted in the identification of novel proteins with antigenic potential viz. Bacteriocins, Choline Binding Proteins, Cholesterol-dependent cytolysins, Accessory secretory proteins, Collagen-binding surface protein, Type IV secretory system proteins. Some of these proteins have previously not been used as vaccine targets; therefore, we propose that they may have applications in development of effective vaccines to combat pneumonia. Further, in contrast to the classical approach the cytoplasmic proteins were not discarded and were categorized under non-classically secreted proteins if they passed all the criteria for non-classically secreted proteins set our pipeline. In addition, we were able to functionally annotate several uncharacterized proteins in all the nine pathogenic strains of *S. pneumoniae* under examination.

Among the possible limitation of RV method is the selection of false positives, due to accuracies of the software used, which is not optimal. Hence, further detail *in vitro* and *in vivo* studies need to be carried out to check the immune response of the predicted proteins for their efficient use as vaccines.

Author Contributions

AW curated the datasets and conducted the analysis work; VK directed the study and both authors prepared and approved the manuscript.

Conflict of Interest

None declared.

Acknowledgement

Authors would like to acknowledge Center for Nanotechnology Research and Applications (CENTRA) funded by The Gujarat Institute of Chemical Technology (Grant no. ILS/GICT/2013/003).

References

1. Bogaert D, De Groot R, Hermans PWM (2004) *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. Lancet Infect. Dis 4: 144–154.
2. Pneumonia and diarrhea (2015) Tackling the deadliest diseases for the world's poorest children | UNICEF Publications | UNICEF.
3. Walker CLF, Rudan I, Liu L, Nair H, Theodoratou E, et al. (2013) Global burden of childhood pneumonia and diarrhoea. Lancet Lond Engl 381: 1405–1416.
4. Jenkins SG, Brown SD, Farrell DJ (2008) Trends in antibacterial resistance among *Streptococcus pneumoniae* isolated in the USA: update from PROTEKT US Years 1–4. Ann. Clin. Microbiol. Antimicrob 7: 1.
5. Vasoo S, Singh K, Hsu LY, Chiew YF, Chow C, et al. (2011) Increasing antibiotic resistance in *Streptococcus pneumoniae* colonizing children attending day-care centres in Singapore. Respirol. Carlton Vic 16: 1241–1248.
6. Moberley SA, Holden J, Tatham DP, Andrews RM (2008) Vaccines for preventing pneumococcal infection in adults. Cochrane Database Syst. Rev CD000422.
7. Weinberger DM, Malley R, Lipsitch M (2011) Serotype replacement in disease after pneumococcal vaccination. Lancet Lond Engl 378: 1962–1973.
8. Lufesi NN, Andrew M, Aursnes I (2007) Deficient supplies of drugs for life threatening diseases in an African community. BMC Health Serv. Res 7: 86.
9. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, et al. (2006) In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. In Silico Biol 6: 43–47.
10. Seib KL, Dougan G, Rappuoli R (2009) The key role of genomics in modern vaccine and drug design for emerging infectious diseases. PLoS Genet 5: e1000612.
11. Christensen H, Hickman M, Edmunds WJ, Trotter CL (2013) Introducing vaccination against serogroup B meningococcal disease: an economic and mathematical modelling study of potential impact. Vaccine 31: 2638–2646.
12. Jones D (2012) Reverse vaccinology on the cusp. Nat. Rev. Drug Discov 11: 175–176.

13. Maritz-Olivier C, van Zyl W, Stutzer C (2012) A systematic, functional genomics, and reverse vaccinology approach to the identification of vaccine candidates in the cattle tick, *Rhipicephalus microplus* Ticks Tick-Borne Dis 3: 179–187.

14. John L, John GJ, Kholia T (2012) A reverse vaccinology approach for the identification of potential vaccine candidates from *Leishmania* spp. Appl. Biochem. Biotechnol 167: 1340–1350.

15. Baker MP, Reynolds HM, Lumicisi B, Bryson CJ (2010) Immunogenicity of protein therapeutics. Self Nonself 1: 314–322.

16. Vivona S, Bernante F, Filippini F (2006) NERVE: New Enhanced Reverse Vaccinology Environment. BMC Biotechnol 6: 35.

17. Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, et al. (2010) Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. Proc. Natl. Acad. Sci 107: 9072–9027.

18. Henderson B, Martin A (2011) Bacterial virulence in the moonlight: multitasking bacterial moonlighting proteins are virulence determinants in infectious disease. Infect. Immun 79: 3476–3491.

19. Ling E, Feldman G, Portnoi M, Dagan R, Overweg K, et al. (2004) Glycolytic enzymes associated with the cell surface of *Streptococcus pneumoniae* are antigenic in humans and elicit protective immune responses in the mouse. Clin. Exp. Immunol 138: 290–298.

20. Kolberg J, Aase A, Bergmann S, Herstad TK, Rødal G, et al. (2006) *Streptococcus pneumoniae* enolase is important for plasminogen binding despite low abundance of enolase protein on the bacterial cell surface. Microbiol. Read. Engl 152:1307–1317.

21. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43: D204–212.

22. Hoskins J, Alborn WE, Arnold J, Blaszczak LC, Burgett S, et al. (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol 183: 5709–5717.

23. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293: 498–506.

24. Lanie JA, Ng W-L, Kazmierczak KM, Andrzejewski TM, Davidsen TM, et al. (2007) Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. J. Bacteriol 189: 38–51.

25. Ding F, Tang P, Hsu M-H, Cui P, Hu S, et al. (2009) Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14. BMC Genomics 10: 158.

26. Camilli R, Bonnal RJ, Del Grosso M, Iacono M, Corti G, et al. (2011) Complete genome sequence of a serotype 11A, ST62 *Streptococcus pneumoniae* invasive isolate. BMC Microbiol 11: 25.

27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol 215:403–10.

28. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, et al. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinforma Oxf Engl 26: 1608–1615.

29. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S (2005) Non-classical protein secretion in bacteria. BMC Microbiol 5: 1–13.

30. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9: 62.

31. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, et al. (2007) Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol 189: 8186–8195.

32. Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.

33. Talukdar S, Zutshi S, Prashanth KS, Saikia KK, Kumar P (2014) Identification of potential vaccine candidates against *Streptococcus pneumoniae* by reverse vaccinology approach. Appl Biochem Biotechnol 172: 3026–3041.

34. Argondizzo APC, da Mota FF, Pestana CP, Reis JN, de Miranda AB, et al. (2015) Identification of proteins in *Streptococcus pneumoniae* by reverse vaccinology and genetic diversity of these proteins in clinical isolates. Appl Biochem Biotechnol 175: 2124–2165.

35. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, et al. (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect. Immun 69: 1593–1598.

36. Eijsink VGH, Axelsson L, Diep DB, Håvarstein LS, Holo H, et al. (2002) Production of class II bacteriocins by lactic acid bacteria; an example of biological warfare and communication. Antonie Van Leeuwenhoek 81: 639–654.

37. Dawid S, Roche AM, Weiser JN (2007) The blp Bacteriocins of *Streptococcus pneumoniae* Mediate Intraspecies Competition both *In Vitro* and *In Vivo* Infect. Immun 75: 443–451.

38. Gosink KK, Mann ER, Guglielmo C, Tuomanen EI, Masure HR (2000) Role of Novel Choline Binding Proteins in Virulence of *Streptococcus pneumoniae* Infect. Immun 68:5690–5.

39. Kadioglu A, Weiser JN, Paton JC, Andrew PW (2008) The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. Nat Rev Microbiol 6:288–301.

40. Allen RC, Popat R, Diggle SP, Brown SP (2014) Targeting virulence: can we make evolution-proof drugs? Nat. Rev. Microbiol 12:300–8.

41. Hung DT, Shakhnovich EA, Pierson E, Mekalanos JJ (2005) Small-molecule inhibitor of *Vibrio cholerae* virulence and intestinal colonization. Science 310:670–674.

42. Ogunniyi AD, Paton JC (2015) Vaccine Potential of Pneumococcal Proteins. *Streptococcus pneumoniae* 59–78.

43. Bumbaca D, Littlejohn JE, Nayakanti H, Lucas AH, Rigden DJ, et al. (2007) Genome-based identification and characterization of a putative mucin-binding protein from the surface of *Streptococcus pneumoniae*. Proteins 66: 547–558.

44. Du Y, He YX, Zhang ZY, Yang YH, Shi WW, et al. (2011) Crystal structure of the mucin-binding domain of Spr1345 from *Streptococcus pneumoniae* J Struct Biol 174: 252–257.

45. Rigel NW, Braunstein M (2008) A new twist on an old pathway – accessory Sec systems. Mol Microbiol 69: 291–302.

46. Deivanayagam CC, Rich RL, Carson M, Owens RT, Danthuluri S, et al. (2000) Novel fold and assembly of the repetitive B region of the *Staphylococcus aureus* collagen-binding surface protein. Structure 8: 67–78.

47. Kang M, Ko Y-P, Liang X, Ross CL, Liu Q, et al. (2013) Collagen-binding Microbial Surface Components Recognizing Adhesive Matrix Molecule (MSCRAMM) of Gram-positive Bacteria Inhibit Complement Activation via the C^{lassical} Pathway. J Biol Chem 288: 20520–20531.

48. Juhas M, Crook DW, Hood DW (2008) Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. Cell. Microbiol 10: 2377–2386.

49. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, et al. (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11: 107.

50. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, et al. (2010) A Catalog of Reference Genomes from the Human Microbiome. Science 328: 994–999.

51. Kerr AR, Paterson GK, McCluskey J, Iannelli F, Oggioni MR, et al. (2006) The Contribution of PspC to Pneumococcal virulence varies between strains and is accomplished by both complement evasion and complement-independent mechanisms. Infect Immun 74: 5319–5324.

52. Ogunniyi AD, Grabowicz M, Briles DE, Cook J, Paton JC (2007) Development of a Vaccine against Invasive Pneumococcal Disease based on combinations of virulence proteins of *Streptococcus pneumoniae* Infect. Immun 75: 350–357.

53. Hsieh YC, Lin TL, Lin CM, Wang -T (2015) Identification of PhlB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone. Sci Rep 5: 12265.

This article was originally published in a special issue, entitled: **"Industrial and Data Mining"**, Edited by S1