Canadian Journal of Microbiology

Vaccinomics driven proteome-wide screening of Haemophilus influenzae for the prediction of common putative vaccine candidates

Journal:	<i>Canadian Journal of Microbiology</i>
Manuscript ID	cjm-2020-0535.R2
Manuscript Type:	Article
Date Submitted by the Author:	23-May-2021
Complete List of Authors:	Bibi, Naseeha; National University of Sciences and Technology, Atta-ur-Rahman School of Applied Biosciences Zaidi, Najam-us-Sahar; National University of Sciences and Technology, Atta-ur-Rahman School of Applied Biosciences Tahir, Muhammad; National University of Sciences and Technology, Atta-ur-Rahman School of Applied Biosciences Babar, Mustafeez; Shifa Tameer-e-Millat University, Shifa College of Pharmaceutical Sciences
Keyword:	H. influenzae, Putative protein target, B and T cell epitopes, Immunoinformatics, Vaccine candidates
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

SCHOLARONE™
Manuscripts

Title:**Vaccinomics driven proteome-wide screening of *Haemophilus influenzae* for the prediction of common putative vaccine candidates****Authors' List and affiliations:** Naseeha Bibi¹, Najam-us-Sahar Sadaf Zaidi¹✉, MuhammadTahir¹, Mustafeez Mujtaba Babar ²¹ Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences 8 and Technology, Islamabad 44000, Pakistan.² Shifa College of Pharmaceutical Sciences, Shifa Tameer-e-Millat University, Islamabad.

44000, Pakistan.

✉ **Corresponding author:** Najam-us-Sahar Sadaf Zaidi, Associate Professor, Atta-ur-

Rahman School of Applied Biosciences (ASAB), National University of Sciences and

Technology, Islamabad 44000, Pakistan. Tel : +92-51-9085-6132

E-mail✉: zaidi.sahar@gmail.com; sadafzaidi@asab.nust.edu.pk

Abstract

Haemophilus influenzae colonizes the respiratory tract and is associated with life-threatening invasive infections. The recent rise in its global prevalence, even in the presence of multiple vaccines, indicate an urgent need for developing cross-strain effective vaccine strategies. Our work focused on identifying the universally conserved antigenic regions of *H. influenzae* that can be used for developing new vaccines. A variety of bioinformatics tools were applied for the comprehensive geno-proteomic analysis of *H. influenzae* type “a” strain, as reference serotype, through which subcellular localization, essentiality, virulence, and non-host homology were determined. B and T-Cell epitope mapping of 3D protein structures were performed. Thereafter, molecular docking with HLA DRB1*0101 and comparative genome analysis established the candidature of identified regions. Based on the established vaccinomics criteria, five target proteins were predicted as novel vaccine candidates. Among these, 9 epitopic regions were identified that could regulate the lymphocyte activity through strong protein-protein interactions. Comparative genomic analysis exhibited that the identified regions were highly conserved among the different strains of *H. influenzae*. Based on multiple immunogenic factors, the five prioritized proteins and their predicted epitopes were identified as the ideal common putative vaccine candidate against typeable strains.

Keywords: *H. influenzae*; Immunoinformatics; Vaccine candidates; Putative protein target; B and T cell epitopes.

1. Introduction

Haemophilus influenzae is a Gram-negative coccobacillus that normally colonizes respiratory tract of healthy individuals. However, under certain conditions, it can also be an important cause of severe invasive infections (Berndsen et al. 2012; Cox et al. 2017; Kelly et al. 2011). The pathogenesis of the human-restricted bacterium *H. influenzae* initiates in the respiratory tract and it can invade other tissues of the body (Pinto et al. 2018; Krishnan et al. 2015). Among the infections caused by this bacterium, community acquired pneumonia (CAP), conjunctivitis, epiglottitis, septicaemia, otitis media, meningitis, bronchitis and sinusitis are of major clinical importance (Tsang et al. 2017; Ulanova et al. 2014) and particularly common in adults and children (Wroblewski, et al., 2013). Its infection has also been associated with cystic fibrosis and chronic obstructive pulmonary disease (Pinto et al. 2018; Mohd-Zain et al. 2015). Globally, respiratory diseases associated with *H. influenzae* cause an estimated 1.9 million deaths per year (Kumavath et al. 2014). According to the Active Bacterial Core surveillance (ABCs) CDC 2018 report the percentage of meningitis, bacteraemia, and pneumonia cases was reported as 7.1%, 25.1% 53.1% respectively with average death rate of 12.6 per 100 cases (CDC Centers for Disease Control, 2018). Rate of permanent sensorineural hearing loss due to meningitis is reported to be 6% of all infected patients. Epiglottitis, secondary to a *H. influenzae* acute respiratory tract obstruction, carries a mortality rate of 5-10% while neonatal *H. influenzae* disease carries a mortality rate of 55% (Joseph Adrian L Buensalido, 2019).

H. influenzae strains can be classified into two types; encapsulated (typeable) and non-encapsulated strains (non-typeable; NTHi) based on the antigenic properties of their capsular polysaccharides (Kelly et al. 2011; Collins et al. 2018). The typeable strains, based on their unique capsular antigens, can further be categorized into six serotypes (serotypes a to f) (Konini et al. 2016; Pinto et al. 2018). Owing to their broad biogeographical distribution, infections by both typeable and non-typeable *H. influenzae* strains have been reported throughout the world (Murphy 2015; Brown et al. 2009). The

most virulent serotype among the six typeable strains is the *H. influenzae* type b (Hib) followed by Hia (*H. influenzae* type a) as the second most virulent serotype (Eton et al., 2017). High frequency of Hib disease is reported among the African Americans, American Indians, and Australian Aborigines. Despite Hib vaccination, the Hib disease rate among rural Alaskan native children is quite high with an overall rate of 5.4 per 100,000, in Queensland, Australia, the incidence rate among children under 5 years of age was reported to be 7.4 per 100,000. Interestingly, the incidence rate among children of indigenous descent is 10.2 per 100,000. (Troeger et al., 2018). Similar trends have been found in the north-American population as well (Tsang et al., 2017).

Till date three vaccines have been approved by the Food and Drug Administration (FDA), USA which include Pedvax HIB, ActHIB and Hiberix which were licensed for clinical use in 1989, 1993 and 2009 respectively (CDC Centers for Disease Control, 2020). These vaccines are either polysaccharide or toxoid based conjugate vaccines that do not confer complete protection against all *H. influenzae* serotypes. Significant increase in morbidity and mortality, due to other serological types besides Hib, strongly suggest strain replacement in the post-Hib vaccines era as the strain has been eradicated from most of the geographical regions (Ulanova, 2013). Invasive *H. influenzae* type a (Hia) disease, for instance, has emerged in Alaska's indigenous population (Bruce et al., 2013). In a study of invasive *H. influenzae* in Northwestern Ontario, Canada, the *H. influenzae* serotype a (Hia) was found to be the most common one indicated in 50% of the samples while around 20% of the infections had a non-typeable *H. influenzae*. The case-fatality rate was 10% for invasive *H. influenzae* type a which exceeded the pre-Hib vaccine incidence rate of *H. influenzae* type b infections (Eton et al., 2017). This high prevalence prompts the need for novel prophylactic strategies (Pinto et al. 2019; Ribeiro et al. 2003). Due to the appearance of drug resistant strains of the bacterium, its infection remains a major health concern worldwide (Tsang et al. 2017). Moreover, the phenomenon of capsule switching or replacement of *H. influenzae* strains following its widespread vaccination has led its strains to undergo mutations to fill the previously occupied ecological niche by Hib (*H. influenzae* type "b").

These factors, therefore, stress the need for designing and developing effective vaccine candidates against the bacterial strains (Ulanova et al. 2012; Behrouzi et al. 2017). Of the various immunogenic biomolecular components, utilizing outer membrane proteins, can serve as an effective candidate for vaccine development (Van Eldere et al. 2014).

Information about the genomic structure of pathogenic bacteria has significantly increased due to availability of whole genome sequencing (WGS) technologies (Pinto et al. 2019; Eutsey et al. 2013; Janto et al. 2014). Additionally, the exploitation of bioinformatics tools and their application in vaccinology have led to the design of effective vaccine candidates (Ali et al. 2019). These techniques in addition to the reverse vaccinology approach has facilitated in performing detailed *in silico* analysis to prioritize the most likely protective antigens effectively in a time and cost-effective manner (Adhikari et al. 2018; Xiao et al. 2018). A properly designed computational workflow can, hence, facilitate the process and the accuracy of prediction of potential antigens that can serve as appropriate vaccine candidates (Droppa-almeida et al. 2018).

In the present work, a detailed and comprehensive analysis of the reference genome of invasive *H. influenzae* type “a” strain has been performed for the identification of putative, prioritized, molecularly stable vaccine candidates. The bioinformatics analyses included identification of subcellular localization to screen out proteins present in outer membrane region followed by their essentiality check and determination of virulent nature of the targeted proteins. Among the other parameters which were analysed for the selection of putative vaccine candidates are 3D structure prediction to locate positions of the B and T-Cell epitopes on surface of the protein targets. Afterwards, putative protein targets were identified using multiple parameters and their candidature to be used as a vaccine candidate was established. The significant findings in this study can support rational vaccine design and enhance our understanding of vaccine design and development processes..

2. Methodology

The overall workflow of the whole methodology has been summarized in Figure 1. Detailed information about the steps has been provided in the following sections.

2.1 *Haemophilus influenzae* Strain Selection and retrieval of primary sequence data as reference genome

Among the six typeable serotypes of *H. influenzae* strains, the reference strain of *H. influenzae* serotype “a” NML-Hia (GenBank accession: NZ_CP017811) was selected to perform detailed immunoinformatic analysis by using different online available bioinformatics tools. In recent years, *H. influenzae* serotype “a” has emerged as a significant cause of invasive diseases. The reference strain has been completely sequenced in 2017. Primary genomic and proteomic data of *H. influenzae* NML-Hia-1 strain was retrieved from NCBI (GenBank accession: NZ_CP017811) as a RefSeq (Iskander et al. 2017).

2.2 Proteome Screening for Subcellular Localization

Determination of Subcellular localization of pathogenic proteins is one of the important considerations for potential vaccine candidates (Naz et al. 2015). *H. influenzae* type “a” proteome was screened for subcellular localization using programs PSORTb v 3.0 and CELLO for the identification of extracellular and outer membrane proteins (Yu et al. 2014; Yu et al. 2010). The program for determining the subcellular location PSORTb 3.0.2 (available at <http://www.expasy.org/proteomics>) consists in multiple analytical modules, each one analysing a biologic resource known for influencing or being typical of subcellular location. A second software CELLO (<http://cello.life.nctu.edu.tw/>) was also used for confirmation of the predicted outer membrane and extracellular protein locations. Based on support vector machines (SVM), CELLO program is trained to localize bacterial proteins within five modules including cytoplasmic, periplasmic, extracellular, outer-membrane and inner-membrane with 89% prediction accuracy. Identification of proteins locations as outer membrane proteins was followed by screening in the next

step for other parameters including essentiality, virulence, antigenicity, allergenicity, physico-chemical analysis (e.g. molecular weight, isoelectric points, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index), 3D structures analysis, non-host homology, transmembrane helices analysis, protein-protein interactions (PPI), epitope mapping, docking and multiple sequence alignment (MSA) analysis.

2.3 Screening of Essential Genes

In the next step, only those proteins which were predicted and localized as outer membrane, periplasmic and extracellular were checked for their essentiality by using Database of Essential Genes (DEG) (<http://tu-bic.tju.edu.cn/deg/>) version 10.4 to extract essential proteins (Luo et al. 2014). Some important essential genes required by pathogen to carry out major functions, e.g. genes for their survival, infection, adhesion, entry and persistence in the host, can be revealed by essentiality check which is considered as an important parameter for identification of potential targets (Naz et al., 2015; Rashid et al., 2017).

2.4 Genetic Diversity of *H. Influenzae* Pathogenesis Associated Genes

The genetic sequences of prioritized surface exposed outer membrane, periplasmic and extracellular essential proteins were analysed for the genetic diversity of genes known to be involved in *H. influenzae* pathogenesis. Virulence Factor Database (VFDB) (<http://www.mgc.ac.cn/VFs/main.htm>) was used for identification of potential virulence genes and proteins. The VFDB server provides various classes of virulence factors: adherence, antimicrobial activity, bio-surfactant, anti-phagocytosis, and protease, toxins, and iron uptake, lipase and secretion systems.

2.5 Physico-Chemical Analysis

Determination of the different physico-chemical properties that indicate molecular stability were computed by using ExPasy ProtParam server (<http://web.expasy.org/protparam>). The computed parameters included the molecular weight, theoretical pI (isoelectric point), amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and

grand average of hydropathicity (GRAVY) for the prioritized surface exposed essential proteins and effective targets due to ease of purification and development process.

2.6 Comparative Analysis with Human Genome

The National Center for Biotechnology Information (NCBI) Protein Basic Local Alignment Search Tool (BLASTp) server was used to identify and exclude potential human host homologues. Ideally, the identified proteins should not be homologous to host proteome as they might cause autoimmunity response. The cut off criteria was set at 0.005 E-value and 35% sequence identity for about 35% or more query coverage. Proteins with greater values than set criteria were considered unsuitable for vaccine targets and removed from data set (Rashid et al. 2017).

2.7 Determination of Antigenic potential and allergenicity of target proteins

To verify the immunogenic potential of the proteins predicted by the various tools cited earlier, their sequences were checked for antigenicity by using VaxiJen v2.0 using threshold value = 0.4 (<http://www.ddgpharmfac.net/VaxiJen/VaxiJen/VaxiJen.html>). This online tool allows independent alignment prediction of protective antigens and is based solely on antigens' physico-chemical properties with 70% to 89% precision based on the target organism. Those proteins having a score of >0.4 are considered to be potentially antigenic. Considering the score of antigenic proteins, they were further analysed for the prediction of B and T-cell epitopes. The allergenicity of the targeted proteins was checked by online tool AllerTop v.2.0 (<https://www.ddg-pharmfac.net/AllerTOP/index.html>). These tools use the method based on auto cross covariance (ACC) transformation of protein sequences into uniform equal-length vectors. (Anal. Chim. Acta 1993;277:239-253).

2.8 Determination of three-dimensional (3D) Protein Structure

An efficient way to visualize, understand and infer the relationship of immunogenic proteins with other proteins is to provide or alternatively determine the 3D structure of a protein. Protein Data Bank (PDB) was checked for availability of 3D structures of the candidate antigenic proteins predicted in the earlier steps of our analysis. However, among the five prioritized proteins only

peptidoglycan associated lipoprotein (pal omp p6) was found to have a protein structure determined through physical methods. The 3D structure of the rest of the proteins was generated by using Protein Fold Recognition Server, commonly referred to as Protein Homology/ analogy Recognition Engine (Phyre 2), I-TASSER server and SWISS-MODEL (Yang et al. 2015; Waterhouse et al. 2018). Moreover, the 3D structure of Pal OMP-P6 available on Protein Databank was compared with the predicted models. It was found that I-TASSER models had the best structural homology. The best models generated by I-TASSER with high identity with the template were selected for transmembrane helices analysis.

2.9 Transmembrane Helices Analysis

The total number of transmembrane helices were analysed within the prioritized targeted proteins by TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>) using default values of their parameters. Proteins having 2 or more transmembrane helices were not considered to be good targets because such proteins due to the presence of more than one transmembrane helical region usually fail to express during *in vitro* and *in vivo* bioprocessing.

2.10 Protein–protein interaction information

Connectivity network of immunogenic proteins needs to be considered for the thorough understanding of the biological phenomena as the backbone of the cellular machinery is formed by proteins and their functional interactions. Interaction network of the prioritized proteins was analysed by using STRING v11 online server (Szklarczyk et al. 2016).

2.11 Epitope mapping B-Cell and T-Cell Epitope Predictions

After performing detailed analysis of the protein structure, the targeted proteins were prioritized based on the best scores and values for all the essential parameters. Their sequences were submitted to different prediction tools for determining their capability to induce the immunogenicity through B-Cell and T-Cell activation. Elli pro server (<http://tools.iedb.org/ellipro/>) was used for prediction of both linear and discontinuous epitopes based on the antigen's 3-D structure (Soria-Guerra et al. 2015).

In addition, BCPreds server was also used to predict 20mer epitopes and those having a score of greater than 0.8 were further analysed for overlapping T-Cell epitope predictions. In comparison to the conventional vaccines, a more effective and safer option is the peptide-based vaccine as it includes only the immunogenic epitopes. Predicting the epitopes with higher binding potential for Major Histocompatibility Complex (MHC) is necessary to design peptide vaccines as the immunogenic potential of an antigen is primarily dependent on the MHC for presentation to the immune cells (Rashid et al., 2017). Hence, predicting epitope binding to both MHC class I and II is imperative for effective vaccine prediction (Rizwan et al., 2017). The MHC class I and II binding prediction tools (<http://tools.iedb.org/mhci/>) were used to predict the T-Cell Epitopes and their interaction with different MHC- I and II alleles by using default IEDB recommended method against available set of reference alleles (Andreatta et al. 2016). The available set includes MHC-I alleles HLA_A*0101, HLA_A*0201, HLA_A*0202, HLA_A*0203, HLA_A*0206, HLA_A*1101, HLA_A*3301, HLA_A*0301, HLA_A*6801, HLA_A*6802 and HLA_A*3501; MHC II alleles DRB1*0101, DRB1*0401 and DRB1*0701. The half-maximal inhibitory concentration (IC₅₀) values were calculated by MHCpred. The IC₅₀ values required for the peptide's binding to the specific MHC-I or II molecules were set at $\leq 100\text{nM}$. Peptides with low IC₅₀ value were proposed to be more potent MHC-I and II epitopes (Droppa-almeida et al. 2018). The epitopes with the strongest binding were visualized on their respective 3D predicted protein models by using Discovery Studio v.4.1 (Dassault Systemes, 2016).

2.12 Docking Technique; Epitopes and DRB1*0101 Allele Receptor Binding

Those T-cell epitopes interacting with more than 15 MHC class I and II molecules, especially DRB1*0101 allele, were pooled out and proceeded for docking. DRB1*0101 is used for docking of the predicated T-cell epitopes as this MHC-II allele is the most frequently bound component and in order to produce better antigenic response epitope should bind to DRB1*0101. To verify the anchoring and to explore the interaction between T-Cell predicted epitopes and DRB1*0101 allele

receptor the homology modelling was performed for all the selected epitopes and DRB1*0101 allele receptor. Protein sequence and PDB ID of DRB1*0101 were retrieved from the NCBI database and submitted to Raptor X server (<http://raptorx.uchicago.edu/>) for homology modelling (Källberg et al. 2016). To display the predicted 3D structure of DRB1*0101 alleles, Chimera Visualisation software 1.13.1 was used (Pettersen et al. 2004). The homology modelling of the MHC I predicted epitopes was performed by using online available server PEP-FOLD3 (<http://bioserv.rpbs.univparisdiderot.fr/services/PEP-FOLD3/>) to predict the linear structures of the amino acid sequences (Lamiable et al. 2016). Best models were selected by using FireDock (<https://bioinfo3d.cs.tau.ac.il/FireDock/php.php>)

Based on the peptide-binding groove affinity, molecular docking was performed between DRB1*0101 allele and the predicted proposed epitopes. DRB1*0101 allele was set as macromolecule receptor and the predicted peptide epitopes were set as ligands. Molecular docking of 3D structure of DRB1*0101 alleles and 3D modelled epitopes was performed by using PyRx (Dallakyan & Olson, 2015). The process was performed by submitting PDB model files of 9 predicted epitopes as ligands and DRB*0101 receptor as macromolecule. For visualization of docked complex in 2D and 3D Discovery studio visualizer tool was used .

2.13 Core-Genome-Based Phylogenetic and Diversity Analysis

Reference sequences of different strains and serotypes were retrieved from NCBI and arranged in FASTA file format for analysis of conservancy of the predicted selected proteins. Alignment of all retrieved sequences was performed by using ClustalW through BioEdit software. Predicted epitopes were then investigated in retrieved reference sequences after MSA for conservancy. Phylogenetic trees of these aligned reference sequences were created by using MEGA X (<https://www.megasoftware.net/>). to analyse the evolutionary divergence of these sequences. The evolutionary history was inferred using the Neighbour-Joining method keeping values for all parameters including the bootstrap replications as default.

3. Results

3.1 Phenotypic and Genomic features

The focus of the current study is on the typeable strains of *Haemophilus influenzae*, a gram-negative coccobacillus bacterium. For the purpose, the reference strain *Haemophilus influenzae* type “a” was selected for the comprehensive analysis for identification of the putative vaccine candidates owing to its high pathogenicity and prevalence. Whole-genome sequence of a *H. influenzae* serotype “a” strain NML-Hia-1, isolated from a blood culture specimen, was used as reference genome (GenBank accession: CP017811) which is a single 1,829,217 bp (1.82Mbp) Hia chromosomal DNA with an average 38.02% GC contents. Currently, Hia is not known to contain plasmids. The annotation predicted a total of 1,781 genes with only 1,643 coding genes sequences, 19 rRNA operons, and 59 tRNA genes (Iskander et al. 2017).

3.2 Subcellular Localization

Synergistic use of CELLO and PSORTb tools generated results with slight variations in the location of the identified proteins. Localization analysis results by using CELLO found that 45.26% proteins are located in cytoplasm. In comparison only 2.59%, 2.77% and 0.31% were in outer membrane, periplasmic space and extracellular regions respectively while the exact location of 21.31% of proteins could not be predicted. In comparison PSORTb localization analysis results found 44.70% of the proteins located in cytoplasm while 2.46%, 2.34% and 0.18% were present in the outer membrane, periplasm and extracellular regions. The location of 20.32% was unknown (Figures 2a & 2b). Around 40 proteins were detected as outer membrane proteins by the synergistic use of both tools. The distribution pattern and functional characterization of the bacterial proteome have been presented in Supplementary Figure 1 and 2.

3.3 Essentiality and Virulence

The screening and scrutinizing process of the 40 outer membrane proteins identified in the analysis, 37 were found to be essential whereas 41 of the 47 extra cellular and outer membrane proteins were

essential for the survival and metabolism of the bacteria. Furthermore, the 37 essential outer membrane proteins were further checked for their virulent nature. The extracellular essential proteins were excluded from the study. Among the outer membrane essential proteins only 5 were predicted as virulent. These 5 proteins were screened out and selected for further analysis targets. They include peptidoglycan associated lipoprotein (Pal OMP P6), glutamate dehydrogenase (Gdh), Transferrin binding protein (tbp1), Type IV pilus biogenesis/stability Protein (PilW) and porin ompA (Supplementary Table 3).

3.4 Comparative Analysis with Human Genome; Analysis of Homology Models

The shortlisted five outer membrane proteins of *H. influenzae* strain “a” (NML Hia, (RefSeq CP 017811.1) were subjected for homology analysis for human host homologues against *Homo sapiens* (taxid: 9606) using BLASTp to predict their ability to cause autoimmune responses. BLASTp results (Table 1) of the targeted proteins against proteome of the human host revealed them to be non-homolog. The peptidoglycan associated lipoprotein or PAL OMP P6 showed no significant similarity in BLASTp against *Homo sapiens* proteome. The BLASTp results for Glutamate dehydrogenase (Gdh) showed only 28% identity with chain A of human protein Glutamate Dehydrogenase 1. The transferrin binding protein (Tbp1) showed only 35% identity with dynamin 3, isoform CRAg [*Homo sapiens*] and query coverage of only 5%. Type IV pilus biogenesis/stability Protein (PilW) showed only 35% identity with human protein Y-linked transcript variant 115 whereas porin OmpA Protein BLASTp results show no significant similarity against human host proteome.

3.5 Physico-Chemical Analysis

The targeted proteins were further analysed and checked for some physicochemical parameters that can contribute to the stability of the vaccine candidates. These parameters include molecular weight, isoelectric point (pI) value, functional annotation, estimated half-life, instability index, aliphatic index, Grand Average of Hydropathy (GRAVY) and antigenicity scores for the further shortlisting as potential vaccine candidates (Table 1). Molecular weight of <110 kDa makes the targeted proteins

more appropriate potential candidate due to the ease of purification and development process. Molecular weights of the selected five proteins viz. Pal OMP -P6, Gdh, *tbp1*, *pilW* and porin *omp A* were calculated by ExPASy protpram tool. Molecular weight of Pal OMP -P6 was calculated to be 16kDa with pI value 6.09. The molecular weight of Gdh, *PilW* and Porin *OmpA* were calculated to be 48kDa, 20kDa and 37kDa respectively, whereas pI value for these proteins was found to be 6.02, 8.36 and 9.62 respectively. The *Tbp1* protein was of largest size among the five shortlisted proteins bearing a molecular weight of 103kDa and a pI value 9.34. The instability index (II) provides an estimate of the stability of protein. The protein molecules with an instability index smaller than 40 is predicted as stable while a value above 40 predicts that the protein may be unstable. The aliphatic index (AI) of a protein may be regarded as a positive factor for the increase of thermostability of globular proteins. The half-life is a prediction of the time it takes for half of the amount of protein in a cell to degrade in the cells like reticulocytes for *in vitro* analysis and for *in vivo* analysis (Gasteiger et al., 2005). The values for each of the mentioned parameters are given in the Table 1.

3.6 Determination of Antigenicity Score and allergenicity for the Targeted Proteins

The antigenicity score of the selected proteins was determined by using VaxiJen 3.0 server. The antigenicity scores for the identified proteins have been presented in Table 1. A threshold base value of 0.4 was set for considering the bacterial proteins as probable antigens. The highest score was obtained for porin *OmpA* with 0.6150 value followed by Pal OMP -P6 Protein with value of 0.6148 score. Antigenicity scores for *PilW* and *Tbp1* protein were 0.6058 and 0.5989 respectively, whereas the lowest antigenicity score was found to be 0.5364 for Gdh protein. All the selected proteins proved their candidature to be probable antigens for eliciting immune responses. Bioinformatics tool for allergenicity prediction AllerTOPv. 2.0 was used to determine the nature of the candidate proteins. All of the proteins were classified as Non-allergen except the Pal OMP P6 which was classified as a probable allergen with its resemblance to the UniProtKB accession number P10324.

3.7 Determination of three-dimensional (3D) Protein Structure

Three-dimensional protein structures are crucial to understand and infer their role and interaction with other ligands, receptors, and protein molecules. 3D structure prediction tools I-TASSER, and Phyre servers were used to predict and compare the build models (Figure 3). Pal OMP-P6 protein was found to have an experimentally validated 3D structure submitted in PDB. Structures of the other four shortlisted proteins were predicted by using I-TASSER online server (Yang et al. 2015) and have been provided in Figure 3. Iterative Threading Assembly Refinement (I-TASSER) is a hierarchical approach for prediction of protein structure and function by first identifying structural templates from the PDB followed by derivation of function insights of the target by threading the 3D models through protein function database ([BioLiP](#)). The PDB model files generated by modelling tools for all the five proteins structure were then imported to Discovery studio for visualization (Figure 3).

3.8 Prediction of Transmembrane Helices in Proteins

TMHMM server 2.0 online tool was used to predict number of transmembrane helices which should ideally be less than 2 and for best targets, they should not exceed 1. It is established that the colonization and expression become harder and difficult for proteins with large number of transmembrane helices. All five proteins Pal OMP -P6, Gdh, PilW, Tbp1 and porin omp A were predicted with no transmembrane helices and the whole sequences were labelled as 'outside' implying thereby that the proteins contain no membrane helices.

3.9. Protein-Protein Interaction Analysis and Functional Enrichment Networks

Functional interactions of Proteins form the backbone of the cellular machinery. Full understanding of biological phenomena needs to consider their interaction network (Szklarczyk et al. 2019). Important signalling pathways and crucial molecular cross talks can reveal many stimulating factors between potential therapeutic targets at cellular level (Naz et al. 2015). STRING v.11.0 was used to construct interaction maps for all prioritized proteins (Figure 4a-4e). It is a database containing the

established and predicted protein-protein interactions including direct i.e. physical and indirect i.e. functional associations. These interaction networks stem from various computational predictions (Szkłarczyk et al. 2019). Pal OMP -P6 protein interaction analysis results showed direct interaction of Pal OMP -P6 with other members of ompA including tolA, tolB, tolQ and tolR as well as omp2 and ompA family (Figure 4b). Bio functional analysis of PAL OMP -P6 revealed it to be involved in many important biological activities such as transport of peptides (GO: 0015833), proteins (GO: 0015031), Bacteriocins (GO: 0043213) and lipoproteins (GO: 0042953). Among other biological functions of PAL OMP -P6, cell division (GO: 0051301), Cell cycle (GO: 0007049) and protein import (GO: 00170388) are worth-mentioning. The topological analysis of the Tbp1 constructed network indicated interaction with *tbp2*, *omp2* and other *tonB* like essential proteins (Figure 4e). Its involvement in iron-ion transmembrane transport (GO: 0034255) and pathogenesis (GO: 0009405) clearly call for their prioritization that may lead to the identification of novel potential vaccine targets. Functional network analysis for Gdh interaction map (Figure 4a) revealed that it is involved in important biological process including biosynthesis of small molecules (GO: 0044283) and carboxylic acid (GO: 0046394), and metabolism of small molecules (GO: 0044281) and carboxylic acid (GO: 0019752). KEGG pathways analyses of Gdh showed its involvement in many important pathways including arginine biosynthesis pathway (*hix00220*), metabolic pathways of nitrogen (*hix00916*), alanine, aspartate and glutamate biosynthesis pathways (*hix00250*) and TCA cycle (*hix00020*). Similar topology for other targeted protein porin ompA (P5) was also predicted (Figure 4c) including biological processes of peptide metabolism (GO: 0006518), translation process (GO:0006412), Binding of RNA (GO:0003723) and rRNA (GO: 0019843). While the interacting map for *pilW* protein particular protein did not show any specific biological process due to a lack of data record, however the constructed map showed its interaction with some of the important proteins (Figure 4d) such as those involved in DNA transformation, Protein transport, transformation protein TfoX, protein HofC homolog, Competence protein E; and transformation processes.

3.10. Epitope Mapping

Validation of all the essential parameters and 3D structure model prediction resulted in selection of five prioritized proteins as target vaccine candidates (Figure 3). Antigenicity analysis of the selected proteins already revealed them to be antigenic. In principle, targeted proteins should initiate B-cell and T-cell mediated immunity. Elli pro server used for B-Cell epitope predicted linear and discontinuous epitopes along with their respective positions on 3D predicted model. Ellipro associates each predicted epitope with a score defined as a PI (Protrusion Index) value. Those epitopes with a score of greater than 0.75 were shortlisted for further T-Cell epitope prediction analysis (Supplementary Table 4). BCpred server was also used to predict overlapping B and T-Cell epitopes and those B-Cell epitopes having a score greater than 0.8 were shortlisted for T-Cell Epitope analysis (Supplementary Table 4). Thirty epitopes from BCpred prediction results of prioritized protein having the highest antigenicity score of ≥ 0.8 were further analysed for their high binding affinity with MHC-class I and II alleles specifically with DRB1*0101 by using MHCpred tool. IC₅₀ i.e. half-maximal inhibitory concentration values, for the selected peptides were determined by MHCpred based on the binding predictions as per the input parameters provided. In case of peptide selection based on IC₅₀ values, lower value indicates better binding affinity (Paul et al. 2016). Epitopes showing high binding affinity for DRB1*0101 allele may produce a good immunogenic response due to its wider distribution among human population groups (Naz et al. 2015). The IC₅₀ values required for the peptide's binding to the specific MHC-I molecules were set less than or equal to 100nM. Of these nine best scoring overlapping B and T-Cell epitopes were selected and have been presented in Table 2. These have been based on their ability to bind with maximum number of MHC alleles (total MHC I and II binding alleles = 30), lowest IC₅₀ value for DRB1*0101 and antigenicity (VaxiJen scores). Visualization of the nine epitopic regions with one letter code and specifier on

surfaces of their respective 3D protein structure was performed by using Chimera visualization tool 1.13.1. (Figure 5).

3.11. Docking of DRB1*0101 Human Allele and Proposed Epitopes

Protein sequence of human DRB1*0101 allele was retrieved from Uniprot with PDB ID: Iaqd and I-TASSER was used for its 3D structure prediction. For molecular docking of the DRB1*0101 allele, only beta chain sequence from UniProt database was used to predict its 3D structure due to ligand binding positions in Beta chain (Figure 6a & b). For docking, PyRx tool was used to check the binding interactions of the nine selected epitopic regions with the binding sites on the surface of the beta chain of the DRB*0101 allele (Figure 6a & 6b). The ligand and receptor input files were used in the PDB format. The homology modelling of the selected predicted epitopes was performed with PEP FOLD3 to predict the linear structures from amino acid sequences. Molecular docking was performed according to peptide-binding groove affinity, between DRB1*0101 and selected proposed peptides, the active pockets residues from beta chain of the DRB*0101 allele for TE3 peptide ligand are shown in Figure 6b which involve CYS (B:44), ASN (B: 111), PHE (B:46), GLY (B:115), VAL (B:120), TYR (B:152), GLY (B:49) and TYR (B:112). Human DRB1*0101 allele was set as receptor, while the screened peptides were set as ligands. The analysis was performed by using the default values for the exhaustiveness as 8. The best binding score was obtained for the TbE2 as -7.8kcal/mol and the least value for the gdhE2 as -5.1Kcal/mol. The binding energy scores are mentioned in the Table 2. Excellent binding energy score values with the DRB1*0101 allele were obtained for almost all the alleles used in the study. Visualization of the binding patterns and the active site residues was performed by using Discovery Visualization offline tool, which indicated excellent binding ability of the peptide with the beta chain of the selected allele (Figure 6a&b).

3.12 Genetic diversity and meta-analysis by Sequences Alignment

The five prioritized protein sequences were proceeded for evaluating the degree of sequence alignment by using NCBI BLASTp results for different strains of *H. influenzae*, (Supplementary

Tables 5-9). The alignment also helped in predicting the conserved regions (Supplementary Figures 3-14). The alignment of all retrieved sequences using ClustalW through BioEdit software showed high conservancy between the aligned sequences of different strains. The conserved regions were recognized by identity and similarity of amino acid sequences. Phylogenetic trees were created using MEGA X (<https://www.megasoftware.net/>). The evolutionary history for the prioritized proteins was inferred using the Neighbor-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) were selected for the analysis. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The evolutionary divergence analysis of the peptidoglycan associated protein (Pal OMP P6) sequences of the different strains of *H. influenzae* resulted in clustering of different strains sequences in a single clade which indicated its monophyly and conservancy across its evolutionary history. Similar results were generated by phylogenetic trees for glutamate dehydrogenase and pilus type IV proteins, where different strains were clustered as a single clade representing them as monophyletic groups. However, evolutionary divergence analysis of the transferrin binding protein and porin omp A did not generate single clades for large number of strains. The phylogenetic analysis, hence, indicated that Pal OMP -P6, gdh and pilus type IV proteins are highly conserved among large number of strains of *H. influenzae* and are ideal vaccine candidates.

4. Discussion

H. influenzae is a medically important bacterium that is responsible for a high proportion of childhood pneumonia, meningitis, and sepsis (Zaidi et al. 2010). According to the recent epidemiological data, *Hib* is the leading cause of childhood meningitis in most Asian countries. Since the inclusion of Hib vaccine in the World Health Organization's Expanded Program on Immunization (EPI) in the late 1980s, the incidence of Hib disease has dramatically declined. However, recent international surveillance programs indicated an emergence of invasive *H. influenzae* diseases due to non-*Hib* strains during the last decade (Cox et al. 2017). Therefore, the development of a cross-strain effective

vaccine is essential. In the present study detailed analysis of 1.8 MB genome of the *H. influenzae* strain 'a' has been performed for the prediction of common putative vaccine candidate targets. The prediction has been carried out by checking different parameters such as virulence, antigenic factors, and determination of the 3D structure.

The serotype replacement or capsule switching after Hib vaccine introduction has been reported from various parts of the world. Significant increase in meningitis incidence due to type a was documented even in those who had already been vaccinated with the Hib conjugate vaccine. (Ribeiro et al., 2003). Therefore, we need new common putative vaccine candidates to induce immunity against different strains of *H. influenzae*. The present study revealed some highly conserved antigenic epitopes of virulent and essential proteins which can be capable to successfully elicit immune response against different strains thus making them ideal targets for common peptide vaccines.

Most of the studies on human immunity against *H. influenzae* infections have focused on identification of antigenic proteins that are essential for colonization, invasion, and survival within the host (Latz et al. 2014). In the present study different immunoinformatics tools have been used for prediction of common putative vaccine candidates against typeable strains of *H. influenzae*. This was followed by the prediction of the location of the highly antigenic epitopes within the cellular structure and docking of selected epitopes with common human allele DRB1*0101. Immunological studies revealed the detection of antibodies against OMPs and LPSs in human serum, thereby, indicating them to be ideal candidates for the activation of humoral immunity hence making them excellent candidates for vaccine development (Behrouzi et al. 2017). In the present study, focus was on the prioritization of the outer membrane proteins as vaccine candidates. Therefore, only those proteins which were predicted in outer membrane or extracellular spaces were further analysed for screening of putative vaccine candidates. Targeting these proteins was important because of their significant role as virulence factors because they aid the pathogen in the invasion, adherence, survival and proliferation in the infected tissue (Naz et al. 2015). The identified proteins were then proceeded for

the essentiality check. This analysis is an important parameter for identification of potential vaccine targets as they control major cellular functions of microbes that are essential for the bacterial survival and pathogenesis (Rizwan et al. 2017). The exploitation of Database of Essential Genes (DEG) yielded 41 of the 47 outer membrane and extracellular proteins as essential. Those proteins evaluated to be essential were further analysed for their role as virulence factors by using the virulent factors databases. Elucidating the molecular mechanisms of virulent factors (VFs) can improve the understanding of the cellular and molecular basis of pathogenesis. Many important virulence factors interact with host cells and modulate their functions. The identified essential and virulent proteins pooled out of extracellular and OMPs of *H. influenzae* were subjected to homology analysis against the human proteome. A high degree of molecular mimicry of administered vaccine or drug molecule can lead to autoimmunity, thereby, resulting in the failure of the prophylactic and therapeutic strategies. To avoid such situations, the potential vaccine targets should be non-homologous to human proteome.

The higher order tertiary and quaternary structures of proteins provide an effective way to visualize, understand and infer their relationship with other proteins. A combination of modelling tools such as Swiss port, I-TASSER, and Phyre were used for 3D structure modelling. Total 9 best scoring linear epitopes were selected for final analysis of T-Cell epitope predictions by MHC Pred servers (Table 2). These epitope identification analyses, hence, yielded the regions that were capable of interacting with the B and T lymphocytes. The binding energy scores after docking revealed the excellent binding affinity of all the epitopes with beta chain of the DRB1*0101 allele. The higher binding affinity values are considered effective. In the current analysis the epitopic binding scores range from -5.0 kcal/mol to -7.8 kcal/mol indicating excellent binding affinity.

To date, there is no recommended common vaccine against *H. influenzae* typeable strains. The concept of prevention of infections caused by typeable strains by designing a common, epitope-based vaccine has not been reported previously. In the developing countries the cost of conjugate vaccine

against Hib has precluded its use and even after years of introduction of Hib conjugate vaccines, it has contributed to <2% reduction in the global Hib incidence. (Ribeiro et al., 2003). In such situation the capsule switching and strain replacement leading to spread of invasive disease due to other typeable types strongly suggest identifications and development of common putative vaccine targets against typeable strains. This idea of development of common vaccines against typeable strains will greatly reduce the burden of such invasive infections against other types and can easily be included in global immunization schedule to reduce potential increases in invasive disease due to serotype replacement. Moreover, such a vaccine offers the ability to stimulate an effective specific immune response while avoiding potential undesirable effects. In this study among the predicted five putative protein candidates, Pal OMP P6 is found to be highly conserved and meet critical requirements to be used as a common vaccine candidate. This reflects a promising scope to use these proteins as common protective and therapeutic option against typeable strains of *H. influenzae*. Further investigations to establish the immunogenicity, stability, efficacy and their delivery strategy need to be carried out through *in vitro* and *in vivo* experimentation.

Conclusion

This is the first *in silico* study for the reference strain to predict common target vaccine candidates against typeable strains of *H. influenzae*. The predicted epitopes of the conserved protein targets can be used to develop effective, trustable prophylactic and therapeutic agents. Among the five predicted proteins, Omp p6 is highly conserved with stable molecular evolution and, thus, is an ideal target as putative vaccine candidate for developing a cross-strain effective vaccine against the *H. influenzae* typeable strains. The identified epitopic regions can form part of the cross-strain effective conjugate vaccines against the bacterium.

Authors' contributions

NB and NSZ conceptualized the study and experimental design, carried out the experimentation, data analysis and data interpretation. MT and MMB interpreted the data and reviewed the

experimentation. All the authors contributed to the drafting and revision of the manuscript at all the stages and approved the final manuscript.

References

- Adhikari U.K., Tayebi M. and Rahman M.M. 2018. Immunoinformatics Approach for Epitope-Based Peptide Vaccine Design and Active Site Prediction against Polyprotein of Emerging Oropouche Virus. *Journal of Immunology Research*, **2018**: 6718083 (1-22). <https://doi.org/10.1155/2018/6718083>.
- Ali S.A., Almofti Y.A. and Abd-elrahman K.A. 2019. Immunoinformatics Approach for Multiepitopes Vaccine Prediction against Glycoprotein B of Avian Infectious Laryngotracheitis Virus. *Advances in Bioinformatics*, **2019**: 1–23. <https://doi.org/10.1155/2019/1270485>.
- Andreatta M. and Nielsen M. 2016. Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics*, **32**(4):511-7. <https://doi.org/10.1093/bioinformatics/btv639>.
- Behrouzi A., Vaziri F., Rahimi-Jamnani F., Afrough P., Rahbar M., Satarian F. and Siadat S.D. 2017. Vaccine Candidates against Nontypeable *Haemophilus influenzae*: A Review. *Iranian Biomedical Journal*, **21**(2): 69–76. <https://doi.10.18869/acadpub.ibj.21.2.69>.
- Berndsen M.R., Erlendsdóttir H. and Gottfredsson M. 2012. Evolving epidemiology of invasive *Haemophilus* infections in the post-vaccination era: Results from a long-term population-based study. *Clinical Microbiology and Infection*, **18**(9): 918–923. <https://doi.org/10.1111/j.1469-0691.2011.03700.x>.
- Brown V.M., Madden S., Kelly L., Jamieson F.B., Tsang R.S.W. and Ulanova M. 2009. Invasive *Haemophilus influenzae* Disease Caused by Non-Type b Strains in Northwestern Ontario,

- Canada, 2002–2008. *Clinical Infectious Diseases*, **49**(8): 1240–1243. <https://doi.org/10.1086/605671>.
- Collins S., Litt D., Almond R., Findlow J., Linley E., Ramsay M., Borrow R. and Ladhani S. 2018. *Haemophilus influenzae* type b (Hib) seroprevalence and current epidemiology in England and Wales. *Journal of Infection*, **76**(4): 335–341. <https://doi.org/10.1016/j.jinf.2017.12.010>.
- Cox A.D., Williams D., Cairns C., Michael F., Fleming P., Vinogradov E., Arbour M., Masson L. and Zou W. 2017. Investigating the candidacy of a capsular polysaccharide-based glycoconjugate as a vaccine to combat *Haemophilus influenzae* type a disease: A solution for an unmet public health need. *Vaccine*, **35**(45): 6129–6136. <https://doi.org/10.1016/j.vaccine.2017.09.055>.
- Dassault Systèmes BIOVIA. 2016. Discovery Studio Modeling Environment, Release 2017, San Diego: Dassault Systèmes, 2016.
- Droppa-almeida D., Franceschi E. and Padilha F.F. 2018. Immune-Informatic Analysis and Design of Peptide Vaccine from Multi-epitopes Against *Corynebacterium pseudotuberculosis*. *Bioinformatics and Biology Insights*, **12**(1): 1-9 <https://doi.org/10.1177/1177932218755337>.
- Eutsey R.A., Hiller N.L., Earl J.P., Janto B.A., Dahlgren M.E., Ahmed A., Powell E., Schultz M.P., Gilsdorf J.R., Zhang L., Smith A., Murphy T.F., Sethi S., Shen K., Post J.C., Hu F.Z. and Ehrlich G.D. 2013. Design and validation of a supragenome array for determination of the genomic content of *Haemophilus influenzae* isolates. *BMC Genomics* 2013, **14**(1): .2013 Jul 17,14:484 (1-16). <https://doi.org/10.1186/1471-2164-14-484>.
- Iskander M., Hayden K., van Domselaar G. and Tsang R. 2017. First Complete Genome Sequence of *Haemophilus influenzae* Serotype a. *Genome Announcements*, **5**(3): e01506-16. <https://doi.org/10.1128/genomeA.01506-16>.

- Janto B.A., Hiller N.L., Eutsey R.A., Dahlgren M.E., Earl J.P., Powell E., Ahmed A., Hu F.Z. and Ehrlich G.D. 2014. Development and validation of an *Haemophilus influenzae* supragenome hybridization (SGH) array for transcriptomic analyses. PLoS One, **9**(10): e105493. <https://doi.org/10.1371/journal.pone.0105493>.
- Källberg M., Wang H., Wang S., Peng J., Wang Z., Lu H. and Xu J. 2016. Template-based protein structure modeling using the RaptorX web server. Nature Protocols, **7**(8): 1511–1522. <https://doi.org/10.1038/nprot.2012.085>.
- Kelly L., Tsang R.S.W., Morgan A., Jamieson F.B. and Ulanova M. 2018. Invasive disease caused by *Haemophilus influenzae* type a in Northern Ontario First Nations communities. Journal of Medical Microbiology, **60** (2011): 384–390. <https://doi.org/10.1099/jmm.0.026914-0>.
- Konini A., Nix E., Ulanova M. and Moghadas S.M. 2016. Dynamics of naturally acquired antibody against *Haemophilus influenzae* type a capsular polysaccharide in a Canadian Aboriginal population. Preventive Medicine Reports, **3**: 145–150. <https://doi.org/10.1016/j.pmedr.2016.01.004>.
- Krishnan A., Amarchand R., Gupta V., Lafond K.E., Suliankatchi R.A., Saha S., Rai S., Misra P., Purakayastha D.R., Wahi A., Sreenivas V., Kapil A., Dawood F., Pandav C.S., Broor S., Kapoor S.K., Lal R. and Widdowson M. 2015. Epidemiology of acute respiratory infections in children - preliminary results of a cohort in a rural north Indian community. BMC Infectious Diseases, **15**: 462 (1-10). <https://doi.org/10.1186/s12879-015-1188-1>.
- Kumavath R., Prasad S., Devarapalli P. and Barh D. 2014. In Silico Identification of Novel Candidate Drug Targets in *Haemophilus Influenzae* Rd KW20. International Journal of Genetics and Genomics, **2** (4): 62-67. <https://doi.org/10.11648/j.ijgg.20140204.13>.

- Lamiable A., Thevenet P., Rey J., Vavrusa M., Derreumaux P. and Tuffery P. 2016. PEP-FOLD3: faster denovo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Research*, **44**: 449–454. <https://doi.org/10.1093/nar/gkw329>.
- Latz E., Franko J., Golenbock D.T. and Schreiber J.R. 2014. Haemophilus influenzae Type b-Outer Membrane Protein Complex Glycoconjugate Vaccine Induces Cytokine Production by Engaging Human Toll-Like Receptor 2 (TLR2) and Requires the Presence of TLR2 for Optimal Immunogenicity. *Journal of Immunology*, **172**(4):2431-8. <https://doi.org/10.4049/jimmunol.172.4.2431>.
- Luo H., Lin Y., Gao F., Zhang C.T. and Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research*, **42**(D1): D574-80. <https://doi.org/10.1093/nar/gkt1131>.
- Mohd-Zain Z., Kamsani N.H., Ahmad N. and Clarke S.C. 2015. Phylogenetic relationship of non-typeable Haemophilus influenzae isolated in Malaysia. *Infection, Genetics and Evolution*, **36**: 240–243. <https://doi.org/10.1016/j.meegid.2015.09.017>.
- Murphy T.F. 2015. Vaccines for Nontypeable Haemophilus influenzae: The Future Is Now. *Clinical and Vaccine Immunology*, **22**(5): 459–466. <https://doi.org/10.1128/CVI.00089-15>.
- Naz A., Awan F.M., Obaid A., Muhammad S.A., Paracha R.Z., Ahmad J. and Ali A. 2015. Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: A reverse vaccinology-based approach. *Infection, Genetics and Evolution*, **32**:280-91. <https://doi.org/10.1016/j.meegid.2015.03.027>.
- Paul S., Sidney J., Sette A. and Peters B. 2016. TepiTool: A pipeline for computational prediction of T cell epitope candidates. *Current Protocols in Immunology*, **114**: 18.19.1–18.19.24. <https://doi.org/10.1002/cpim.12>.

- Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C. and Ferrin T.E. 2004. UCSF Chimera-a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13):1605-12. <https://doi.org/10.1002/jcc.20084>.
- Pinto M., González-Díaz A., Machado M.P., Duarte S., Vieira L., Carriço J.A., Marti S., Bajanca-Lavado M.P. and Gomes J.P. 2019. Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infection, Genetics and Evolution*, **67** (10/2018): 126–135. <https://doi.org/10.1016/j.meegid.2018.10.025>.
- Rashid M.I., Naz A., Ali A. and Andleeb S. 2017. Prediction of vaccine candidates against *Pseudomonas aeruginosa*: An integrated genomics and proteomics approach. *Genomics*, **109**(3-4):274-283. <https://doi.org/10.1016/j.ygeno.2017.05.001>.
- Ribeiro G.S., Reis J.N., Cordeiro S.M., Lima J.B.T., Gouveia E.L., Petersen M., Salgado K., Silva H.R., Zanella R.C., Almeida S.C.G., Brandileone M.C., Reis M.G. and Ko A.I. 2003. Prevention of *Haemophilus influenzae* Type b (Hib) Meningitis and Emergence of Serotype Replacement with Type a Strains after Introduction of Hib Immunization in Brazil. *Journal of Infectious Diseases*, **187**(1): 109–116. <https://doi.org/10.1086/345863>.
- Rizwan M., Naz A., Ahmad J., Naz K., Obaid A., Parveen T., Ahsan M. and Ali A. 2017. VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. *BMC Bioinformatics*, **18**:106 (1–7). <https://doi.org/10.1186/s12859-017-1540-0>.
- Soria-Guerra R.E., Nieto-Gomez R., Govea-Alonso D.O. and Rosales-Mendoza S. 2015. An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. *Journal of Biomedical Informatics*, **53**:405-14. <https://doi.org/10.1016/j.jbi.2014.11.003>.
- Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J. and von Mering C. 2019. STRING v11: Protein-protein

- association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47(D1)**: D607-D613. <https://doi.org/10.1093/nar/gky1131>.
- Tsang R.S.W., Proulx J.F., Hayden K., Shuel M., Lefebvre B., Boisvert A. and Moore D. 2017. Characteristics of invasive *Haemophilus influenzae* serotype a (Hia) from Nunavik, Canada and comparison with Hia strains in other North American Arctic regions. *International Journal of Infectious Diseases*, **57**: 104–107. <https://doi.org/10.1016/j.ijid.2017.02.003>.
- Tsang R.S.W., Shuel M., Whyte K., Hoang L., Tyrrell G., Horsman G., Wylie J., Jamieson F., Lefebvre B., Haldane D., Gad R.R., German G.J. and Needle R. 2017. Antibiotic susceptibility and molecular analysis of invasive *Haemophilus influenzae* in Canada, 2007 to 2014. *Journal of Antimicrobial Chemotherapy*: **72(5)**: 1314–1319. <https://doi.org/10.1093/jac/dkw565>.
- Ulanova M., Tsang R. and Altman E. 2012. Neglected infectious diseases in Aboriginal communities: *Haemophilus influenzae* serotype A and *Helicobacter pylori*. *Vaccine*, **30(49)**: 6960–6966. <https://doi.org/10.1016/j.vaccine.2012.09.063>.
- Ulanova M. and Tsang R.S.W. 2014. *Haemophilus influenzae* serotype a as a cause of serious invasive infections. *Lancet Infectious Diseases*, **14(1)**: 70–82. [https://doi.org/10.1016/S1473-3099\(13\)70170-1](https://doi.org/10.1016/S1473-3099(13)70170-1).
- van Eldere J., Slack M.P.E., Ladhani S. and Cripps A.W. 2014. Non-typeable *Haemophilus influenzae*, an under-recognised pathogen. *Lancet Infectious Diseases*, **14(12)**: 1281–1292. [https://doi.org/10.1016/S1473-3099\(14\)70734-0](https://doi.org/10.1016/S1473-3099(14)70734-0).
- Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., Heer F.T., de Beer T.A.P., Rempfer C., Bordoli L., Lepore R. and Schwede T. 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, **46(W1)**: W296-W303. <https://doi.org/10.1093/nar/gky427>.

- Wattam A.R., Davis J.J., Assaf R., Boisvert S., Brettin T., Bun C., Conrad N., Dietrich E.M., Disz T., Gabbard J.L., Gerdes S., Henry C.S., Kenyon R.W., Machi D., Mao C., Nordberg E.K., Olsen G.J., Murphy-Olson D.E., Olson R., Overbeek R., Parrello B., Pusch G.D., Shukla M., Vonstein V., Warren A., Xia F., Yoo H. and Stevens R.L. 2017. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research*, **45**(D1): D535–D542. <https://doi.org/10.1093/nar/gkw1017>.
- Xiao Z., Zhang Y., Yu R., Chen Y., Jiang X., Wang Z. and Li S. 2018. In silico design of MHC class I high binding affinity peptides through motifs activation map. *BMC Bioinformatics*, **19** (Suppl 19): 516 (13-24). <https://doi.org/10.1186/s12859-018-2517-3>.
- Yang J., Yan R., Roy A., Xu D., Poisson J. and Zhang Y. 2015. The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, **12**(1): 7–8. <https://doi.org/10.1038/nmeth.3213>.
- Yu C.S., Cheng C.W., Su W.C., Chang K.C., Huang S.W., Hwang J.K. and Lu C.H. 2014. CELLO: A web server for protein subcellular localization prediction with functional gene ontology annotation. *PLoS One*, **9**(6): e99368 (1-9). <https://doi.org/10.1371/journal.pone.0099368>.
- Yu N.Y., Wagner J.R., Laird M.R., Melli G., Rey S., Lo R., Dao P., Cenk Sahinalp S., Ester M., Foster L.J. and Brinkman F.S.L. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**(13): 1608–1615. <https://doi.org/10.1093/bioinformatics/btq249>.
- Zaidi A.K., Khan H., Sherali A.R., Lasi R. and Group S.M.S. 2010. Burden of Haemophilus influenzae type b disease in Pakistani children. *Eastern Mediterranean Health Journal*, **16**(6): 460–464.

Table 1. Summary of Physical and Chemical Analysis Results

Parameters	Pal omp6	Gdh	Tbp1	PilW typeIV	Porin OmpA
Gene Size (bp)	462	1350	2739	540	1062
Protein Length (aa)	153	449	912	179	353
M. W (kDa)	16	48	103	20	37
pI Value	6.09	6.20	9.34	8.36	9.62
VaxiJen score	0.1642	0.5364	0.5989	0.6058	0.6150
Allergen non-allergen	Allergen	Non-Allergen	Non-Allergen	Non-Allergen	Non-Allergen
Helices <2	zero	zero	zero	zero	zero
Estimated half-life	30hrs	30hrs	30hrs	30hrs	30hrs
Instability index	24.62	23.55	38.14	46.99	20.48
Aliphatic index	79.87	82.38	71.24	82.40	82.75
Non Homologous	yes	yes	yes	yes	yes

Draft

Table 2. Nine selected epitopes based on the best score values from the prediction tool.

Protein	Epitope	B-Cell Epitopes	T cell epitope	Binding energy score kcal/mol	VaxiJen score	Ic50< 100nm
Pal OMP P6	P E1	AACSSSNNDAAAGNGAAQTFG	CSSSNNDAA	-5.4	1.4973	31
Gdh	G E1	ADVLFPGPKAANAGGVATSG	VLFGPGKAA	-6.2	0.9723	38
	G E2	TTLPMGGAKGGSDFDPKGKS	TTLPMGGAK	-5.1	1.0268	38
	G E3	HVGADTDVPAGDIGVGGREV	DTDVPAGDI	-6.4	1.3879	39
	T E1	CTNGYESCKKSPKPPAKLST	CTNGYESCK	-7.2	0.9685	40
Tbp1	T E2	SNPYLYPKPEPYFAGGDLCN	YLYPKPEPY	-7.8	0.4609	40
	T E3	LSKNGTGKGNYGYNHNAQNAK	GTGKGNYG	-6.8	1.3176	35
PilW	PW E1	EQAQQQFELALNSPNYYHQA	AQQQFELAL	-5.4	0.7423	41
Porin	PorinE1	DAISATGYGKANPVTGATCD	KANVTGAT	-5.8	0.9864	41

Figure Legends

Figure 1: Immunoinformatics workflow for the prediction of common putative vaccine candidates for *Haemophilus influenzae*.

Figure 2: The pie-chart represents percentage and number of predicted proteins in five different localizations. (a) Results generated by CELLO server, 735 cytoplasmic, 346 inner membrane, 42 outer membrane, 45 periplasmic and 5 extracellular, whereas location for 451 proteins couldn't predicted. (b) PSORTb generated localization results: 726 cytoplasmic, 330 inner membrane, 40 outer membrane, 38 periplasmic and 3 extracellular, whereas location for 487 proteins couldn't predicted.

Figure 3: 3D structure of the five prioritized proteins predicted by I-TASSER (a-e). The predicted structures show whole sequence topology which is requirement for epitope mapping.

Figure 4: Protein-Protein Interaction analysis of the five screened proteins: (a) Gdh interaction network, (b) Pal OMP P6 interaction network, (c) porin ompA interaction analysis. (d) PilW type IV interaction network, due to lack of record in STRING Databases a hypothetical map has been

constructed based on the closely related proteins of the same family and domain, and (e) Tbp1 Interaction network, Blue lines indicate gene co-occurrence while green lines represent gene neighbourhood.

Figure 5: Topological view of the five prioritized proteins with their corresponding antigenic epitopes. Protein structures and their epitopic regions have been illustrated in pictorial form in different colours with 1 letter code + specifier on their respective 3D structure of proteins. (a) Pal Omp P6 Epitope 1 (PE1) has been highlighted in blue colour with 1 letter code sequence and specifier on 3D structure of Pal Omp P6 protien. (b) PW E1 Epitope of Pilw type IV protein is highlighted on surface of its 3D structure in yellow colour with 1 letter code sequence and specifier. (c) Porin E1 Epitope of porin protein is highlighted on surface of its 3D structure in blue colour with 1 letter code sequence and specifier. (d) TE1 Epitope of Tbp1 protein is highlighted in red colour, TE2 in purple colour and TE3 in blue colour with 1 letter code sequence and specifier on surface of its 3D structure(e) Gdh E1 Epitope is highlighted in red colour, Gdh E2 in purple colour and Gdh E3 in blue colour with 1 letter code sequence and specifier on surface of its 3D structure.

Figure 6: Spatial representation of the dock molecule (ligand and DRB1*0101 allele) (a) Active pockets residues on beta chain of receptor molecule DRB1*0101 allele, Showing the ligand binding site (b) Docking of ligand (TE3) and receptor molecule (beta chain of DRB1*0101 allele).

Table Legends

Table 1 Summary of Physical and Chemical Analysis of selected proteins.

Table 2 Nine selected epitopes based on the best score values from the prediction tool.

Supplementary Figures

Supplementary Figure 1: The representation of 5 prioritized target proteins: (a) the proteins characterized as outer membrane, essential, virulent, non-human homologues, antigenic and having appropriate molecular weight (<110 kDa) are shown using different colours in the Venn diagram. (b)

Those proteins satisfying a particular parameter are shown in each category of the Venn diagram. Thus, in total five proteins were prioritized. The image has been generated by Jvenn (URL: <http://bioinfo.genotoul.fr/jvenn/usermanual.html>) (Workbench et al., 2010).

Supplementary Figure 2. Genome of *H. influenzae* serotype “a” NML Hia 1. The outer most ring (grey colour) and second ring represents ORFs plus strands, with information of COG (1st ring from outside) and subcellular localizations (2nd ring). (green: cytoplasmic, yellow: cytoplasmic membrane, blue: outer membrane, sky blue: extracellular, pink periplasmic and grey: unknown) third and fourth rings represents the ORFs minus strands with information of COG (3rd ring) and subcellular localizations (4th ring). The 5th ring (inner most) shows GC percentage shown by the peak length with increase in percentage.

Supplementary Figure 3: Relative positions of screened proteins on *H. influenzae* genome: which were analysed for epitope mapping. The respective positions of selected potential vaccine candidates (represented by purple dots). The integers on the outer circle represent the corresponding genome positions. The Figure was generated using Island Viewer 3 (<http://www.pathogenomics.sfu.ca/islandviewer/>)

Supplementary Figure 4: Multiple Sequence Alignment (MSA) of Gdh protein of the retrieved strains using BioEdit software and ClustalW. Colors indicated the conservancy and letters in highlighted bars showed the alteration in amino acid

Supplementary Figure 5: Evolutionary divergence analysis of Glutamate dehydrogenase protein of different strains of *H. influenzae*. The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. This analysis involved 50 amino acid sequences. All positions with less than 95% site coverage were eliminated,

i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There were a total of 449 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4]

Supplementary Figure 6: Multiple Sequence Alignment (MSA) of Pilus type IV Protein of the retrieved strains using BioEdit software and ClustalW. Colors indicated the conservancy and letters in highlighted bars showed the alteration in amino acid

Supplementary Figure 7: Evolutionary divergence analysis of Pilus type IV protein of different strains of *H. influenzae*. The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. This analysis involved 30 amino acid sequences. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There were a total of 179 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4].

Supplementary Figure 8: Multiple Sequence Alignment (MSA) of PORIN Protein of the retrieved strains using BioEdit software and ClustalW. Colors indicated the conservancy and letters in highlighted bars showed the alteration in amino acid.

Supplementary Figure 9: Evolutionary divergence analysis of porin protein of different strains of *H. influenzae*. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [1]. The tree with the highest log likelihood (-3636.87) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting

the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 30 amino acid sequences. There were a total of 358 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2].

Supplementary Figure 10: Multiple Sequence Alignment (MSA) of TBP Protein of the retrieved strains using BioEdit software and ClustalW. Colors indicated the conservancy and letters in highlighted bars showed the alteration in amino acid.

Supplementary Figure 11: Evolutionary divergence analysis of *tbp1* protein of different strains of *H. influenzae*. The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. This analysis involved 30 amino acid sequences. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There were a total of 910 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4].

Supplementary Figure 12: Multiple Sequence Alignment (MSA) of Pal OMP P6 Protein of the retrieved strains using BioEdit software and ClustalW. Colors indicated the conservancy and letters in highlighted bars showed the alteration in amino acid.

Supplementary Figure 13: Evolutionary divergence analysis of Pal OMP P6 protein of different strains of *H. influenzae*. The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to

scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. This analysis involved 51 amino acid sequences. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There were a total of 153 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4]

Supplementary Tables

Supplementary Table 1: B cell Epitopes Prediction Results by Ellipro Server (a) Gdh (b) Porin Omp (c) Tbp 1 (d) Pil W (e) Pal

Supplementary Table 2: B cell Epitopes Prediction Results by using BCPreds server (a) Gdh (b) Porin Omp (c) Tbp 1 (d) Pil W (e) Pal

Supplementary Table 3: Prioritized five proteins after analysis of predicted 40 outer membrane for essentiality and virulence check parameters.

Supplementary Table 4: B-Cell predicted Epitope number by two different tool Ellipro for linear and discontinuous and BCPred for overlapping epitope prediction

Supplementary Table 5: Accession number of Gdh Protein sequences of 50 different strains of *H. influenzae* with strain name and year

Supplementary Table 6: Accession number of Pilus Type IV Protein sequences of 30 different strains of *H. influenzae* with strain name and year.

Supplementary Table 7: Accession number of tbp1 protein of 30 different strains of *H. influenzae* with strain name and year.

Supplementary Table 8: Accession number of porin protein sequences of 30 different strains of *H. influenzae* with strain name and year

Supplementary Table 9: Accession number Pal OMP P6 protein sequences of 50 different strains of *H. influenzae* with strain name and year.

Supplementary Table 10: List of essential outer membrane proteins.

Draft

Figure 1: Immunoinformatics workflow for the prediction of common putative vaccine candidates for *Haemophilus influenzae*.

81x45mm (300 x 300 DPI)

Figure 2: The pie-chart represents percentage and number of predicted proteins in five different localizations. (a) Results generated by CELLO server, 735 cytoplasmic, 346 inner membrane, 42 outer membrane, 45 periplasmic and 5 extracellular, whereas location for 451 proteins couldn't predicted. (b) PSORTb generated localization results: 726 cytoplasmic, 330 inner membrane, 40 outer membrane, 38 periplasmic and 3 extracellular, whereas location for 487 proteins couldn't predicted.

96x28mm (300 x 300 DPI)

Figure 3: 3D structure of the five prioritized proteins predicted by I-TASSER (a-e). The predicted structures show whole sequence topology which is requirement for epitope mapping.

135x83mm (300 x 300 DPI)

Figure 4: Protein-Protein Interaction analysis of the five screened proteins: (a) Gdh interaction network, (b) Pal OMP P6 interaction network, (c) porin ompA interaction analysis. (d) PilW type IV interaction network, due to lack of record in STRING Databases a hypothetical map has been constructed based on the closely related proteins of the same family and domain, and (e) Tbp1 Interaction network, Blue lines indicate gene co-occurrence while green lines represent gene neighbourhood.

96x72mm (300 x 300 DPI)

Figure 5: Topological view of the five prioritized proteins with their corresponding antigenic epitopes. Protein structures and their epitopic regions have been illustrated in pictorial form in different colours with 1 letter code + specifier on their respective 3D structure of proteins. (a) Pal Omp P6 Epitope 1 (PE1) has been highlighted in blue colour with 1 letter code sequence and specifier on 3D structure of Pal Omp P6 protien. (b) PW E1 Epitope of Pilw type IV protein is highlighted on surface of its 3D structure in yellow colour with 1 letter code sequence and specifier. (c) Porin E1 Epitope of porin protein is highlighted on surface of its 3D structure in blue colour with 1 letter code sequence and specifier. (d) TE1 Epitope of Tbp1 protein is highlighted in red colour, TE2 in purple colour and TE3 in blue colour with 1 letter code sequence and specifier on surface of its 3D structure(e) Gdh E1 Epitope is highlighted in red colour, Gdh E2 in purple colour and Gdh E3 in blue colour with 1 letter code sequence and specifier on surface of its 3D structure.

106x98mm (300 x 300 DPI)

Figure 6: Spatial representation of the dock molecule (ligand and DRB1*0101 allele) (a) Active pockets residues on beta chain of receptor molecule DRB1*0101 allele, Showing the ligand binding site (b) Docking of ligand (TE3) and receptor molecule (beta chain of DRB1*0101 allele).

127x45mm (300 x 300 DPI)