Database on bacterial vaccines and prediction of bacterial protective antigens

Submitted by

Satakshi Gupta (MT20215)

Under the guidance of

Prof. Gajendra P.S. Raghava

Head & Professor

in partial fulfillment of the requirements for the degree of Master of Technology in Computational Biology

to

Department of Computational Biology,
Indraprastha Institute of Information Technology,
New Delhi

June 2022

Certificate

This is to certify that the thesis titled "**Database on bacterial vaccines and prediction of bacterial protective antigens**" being submitted by **Satakshi Gupta** to the Indraprastha Institute of Information Technology, Delhi for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

June, 2022

Prof. Gajendra P.S. Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology
New Delhi 110020

Acknowledgement

I would like to express my sincere gratitude and respect towards Prof. Gajendra P.S. Raghava from Indraprastha Institute of Information Technology, Delhi for being my supervisor and for exposing me to this wonderful topic of research and guiding me throughout. Besides my supervisor, I would like to thank the Ph.D. scholar, Neelam Sharma and Research Associate Dr. Naorem Leimarembi Devi from his esteemed lab for their constant guidance, support, and motivation throughout the project. I would also like to thank the Department of Computational Biology and IT administrators at IIIT Delhi for providing me with all the necessary resources. Lastly, I would also like to thank my family and friends for providing much-needed support and motivating me from time to time throughout the course of my thesis which enabled me to pursue my research in an efficient and structured manner.

Satakshi Gupta

M.Tech (CB)

Table of Contents

List of Abbreviations

AMR: Antimicrobial Resistance

WHO: World Health Organization

CDC: Center for Disease Control and Prevention

FDA: Food and Drug Administration

emc: Electronic Medicines Compendium

BCG: Bacille Calmette-Guerin

Hib: Haemophilus influenzae type b

TB: Pulmonary Tuberculosis

VIOLIN: Vaccine Investigation and Online Information Network

ML: Machine Learning

DT: Decision Tree

RF: Random Forest

KNN: k-Nearest Neighbors

SVC: Support Vector Classification

GNB: Gaussian Naïve Bayes

XGB: XGBoost

LR: Logistic Regression

Sen: Sensitivity

Spc: Specificity

Acc: Accuracy

MCC: Matthews Correlation Coefficient

AUROC or AUC: Area Under Receiver Operating Characteristics

BPAgs: Bacterial Protective Antigens

AAC: Amino Acid Composition

DPC: Dipeptide Composition

TPC: Tripeptide Composition

PAAC: Pseudo Amino Acid Composition

APAAC: Amphiphilic Pseudo Amino Acid Composition

DDR: Distance Distribution of Repeats

QSO: Quasi-Sequence Order

RRI: Residue Repeats Index

ATC: Atom Type Composition

BTC: Bond Type Composition

SEP: Shannon Entropy of a Protein

SER: Shannon Entropy of a Residue

SPC: Shannon Entropy of Physicochemical Property

PCP: Physico-Chemical Properties

List of Figures

List of Tables

Abstract

Bacterial diseases are the reason for millions of deaths worldwide thus preventing them is very important. Vaccines are the most cost-effective prevention against many infectious diseases. There are different kinds of bacterial vaccines like live attenuated, inactivated, subunit, toxoids and conjugate with each type having its own way of providing immunity for preventing human bacterial diseases such as pulmonary tuberculosis, diphtheria and many others. In literature, a compiled resource, providing relevant information about vaccines against bacterial diseases was not available. In this study, we have developed a manually curated exhaustive database of bacterial diseases vaccines, to aid researchers in developing novel vaccine candidates. We have created BacVacDB, which is a web-based freely accessible database of bacterial vaccines maintaining comprehensive information related to vaccines. This database comprises 371 vaccine entries covering 30 human bacterial diseases manually extracted from research articles, websites and public databases. Each entry provides detailed information about vaccine name, type, age, description, manufacturer, manufacturing country, year of manufacture, clinical phase status, etc. We have covered details of both approved vaccines as well as the vaccines undergoing different human clinical trials. It provides all the bacterial vaccines information on a single platform to perform efficient and time saving search. The database is accessible at https://webs.iiitd.edu.in/raghava/bacvacdb/. Protective antigens are very important in the research for the development of new and improved vaccines against infectious and non-infectious diseases. Protective antigens are those antigens that are specifically targeted by the acquired immune response of the host. They are capable of stimulating the production of antibodies and induce cell mediated immunity. Identification of protective antigens is the most critical step in the vaccine development process as once the protective antigens are identified, researchers can use these antigens to develop effective subunit and DNA vaccines. In this study, we have developed prediction method to predict if a particular protein could be used as a bacterial protective antigen or not in vaccine development process using different machine learning techniques.

Objectives

1. Collection and compilation of bacterial vaccines for human diseases from the literature.
2. Development of web-server (BacVacDB) to organize human bacterial vaccines information in user friendly format.
3. Implemented different machine learning classification techniques to predict the bacterial protective antigens for the vaccine development process against bacterial diseases.

Chapter 1

Introduction

Introduction to bacterial diseases

Disease is a harmful deviation from the normal state of an organism [1]. There could be multiple causes of disease like pathogenic microbial infections (typhoid, tuberculosis), immune system internal dysfunction (cancer, autoimmune diseases), genetically transferred diseases (down syndrome, thalassemia) [2]. A pathogenic organism can be defined as the harmful microorganism causing disease in the host. These organisms include bacteria, viruses, fungi, parasites and protozoans [3]. They are capable of causing different kinds of infectious diseases and can infect both humans or animals via direct or indirect ways like contaminated food (salmonellosis, botulism), contaminated water (cholera, typhoid fever), airborne (whooping cough, mumps), exposure to infected person (tuberculosis, influenza), insects or animal vectors (plague, malaria) [4][5].

Bacteria are ubiquitous single celled organisms, highly adaptable, the simplest forms of life which are important in maintaining the environment [6]. They lack a nuclear membrane and divide by binary fission, holding their genetic information in a circular double stranded DNA [7]. Most of the bacteria can be classified as gram positive or gram negative depending on their cell wall by a procedure called gram staining [8]. Bacteria could also be categorized based on their requirement of oxygen, for instance, aerobic bacteria (e.g. *Bordetella pertussis*) require oxygen for their growth while anaerobic bacteria (e.g. *Clostridia*) do not need oxygen for their growth and facultative organisms can grow both in the presence or absence of oxygen [5]. Some of the bacterial strains are considered as good since they are not harmful and beneficial for our body, for example, the bacteria present in the gut (*Escherichia coli*, *Lactobacillus acidophilus* and *Bifidobacterium bifidum*) helps in digestion [9] while there are few pathogenic bacterial species that can cause different infectious diseases. Bacteria can cause disease in different ways, some make their entry in the skin via a cut, while for others sexual activity could be the reason. *Staphylococcus* and *Streptococcus* bacteria are mainly associated with skin infections [10]. *Chlamydia trachomatis* and *Neisseria gonorrhoeae* causes sexually transmitted diseases, chlamydia and gonorrhea which can infect both men and women [11]. Food borne diseases can be caused by bacterial species such as *Clostridium botulinum, Bacillus subtilis and Bacillus cereus* [12]. Most of these infections are treated using antibiotics [13].

Role of antibiotics and vaccines

Antibiotics are the medications which are prescribed by doctors for treating most of the bacterial infections as they help in suppressing the growth of the bacteria [14]. However, overuse of antibiotics could lead to antibiotic or antimicrobial resistance (AMR) which is one of the most serious public health threat to humans according to World Health Organization (WHO) [15]. According to the US Center for Disease Control and Prevention (CDC), more than 2.8 million antibiotic resistance infection cases affect people each year in the US [16]. As a result, many bacterial infections such as tuberculosis, gonorrhea, pneumonia and food borne infections are becoming difficult to treat [17].

Bacterial vaccines help in reducing the burden of infections by providing solutions against AMR. They are known for strengthening the immune system and boosting the health of the individuals [18]. Vaccination induces active immunity and activates the immune system's memory and makes antibodies against the pathogen. Vaccines are effective in preventing infections in all age groups [19]. There are vaccines available for preventing many human bacterial diseases such as pneumococcal, meningococcal, typhoid, cholera, tetanus etc. Some of the vaccines provide lifelong protection while for others a booster dose is required after a few years [20]. There are different types of vaccines which are prepared using different processes like they could contain whole bacteria (killed or live attenuated), toxoids, capsular polysaccharides or the proteins isolated from it [21].

Bacille Calmette-Guerin (BCG) vaccine is a live attenuated vaccine for preventing pulmonary tuberculosis where the whole cell bacteria have been weakened. It induces trained innate immunity enhancing the ability of innate effector cells to respond to non-specific stimuli [19]. Vivotif [22] and Typherix [23] are also the examples of live attenuated vaccines which are used for typhoid [24]. Another types of vaccines are the toxoids which are chemically inactivated toxins, for instance, the vaccines for diphtheria (Tdvax)[25] and tetanus (Tripedia) [26] are toxoids. In addition to this, inactivated vaccines contain the killed bacterial cells, inactivated by the use of chemicals, heat or radiation, for example, the vaccine for Q-fever (Q-Vax) [27] falls in this category. While conjugate vaccines contain the capsular polysaccharides covalently bound to a carrier protein such as the vaccines for *Haemophilus influenzae* type b and *Neisseria meningitidis* [28]. Subunit vaccines contain one or more specific antigens to provoke a response from the immune system. A subunit vaccine for preventing Lyme disease (LYMERix, GSK) was licensed in the US in 1998 and was withdrawn in 2002 due to poor market performance [29]. There are different routes for administering these vaccines like intramuscular, intradermal, subcutaneous and oral [30].

History of vaccines

The first laboratory vaccine was developed in 1879 by Louis Pasteur for chicken cholera [31] and this experiment facilitated the development of live attenuated cholera vaccine in 1897 and inactivated anthrax vaccine in 1904 for humans [32]. The live attenuated vaccine for tuberculosis, BCG was for the first time tested in humans in the year 1921 [33] and in 1927 it was used for the first time in new-borns [34]. The live attenuated plague vaccine was reported in 1920 [35]. In 1948, the whole cell pertussis vaccine was first approved in the US for use [36]. The first monovalent (group C) meningococcal polysaccharide vaccine was licensed in 1974 [37] while the pneumococcal polysaccharide vaccine was first licensed for use in the US in 1977 and a 23-valent polysaccharide vaccine (Pneumovax 23, PPSV23) was licensed in 1983 [38]. The first Hib polysaccharide vaccine containing type b purified polysaccharide capsule polyribosylribitol phosphate (PRP) was licensed for use in 1985 followed by conjugate Hib vaccine ProHIBiT in 1987 and PedvaxHIB in 1989 [39]. In 2005, the first conjugate meningococcal vaccine, MCV4 (Menactra), was licensed in the United States and a second, MenACWY-CRM (Menveo) licensed in 2010 [40]. The overall vaccines timeline with different years is described in Figure 1.

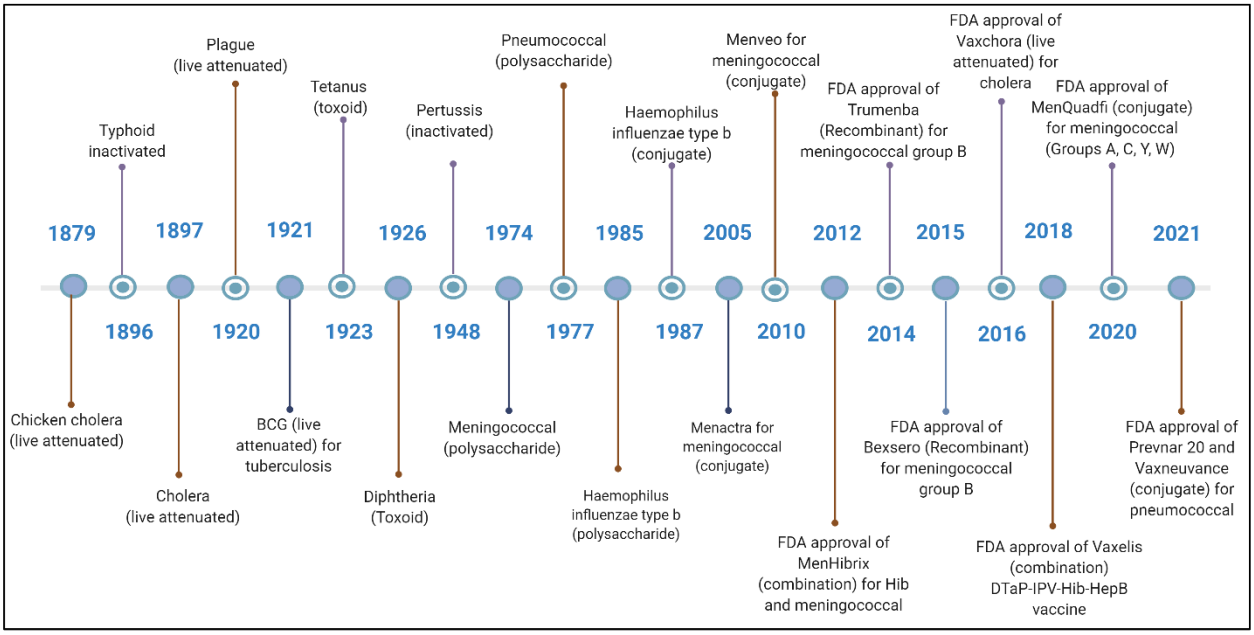

Figure 1: Timeline of bacterial vaccines production and approval

Mechanism of vaccines

Vaccines work by imitating an infection in the body, initially triggering an innate immune response, thus generating an antigen specific adaptive immunity. The innate immunity is the first line of defense against any pathogen which includes barriers to prevent harmful substances from entering the body. It is quick and non-specific [41]. While adaptive immunity is the second line of defense, which is also known as acquired or specific immunity. They provide long lasting protection by developing immunologic memory [42] It is marked by clonal expansion of B and T lymphocytes. When B cells encounter any novel antigen, they produce antibodies specific to the antigen in order to destroy it, and also produces memory cells that lasts for decades and can detect the pathogen during subsequent exposure [43].

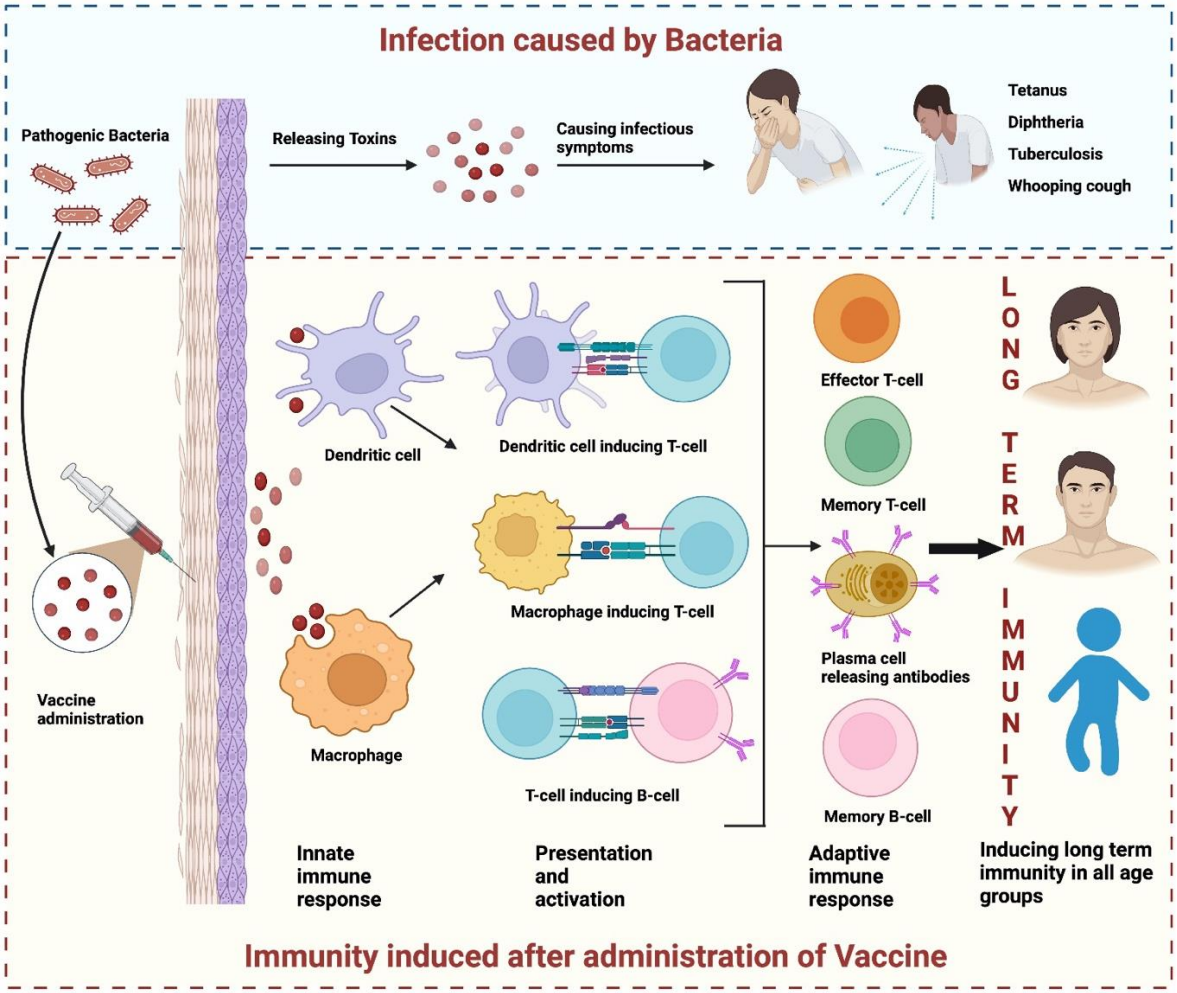

Figure 2: The basic mechanism of action of vaccines.

Introduction to protective antigens

Protective antigens are those antigens that are able to invoke specific and enhanced adaptive immune response of the host [44] and induce protection against infectious and non-infectious diseases since they stimulate the production of antibodies and cell mediated immunity when introduced inside the host. They play an important role in vaccine development, drug design. They could be used as biological markers for the diagnosis of different diseases, used in the analysis of host immunity and to control allergies. To develop new and improved vaccines, identification of protective antigens is an important step in the research and then these protective antigens are further used by researchers to develop subunit and DNA vaccines for preventing infectious diseases. [45].

Objective of thesis

All of the information related to bacterial vaccines, like its name, approval status, route of administration has been dispersed throughout the literature, which makes it time consuming process to access any information for the scientific community working in the field of vaccines and immune-informatics. As per our knowledge, there is no dedicated single platform that compiles all this detailed information. Therefore, the objective of the thesis was to have an in-depth analysis of bacterial vaccines for different diseases, their types, status, mechanism, approval, efficacy, administration route, year and many other important details. After compilation of data, all this information is made available in the form of a freely available database "BacVacDB". We hope that this database could be beneficial for the researchers of the scientific community for the development of novel vaccine candidates.

Since the identification of the protective antigens is most critical step to develop the improved vaccines against diseases, in this study we have developed different machine learning techniques to classify if a particular protein could be used as bacterial protective antigen or not for developing bacterial subunit and DNA vaccines against different bacterial infectious and non-infectious diseases. We hope that this could be useful for the researchers working in the field of vaccine development.

Chapter 2

Collection and compilation of bacterial vaccines

Introduction

Vaccines are one of the most cost-effective devices which are known to prevent many deadly infectious diseases [46]. With the invention of vaccines, our lives have been easier and we are able to defeat many harmful diseases caused by different kinds of microorganisms. Here, we are only focusing on the bacterial vaccines (e.g., BCG, Vaxelis) [47] used for preventing human bacterial diseases such as pulmonary tuberculosis, diphtheria and many others. As we already know, there are different types of bacterial vaccines available in the market, such as inactivated, toxoids, live attenuated, subunit, conjugate etc. The manufacturing process of these vaccines include various bacterial cell constituents. Since all of this information related to bacterial vaccines was scattered in the literature and it is time consuming to gather this information from different platforms, we have created BacVacDB (https://webs.iiitd.edu.in/raghava/bacvacdb), which is a web-based freely accessible database of manually curated bacterial vaccines maintaining comprehensive information. This database comprises 371 vaccine entries covering 30 human bacterial diseases manually extracted from research articles, websites and existing public databases. Each entry provides detailed information about vaccine name, type, age, description, manufacturer, manufacturing country, year of manufacture, clinical phase status, export, approval, dosage, administration site etc. We have covered 167 entries for approved vaccines and 204 entries for vaccines undergoing different human clinical trials. It provides all the bacterial vaccines information on a single platform to perform efficient and time saving search.

Data curation

BacVacDB database was created by manually curating the appropriate information obtained from Google search relevant to bacteria and different bacterial vaccines used for preventing human bacterial diseases. Information was also obtained by searching for relevant literature from PubMed and retrieving it from reliable websites related to vaccines like U.S. Food and Drug Administration (FDA)[https://www.fda.gov/], WHO [https://www.who.int/], CDC [https://www.cdc.gov], Electronic medicines compendium (emc)[https://www.medicines.org.uk/emc/], ClinicalTrials.gov [https://clinicaltrials.gov/] etc. All this information was then represented in tabular format in BacVacDB web server to make it human readable.

Architecture

BacVacDB database has been built using a standard platform on the Linux-Apache-MySQL-PHP (LAMP). MySQL (version 5.7.31) has been used for organizing the data where the main table consisted of 33 different columns or fields related to vaccines, bacteria, disease and other related information and it was associated with 371 entries of the database.

Apache (version 2.4.46) as HTTP server was used for designing this database. HTML (version 5), PHP (version 7.3.21), CSS (version 3) and JavaScript (version 1.8) were used for developing responsive front ends which are compatible with smartphones, tablets and desktops. MySQL have been used for creating the back end and PHP programming language was used to develop a common interface. The complete architecture of this database is explained in Figure 3.

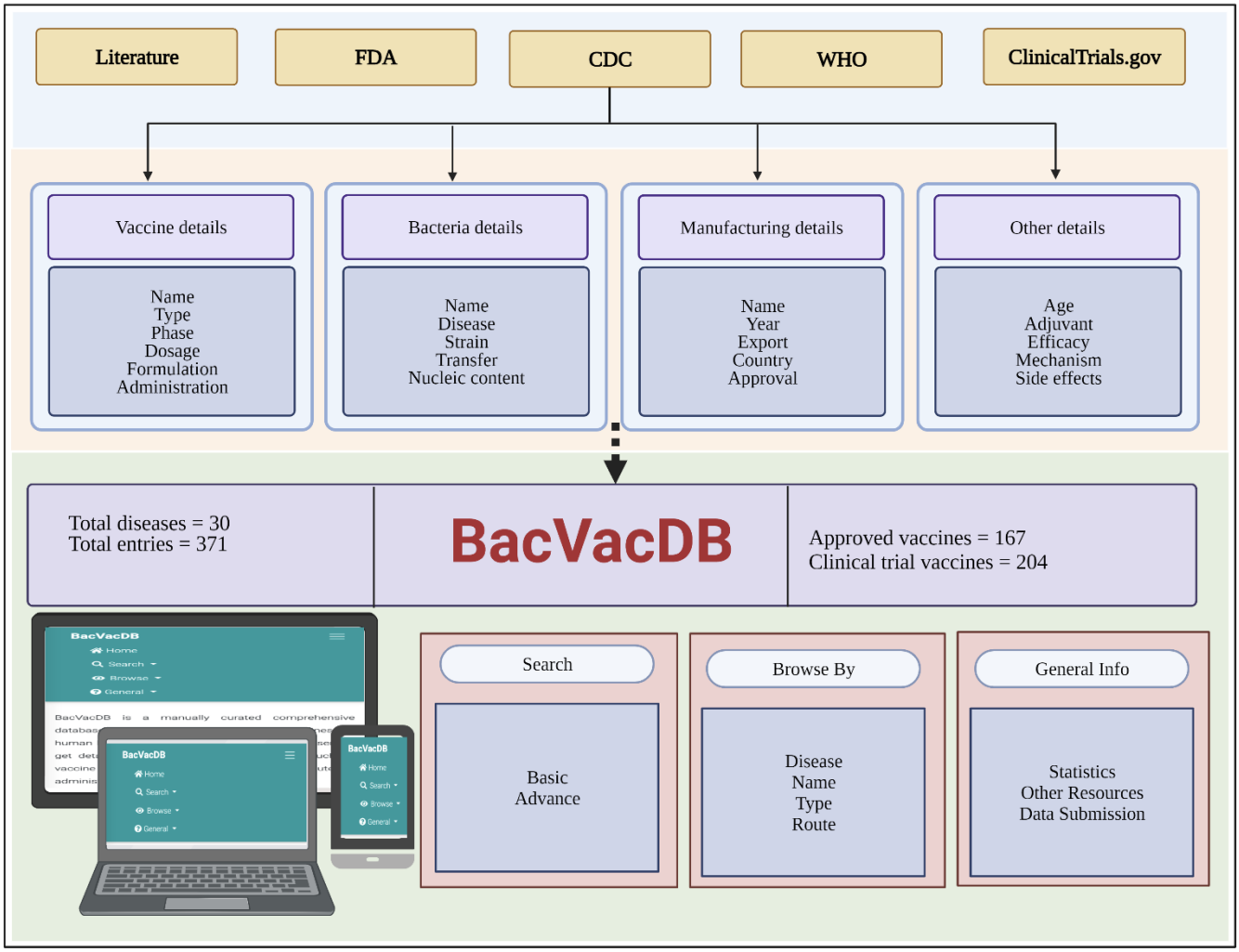

Figure 3: The architecture diagram of BacVacDB database covering different modules

Web server implementation

BacVacDB web server is a freely available server built in order to serve the community. It provides vast information about the vaccines for human bacterial diseases on a single platform.

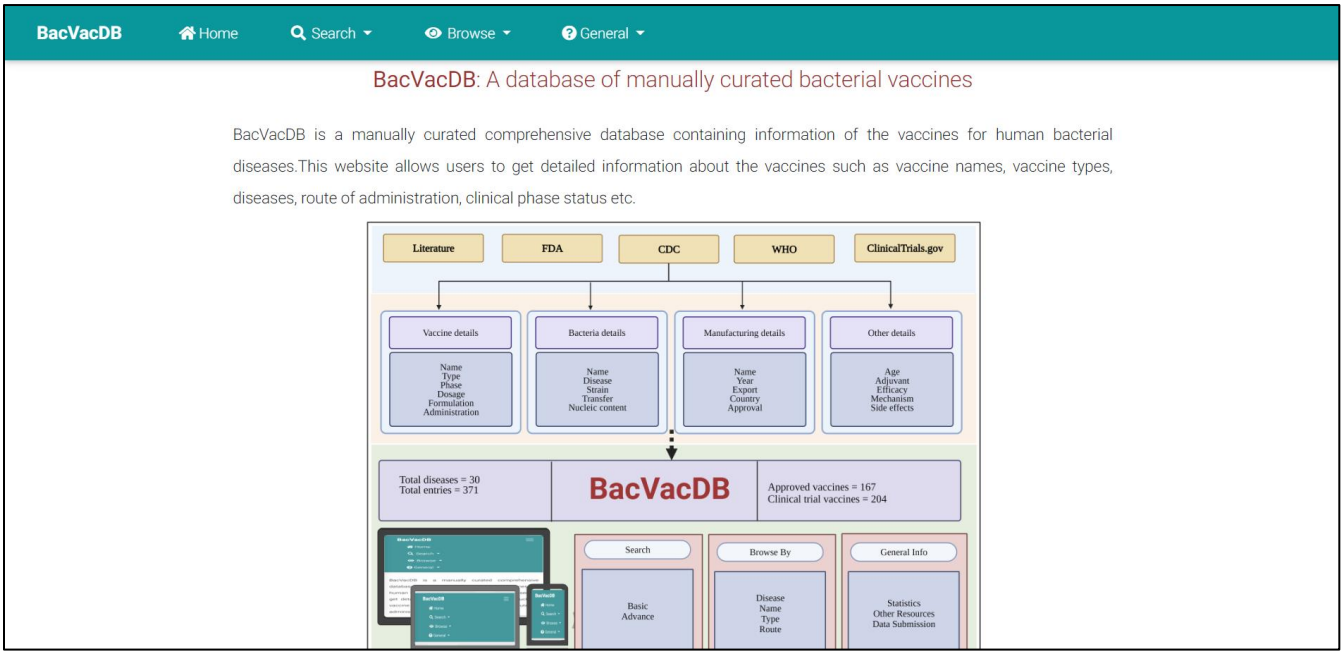

Figure 4: The home page of BacVacDB database

Search Module

This database allows two type of search options for the user – Basic and Advanced search for effortless searching.

Basic Search

This search option allows the user to search in any field or against multiple fields like disease name, vaccine name, bacteria name, type of vaccine, clinical phase status and route of administration. For any selected field, user can choose the fields he/she wants to get displayed on the result screen, thereby allowing user to customize the search. It also provides the facility to copy the data to clipboard and download in csv and excel formats. The Figure 5 represents how the basic search operation can be performed.

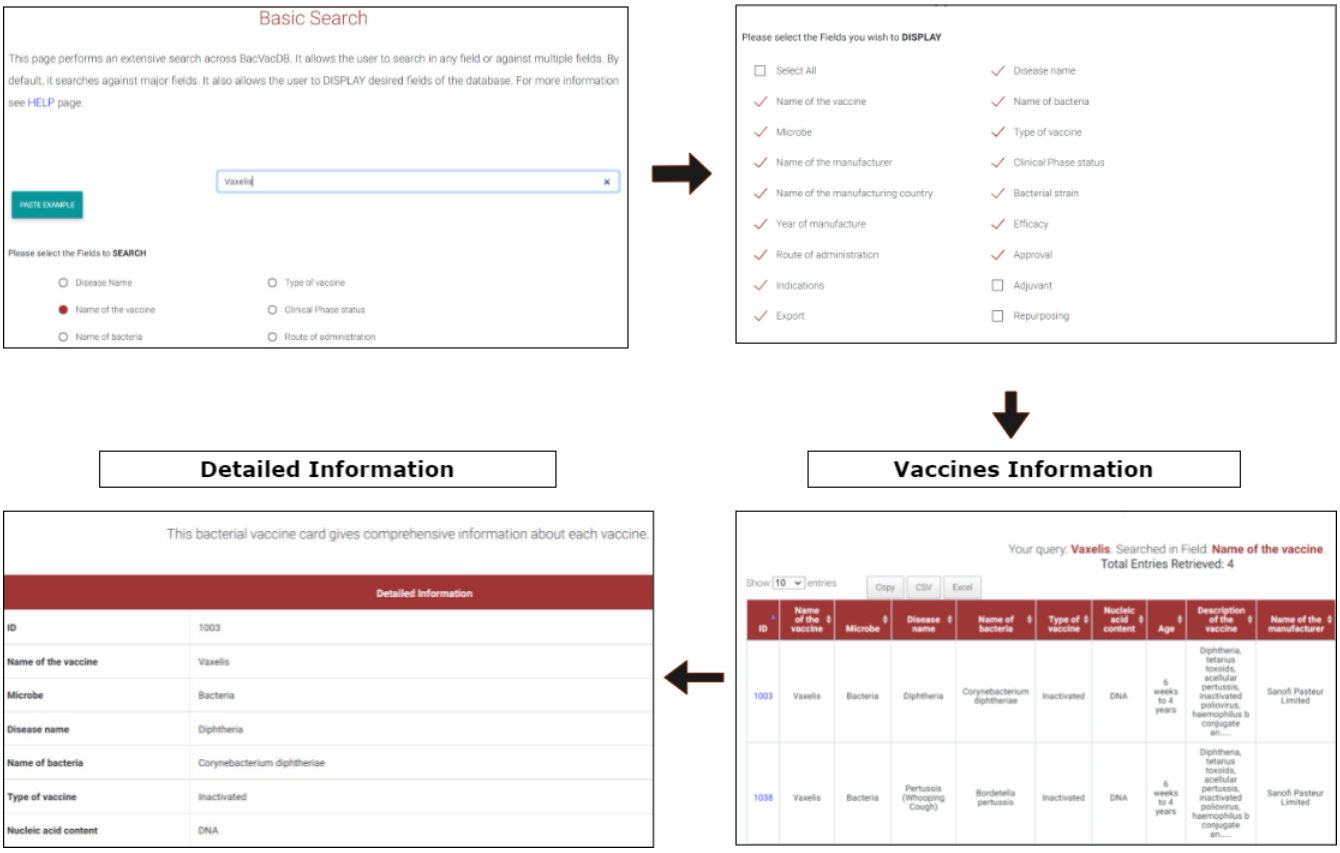

Figure 5: Step by step illustration of the basic search module in BacVacDB

Advanced Search

This search is more advanced and complex. Here, the user can give different queries simultaneously with the help of different Boolean operators such as AND, OR and NOT and the result will be displayed based on these conditions. Any particular query can also be removed using the “Remove” button on the table. Like basic search, the user can copy the data entries to clipboard and can download data in csv and excel formats. Figure 6 is the pictorial representation of the advanced search.

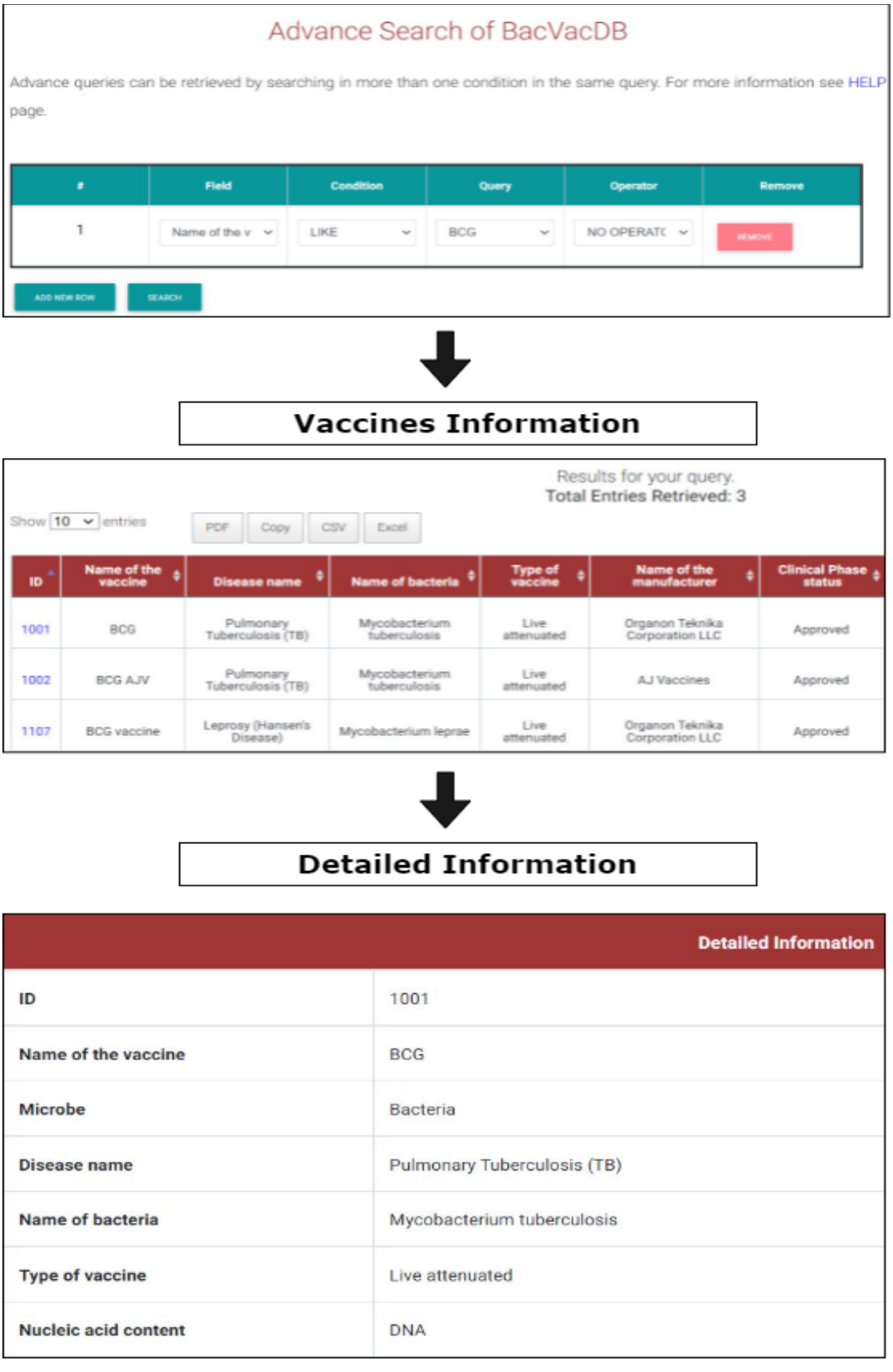

Figure 6: Step by step illustration of the advanced search module in BacVacDB

Browse module

The browse module of this database provides the facility of convenient data navigation within the database. The vaccines information enclosed in this module can also be copied to clipboard or downloaded in csv or excel format. It allows the users to browse vaccines with respect to the fields discussed below. Figure 7 displays different browsing options available in the database.

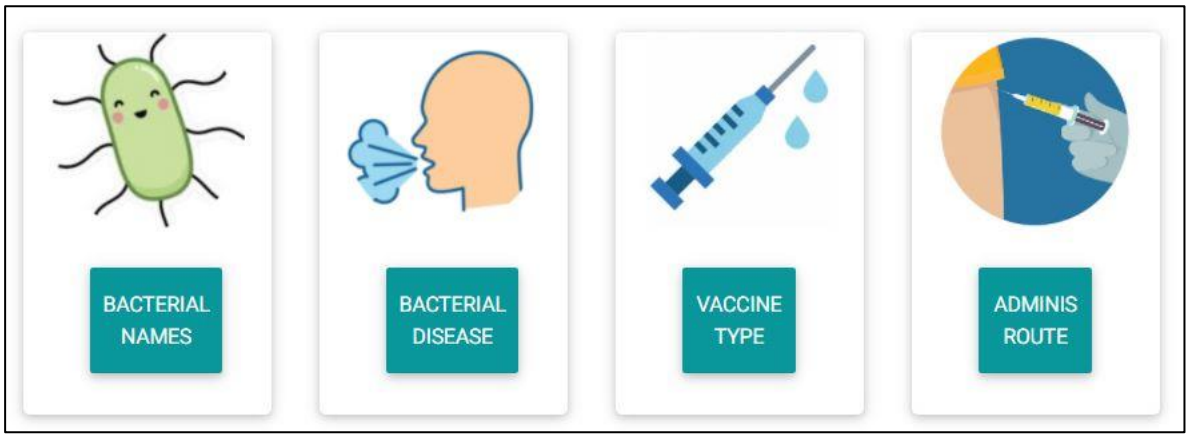

Figure 7: The browse module of BacVacDB with different browsing options

Bacterial Names: It helps the user to browse vaccines based on the bacteria names. This page displays different bacterial names and their vaccine counts. A table is displayed with the names of all the bacteria covered in this database and their corresponding vaccines count as a hyperlink. By clicking on that link, another table will be displayed comprising the information for all those vaccines (approved and clinical trials) that are involved in preventing that particular bacterial disease. This information can also be copied to clipboard or downloaded in csv or excel format.

Bacterial Diseases: This module helps in browsing the vaccines based on the name of the bacterial disease.

Vaccine Types: It allows to browse the vaccines based on their type, for instance, combination, conjugate, inactivated, subunit, recombinant etc.

Route of Administration: Here, the vaccines can be browsed with respect to their administration route, like intramuscular, oral, intravenous etc.

Database statistics

BacVacDB database holds a total of 371 vaccine entries for 30 human bacterial diseases in total extracted from literature, websites and existing databases. In this database, 368 are unique entries for the vaccines. To bifurcate further, out of the 371 bacterial vaccines, 167 entries are for approved vaccines and 204 are the vaccines undergoing different phases of human clinical trials.

Figure 8A depicts the vaccines distribution for different bacterial diseases. There are 66 vaccine entries for diphtheria, which covers around 22% of total vaccines number, then 20% of vaccines are for tetanus (lockjaw) and 15% for pertussis (whooping cough). We analyzed that majority of the vaccine are used for preventing diphtheria, tetanus and pertussis diseases.

Figure 8B presents the contribution of different approved vaccines in preventing bacterial diseases. Here, 36 entries are for tetanus followed by diphtheria (35), pertussis (26), and then other diseases.

Figure 8C represents the different types of bacterial vaccines. Under this, combination vaccines are present in the majority (131), conjugate (64), inactivated (56), recombinant (31) and many others.

Figure 8D represents the vaccines based on their route of administration. It has been observed that most of the vaccines (260) are administered intramuscularly, 37 orally, 26 subcutaneously followed by others.

Our data suggests that combination vaccines have been preventing bacterial diseases such as diphtheria, pertussis, tetanus, Hib and typhoid. While conjugate vaccines prevent diseases like typhoid, diphtheria, pertussis, Meningococcal disease, Pneumococcal, Bacillary Dysentery (Shigellosis) and Campylobacteriosis. Live attenuated vaccines for diseases such as TB, cholera, leprosy, plague, typhoid, Bacillary dysentery (Shigellosis), E. coli Infections and Tularemia (Rabbit fever).

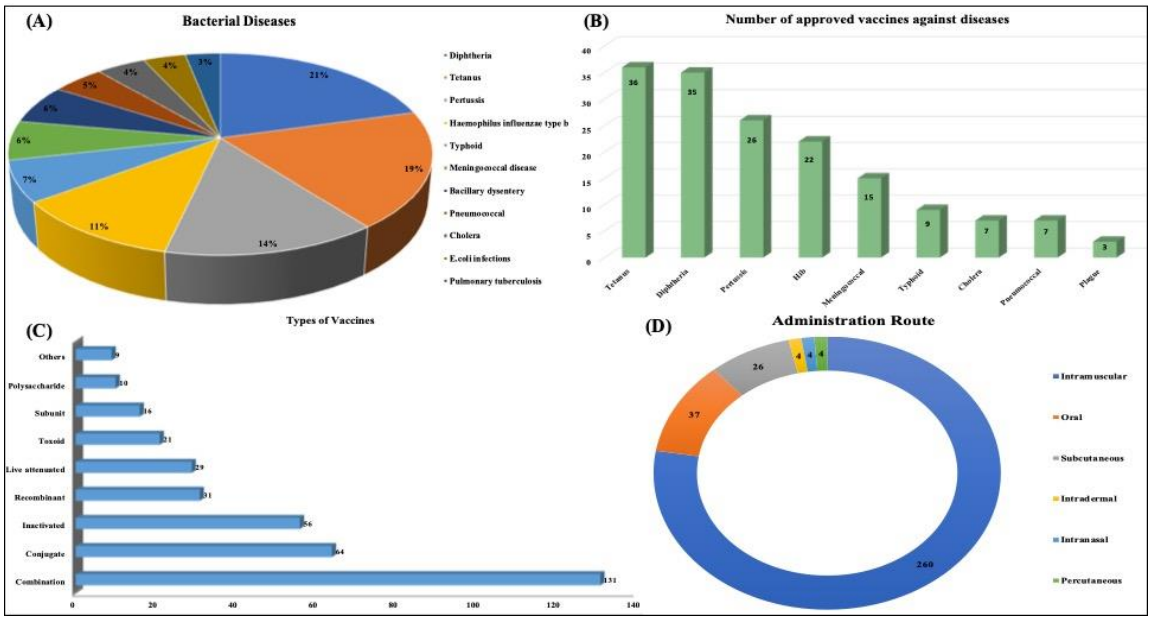

Figure 8: BacVacDB statistics with respect to bacterial diseases, number of approved vaccines, type of vaccines and route of administration

Web server availability

Bacterial vaccines play an important role in preventing diseases caused by harmful strains of bacteria. A number of vaccines are already approved by the regulatory authorities [48][49] and are in use, while others are undergoing different phases of human clinical trials. Under this study, we have curated important information related to vaccines for human bacterial diseases in order to serve the community and all of this information is made available as a freely available web resource named "BacVacDB". This database holds relevant important about vaccines like their name, types, dosage, mechanism of action etc. on a single platform for efficient search. This server is compatible with all the latest gadgets and it could be freely accessed from https://webs.iiitd.edu.in/raghava/bacvacdb/.

Comparison with existing methods

BacVacDB is the first database which is solely dedicated to bacterial vaccines for human use. It consists of 371 vaccine entries for 30 bacterial diseases which is a vast piece of information as compared to other databases such as Vaccine Investigation and Online Information Network (VIOLIN) (http://www.violinet.org/). The Vaxquery component of the VIOLIN database mostly

contain vaccines which are licensed for animals (cattle, pigs) and only covers around 135 vaccines that are for human use for preventing 16 human bacterial diseases. Huvax (https://www.violinet.org/huvax/), another component of VIOLIN database, is a web based human licensed vaccine database that contains 110 human licensed vaccines for 12 bacterial diseases, which is less as compared to BacVacDB. Another knowledgebase, Vaccine Knowledge Project (https://vk.ovg.ox.ac.uk/vk/), managed by Oxford Vaccine Group holds around 31 vaccines for 7 human bacterial diseases. Therefore, in this study, we have made an attempt to compile all the relevant information about bacterial vaccines for treating diseases in humans. We have incorporated all the details on a single platform in the form of a database or web resource named "BacVacDB". Some other information like year of manufacture, form of presentation, approval, repurposing, post-vaccination, dose type is also incorporated.

Applications of database

To the best of author's knowledge, currently no database exists which is solely dedicated to the bacterial vaccines. Thus, a single platform with all the relevant information about bacterial vaccines and its corresponding diseases would make it easier and save time to mine information related to them. It is a small contribution to the scientific community working in the field of vaccines and immune-informatics, pharmaceutical industry and researchers to gather relevant information and will help in the development of novel vaccine candidates.

Discussion

One of the greatest challenges of 21st century is the treatment and preventing of bacterial diseases. The overuse or misuse of antibiotics make the bacteria resistant to it, which is very dangerous for human health since we can't suppress the growth of the bacteria if it becomes resistant. This is when vaccines come in the picture for our rescue. The bacterial vaccines play an important role in our lives and they help in preventing critical diseases caused by the strains of bacteria. It would be of great help if all the bacterial vaccines could be accessed from a single platform by the scientific community or general public to save their time and effort. Taking this perspective in consideration, BacVacDB database was created which is a web-based platform containing massive information related to bacterial vaccines for human use. It is a user-friendly interface and could be freely accessed from the below link https://webs.iiitd.edu.in/raghava/bacvacdb. We believe that this database will be very useful for the scientific community and general public.

Chapter 3

Prediction of bacterial protective antigens

Introduction

Protective antigens can be defined as the antigens that are targeted by the host acquired immunity and able to induce protection in the host against many infectious and non-infectious diseases [50]. When protective antigens are introduced in the host, they stimulate the production of antibodies and induces the cell mediated immunity against the pathogen causing the disease [51] They elicit a protective immune response which is generally experimentally verified on any laboratory animal model [52]. They are an important component of research areas, and plays a vital role in vaccine development, drug design and as biological markers for the diagnosis of different diseases, for example they could be used in the diagnosis of AIDS based on their presence or absence. They are also proven to be helpful in the control of allergies [51]. Identification of protective antigens is the most critical step and after they are identified, researchers use these protective antigens to create new and improved DNA and subunit vaccines for different diseases. The identification process of the protective antigens is illustrated in figure 9. In this study, our aim is to predict whether a particular protein could be used as a bacterial protective antigen (BPAgs) or not. We have developed prediction models using different machine learning (ML) techniques to classify a particular protein as BPAgs. This will help in the development of different vaccines against bacterial diseases by utilizing these BPAgs.

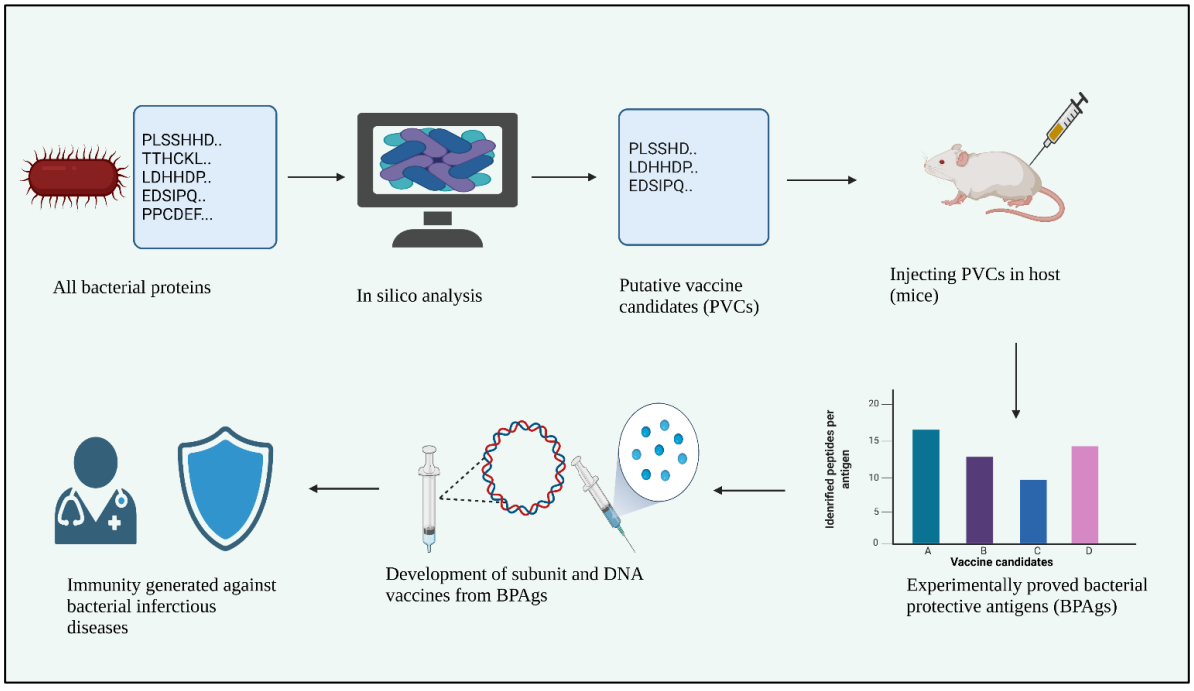

Figure 9: Figure depicting the identification process of BPAgs for vaccine development

Materials and methods

Dataset curation

The positive data for BPAgs was extracted from the Protegen [45](https://www.violinet.org/protegen/), which is a web based central database of protective antigens. It contains 660 experimentally verified BPAgs in total, out of which 195 belong to Gram positive and 465 to Gram negative bacteria. The negative data was collected from AlgPred2 [53] server (https://webs.iiitd.edu.in/raghava/algpred2/) which contains all the proteins that do not cause allergy or any other reaction in humans. The total non-protective sequences were 10, 075 in total.

The positive data for BPAgs consist of 655 sequences from the Protegen database, while the negative data contains 10, 075 sequences which is a very large number of sequences as compared to the positive data. In order to keep the number of sequences as equal, we randomly selected 655 sequences from the negative data to make the total number of sequences as consistent.

Data preprocessing

The negative dataset was already preprocessed but we performed some basic preprocessing step on the positive dataset, like removing the sequences containing non-standard amino acids (BJOUZX). There are 655 positive and 655 negative sequences in the final dataset.

Internal and external validation

The datasets were then randomly divided into two parts: (i) training dataset which holds 80% data and (ii) validation dataset with remaining 20% data [54]. The training files consist of 524 sequences each, while the validation files consist of 131 sequences each. In internal validation, we developed prediction models which were then evaluated using five-fold cross validation technique [55]. In this technique, sequences are divided randomly into five folds, out of which four are used for training and the remaining fifth one is used for testing. This process is repeated five times so that each dataset is used at least once for testing. The performance of all five sets were averaged in order to calculate the final result. In the case of external validation, we evaluated the performance of the model, which was developed on the training dataset on the validation dataset. The architecture of creating dataset and building models for the prediction of BPAgs is shown in Figure 10.

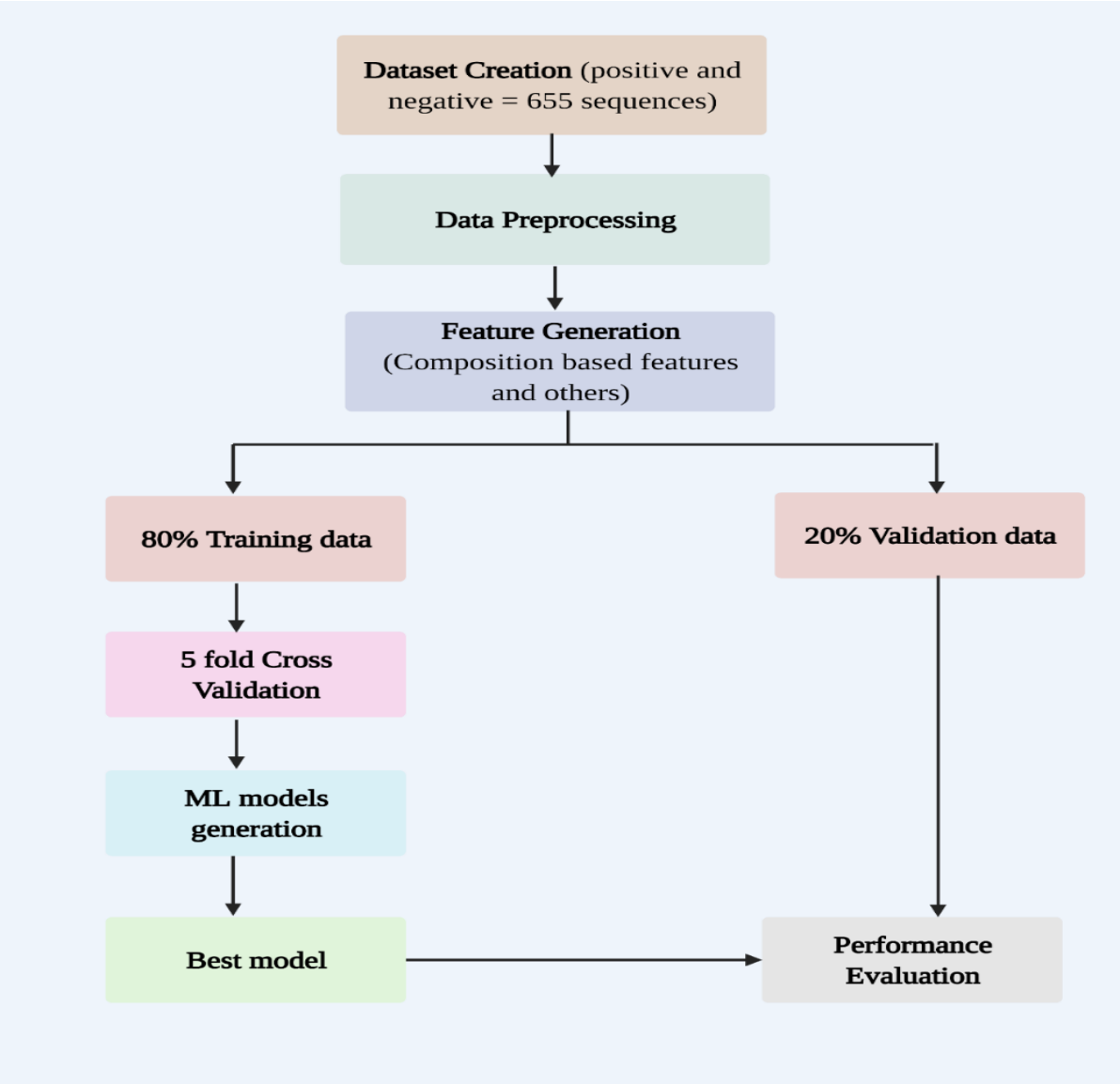

Figure 10: Schematic representation of methodology followed for creating dataset and building models for the prediction of BPAgs

Input features for prediction

We used the standalone version of Pfeature (https://webs.iiitd.edu.in/raghava/pfeature/) which is a web server for computing the peptide and protein features. We computed 14 features in total, which are described in Table 1 with their feature vectors.

Table 1: List of different features used in the study with their vector size

Features	Vector size
Amino Acid Composition (AAC)	20
Dipeptide Composition (DPC)	400
Tripeptide Composition (TPC)	8000
Atom Type Composition (ATC)	5
Bond Type Composition (BTC)	4
Physico-Chemical Properties (PCP)	30
Residue Repeats Index (RRI)	20
Distance Distribution of Repeats (DDR)	20
Shannon Entropy of a Protein (SEP)	1
Shannon Entropy of a Residue (SER)	20
Shannon Entropy of Physicochemical Property (SPC)	25
Pseudo Amino Acid Composition (PAAC)	21
Amphiphilic Pseudo Amino Acid Composition (APAAC)	23
Quasi-Sequence Order (QSO)	42

ML techniques

We implemented various ML techniques using the python scikit-learn package. It features various classification, regression and clustering algorithms. In our study, we used seven ML classifiers from this package for the BPAgs prediction. These classifiers include Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbors (KNN), Support Vector Classification (SVC), Gaussian Naive Bayes (GNB) and XGBoost (XGB). We tuned different parameters which are present on these classifiers during the run and reported the results that are obtained on the best parameters.

Performance measures

The performance of our methods was measured using different threshold dependent and threshold independent parameters. Threshold dependent parameters are Sensitivity (Sen), Specificity (Spc), Accuracy (Acc) and Matthews Correlation Coefficient (MCC). Below is the formula for these parameters:

$$Sen = \frac{TP}{TP + FN} * 100$$

$$Spc = \frac{TN}{TP + FN} * 100$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} * 100$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP = correct positive predictions; TN = correct negative predictions; FP = false positive predictions; FN = false negative predictions. In case of threshold independent parameters, Area Under Receiver Operating Characteristics (AUROC) curve was calculated where a ROC curve was drawn between false positive and false negatives.

Results

Machine learning model performance on various features

AAC based models

It tells us the fraction of each amino acid type within a protein or peptide. It can be used for classifying different proteins and peptides. In our study, we used this feature for developing prediction methods using ML techniques. As per the result, RF based models were better as compared to other models with an AUC of 0.91 on training and 0.92 on validation dataset.

Table 2: Performance of different ML based models using AAC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	71.95	72.71	72.33	0.79	0.45	70.23	77.86	74.05	0.79	0.48
RF	**84.16**	**83.97**	**84.07**	**0.91**	**0.68**	**87.02**	**83.97**	**85.50**	**0.92**	**0.71**
LR	79.77	79.96	79.87	0.88	0.60	84.73	80.15	82.44	0.89	0.65
XGB	81.68	81.87	81.78	0.89	0.64	82.44	81.68	82.06	0.89	0.64
KN	82.63	82.25	82.44	0.90	0.65	83.20	86.26	84.73	0.91	0.70
GNB	78.44	78.05	78.24	0.85	0.57	85.50	80.92	83.21	0.89	0.67
SVC	79.96	79.58	79.77	0.88	0.60	84.73	80.15	82.44	0.89	0.65

DPC based models

It indicates the composition of the residues present in a pair in the peptide. When used for developing prediction methods, RF based models performed better with an AUC of 0.91 on training and 0.92 on validation datasets.

Table 3: Performance of different ML based models using DPC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	63.74	63.93	63.84	0.69	0.28	68.70	68.70	68.70	0.77	0.37
RF	**83.40**	**82.06**	**82.73**	**0.91**	**0.66**	**83.97**	**79.39**	**81.68**	**0.92**	**0.63**
LR	81.68	81.68	81.68	0.89	0.63	83.97	83.21	83.59	0.90	0.67
XGB	80.92	81.10	81.01	0.89	0.62	80.92	84.73	82.82	0.90	0.66
KN	80.15	81.68	80.92	0.89	0.62	76.34	83.21	79.77	0.90	0.60
GNB	78.05	78.05	78.05	0.84	0.56	75.57	81.68	78.63	0.84	0.57
SVC	80.53	84.16	82.35	0.90	0.65	86.26	75.57	80.92	0.89	0.62

TPC based models

A tripeptide is formed from three consecutive amino acids and it provides local order. It was used for developing prediction models for classification where RF turns out to be best with an AUC of 0.90 on training and 0.91 on validation dataset.

Table 4: Performance of different ML based models using TPC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	59.92	60.12	60.02	0.61	0.20	57.25	66.41	61.83	0.66	0.24
RF	**80.15**	**83.78**	**81.97**	**0.90**	**0.64**	**83.20**	**79.39**	**81.30**	**0.91**	**0.63**
LR	82.06	81.49	81.78	0.90	0.64	80.15	81.68	80.92	0.89	0.62
XGB	74.05	73.66	73.86	0.83	0.48	78.63	73.28	75.95	0.85	0.52
KN	80.73	79.58	80.15	0.88	0.60	74.81	78.63	76.72	0.88	0.54
GNB	49.24	90.84	70.04	0.70	0.44	47.33	85.50	66.41	0.66	0.36

SVC	80.15	82.82	81.49	0.90	0.63	80.92	79.39	80.15	0.88	0.60

PAAC based models

It is used to compute the pseudo amino acid composition of a peptide. In our study when we used PAAC based models, then RF and XGB performed better as compared to other models. RF achieved an AUC of 0.91 on training and 0.92 on validation dataset while XGB achieved an AUC of 0.89 on training and 0.92 on validation dataset.

Table 5: Performance of different ML based models using PAAC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	72.14	71.95	72.04	0.79	0.44	69.47	71.76	70.61	0.78	0.41
RF	**84.54**	**84.92**	**84.73**	**0.91**	**0.70**	**84.73**	**84.73**	**84.73**	**0.92**	**0.70**
LR	80.73	80.73	80.73	0.88	0.62	83.97	83.21	83.59	0.89	0.67
XGB	**83.02**	**83.59**	**83.30**	**0.89**	**0.67**	**80.15**	**88.55**	**84.35**	**0.92**	**0.69**
KN	82.63	82.25	82.44	0.90	0.65	83.21	86.26	84.73	0.91	0.70
GNB	78.05	78.24	78.15	0.85	0.56	81.68	80.92	81.30	0.89	0.63
SVC	79.96	79.77	79.87	0.88	0.60	84.73	80.15	82.44	0.89	0.65

APAAC based models

It computes amphiphilic pseudo amino acid composition of a peptide. In our study, RF based models were better as compared to other models with an AUC of 0.91 on training and 0.92 on validation dataset.

Table 6: Performance of different ML based models using APAAC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	68.89	70.23	69.56	0.78	0.39	75.57	75.57	75.57	0.85	0.51
RF	**83.40**	**83.40**	**83.40**	**0.91**	**0.67**	**84.73**	**87.02**	**85.88**	**0.92**	**0.72**
LR	80.73	80.34	80.53	0.88	0.61	83.97	83.21	83.59	0.90	0.67

XGB	81.68	81.49	81.58	0.89	0.63	80.15	83.20	81.68	0.90	0.63
KN	81.68	82.06	81.87	0.90	0.64	83.21	87.79	85.50	0.91	0.71
GNB	78.44	79.01	78.72	0.86	0.57	79.39	82.44	80.92	0.89	0.62
SVC	79.77	80.15	79.96	0.88	0.60	84.73	80.15	82.44	0.89	0.65

DDR based models

It computes the distribution of residue based on the distance from N-terminal, C-terminal and inter distances between same residue within the peptide or protein sequence. When used this feature for developing prediction models, RF performed better with an AUC of 0.90 on training and 0.91 on validation datasets.

Table 7: Performance of different ML based models using DDR

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	68.51	68.70	68.61	0.74	0.37	70.99	78.63	74.81	0.80	0.50
RF	**81.49**	**82.25**	**81.87**	**0.90**	**0.64**	**83.97**	**80.15**	**82.06**	**0.91**	**0.64**
LR	77.29	78.05	77.67	0.84	0.55	77.10	78.63	77.86	0.83	0.56
XGB	81.49	81.49	81.49	0.88	0.63	83.97	77.86	80.92	0.90	0.62
KN	77.10	76.91	77.00	0.86	0.54	77.86	73.28	75.57	0.87	0.51
GNB	32.06	89.12	60.59	0.66	0.26	23.66	93.13	58.40	0.68	0.23
SVC	77.67	77.10	77.39	0.83	0.55	77.10	76.34	76.72	0.83	0.53

QSO based models

It is used to compute Quasi-Sequence Order of a peptide. In our study, RF was better as compared to other models. It achieved an AUC of 0.91 on training and 0.92 on validation dataset.

Table 8: Performance of different ML based models using QSO

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	69.66	70.04	69.85	0.77	0.40	75.57	70.99	73.28	0.80	0.47

RF	**83.21**	**82.06**	**82.63**	**0.91**	**0.65**	**87.02**	**81.68**	**84.35**	**0.92**	**0.69**
LR	80.15	79.96	80.06	0.89	0.60	83.21	81.68	82.44	0.90	0.65
XGB	82.25	82.63	82.44	0.90	0.65	82.44	80.15	81.30	0.89	0.63
KN	82.06	82.63	82.35	0.90	0.65	84.73	84.73	84.73	0.91	0.70
GNB	78.05	78.44	78.24	0.85	0.57	75.57	85.50	80.53	0.88	0.61
SVC	81.68	81.87	81.78	0.90	0.64	81.68	77.86	79.77	0.89	0.60

RRI based models

It measures the number of continuous runs of a residue type in a sequence. When used this feature for developing prediction models, RF proved to be better with an AUC of 0.83 on training and 0.84 on validation dataset.

Table 9: Performance of different ML based models using RRI

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	63.36	63.74	63.55	0.69	0.27	68.70	59.54	64.12	0.70	0.28
RF	**73.66**	**73.09**	**73.38**	**0.83**	**0.47**	**78.63**	**72.52**	**75.57**	**0.84**	**0.51**
LR	69.47	69.85	69.66	0.76	0.39	70.23	70.23	70.23	0.77	0.40
XGB	72.14	71.95	72.04	0.81	0.44	70.99	70.99	70.99	0.79	0.42
KN	71.76	72.33	72.04	0.80	0.44	77.86	74.05	75.95	0.83	0.52
GNB	61.26	68.51	64.86	0.70	0.30	60.31	64.12	62.21	0.67	0.24
SVC	69.08	69.28	69.18	0.76	0.38	74.81	65.65	70.23	0.76	0.41

ATC based models

It is the fraction of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms present in a protein sequence. In our study, KN was better as compared to other models and achieved an AUC of 0.82 on training and 0.86 on validation dataset.

Table 10: Performance of different ML based models using ATC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	71.18	71.37	71.28	0.77	0.43	70.99	73.28	72.14	0.77	0.44
RF	76.72	77.29	77.00	0.83	0.54	80.15	77.10	78.63	0.85	0.57
LR	75.38	75.57	75.48	0.81	0.51	74.81	79.39	77.10	0.85	0.54
XGB	75.57	75.76	75.67	0.82	0.51	77.10	71.76	74.43	0.83	0.49
KN	**75.19**	**75.19**	**75.19**	**0.82**	**0.50**	**80.92**	**74.81**	**77.86**	**0.86**	**0.56**
GNB	72.90	72.52	72.71	0.79	0.45	68.70	79.39	74.05	0.82	0.48
SVC	75.95	74.24	75.10	0.81	0.50	77.86	76.34	77.10	0.85	0.54

BTC based models

It computes the bond composition of each amino acid residue of the sequence. Here, four types of bonds are considered that are total number of bonds, hydrogen bond, single bond and double bond. In our study when we developed prediction models based on this feature, SVC performed better. It achieved an AUC of 0.71 on training and 0.73 on validation dataset.

Table 11: Performance of different ML based models using BTC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	52.67	54.77	53.72	0.56	0.07	59.54	56.49	58.01	0.62	0.16
RF	56.68	55.53	56.11	0.59	0.12	61.07	54.20	57.63	0.61	0.15
LR	67.56	63.74	65.65	0.70	0.31	62.60	66.41	64.50	0.72	0.29
XGB	56.30	57.06	56.68	0.58	0.13	61.83	48.86	55.34	0.60	0.11
KN	54.96	52.10	53.53	0.57	0.07	61.07	55.72	58.40	0.62	0.17
GNB	52.10	50.57	51.34	0.51	0.03	81.68	13.74	47.71	0.44	-0.06
SVC	**64.89**	**65.27**	**65.08**	**0.71**	**0.30**	**61.83**	**66.41**	**64.12**	**0.73**	**0.28**

SEP based models

It computes the Shannon entropy of a protein sequence. In our study, XGB performed better using the

SEP feature. XGB achieved an AUC of 0.69 on training and AUC of 0.69 on the validation dataset.

Table 12: Performance of different ML based models using SEP

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	61.26	62.60	61.93	0.67	0.24	58.78	63.36	61.07	0.65	0.22
RF	59.16	60.12	59.64	0.66	0.19	58.02	59.54	58.78	0.65	0.18
LR	0.00	100.00	50.00	0.63	0.00	0.00	100.00	50.00	0.67	0.00
XGB	**61.64**	**63.36**	**62.50**	**0.69**	**0.25**	**54.20**	**66.41**	**60.31**	**0.69**	**0.21**
KN	63.36	63.36	63.36	0.68	0.27	54.20	61.83	58.02	0.68	0.16
GNB	58.78	61.07	59.92	0.62	0.20	58.02	51.15	54.58	0.58	0.09
SVC	49.24	50.00	49.62	0.53	-0.01	100.00	0.00	50.00	0.67	0.00

SER based models

It computes the Shannon entropy of the residues of the peptide or protein sequence. When we used this feature in our study to develop prediction models, RF and KN both performed well. RF achieved an AUC of 0.91 on training and 0.92 on validation while KN achieved an AUC of 0.90 on training and 0.92 on validation dataset.

Table 13: Performance of different ML based models using SER

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	74.62	74.62	74.62	0.81	0.49	70.99	81.68	76.34	0.82	0.53
RF	**83.59**	**84.35**	**83.97**	**0.91**	**0.68**	**84.73**	**83.21**	**83.97**	**0.92**	**0.68**
LR	80.15	81.30	80.73	0.89	0.62	84.73	79.39	82.06	0.90	0.64
XGB	80.53	80.53	80.53	0.89	0.61	78.63	83.97	81.30	0.89	0.63
KN	**81.88**	**82.25**	**82.06**	**0.90**	**0.64**	**83.21**	**86.26**	**84.73**	**0.92**	**0.70**
GNB	80.53	80.15	80.34	0.88	0.61	83.21	83.97	83.59	0.90	0.67
SVC	80.34	81.30	80.82	0.89	0.62	85.50	80.15	82.82	0.90	0.66

SPC based models

It calculates the Shannon entropy of a particular physicochemical property in a sequence. When we used this descriptor for developing models, RF proved to be better with an AUC of 0.87 on training and 0.89 on validation dataset.

Table 14: Performance of different ML based models using SPC

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	65.65	66.41	66.03	0.72	0.32	74.05	58.02	66.03	0.69	0.33
RF	**78.44**	**77.67**	**78.05**	**0.87**	**0.56**	**80.15**	**80.15**	**80.15**	**0.89**	**0.60**
LR	78.82	78.82	78.82	0.87	0.58	77.86	83.97	80.92	0.88	0.62
XGB	76.53	75.76	76.15	0.86	0.52	80.92	79.39	80.15	0.83	0.60
KN	77.48	78.05	77.77	0.85	0.56	77.10	85.50	81.30	0.87	0.63
GNB	70.23	71.18	70.70	0.77	0.41	71.76	74.81	73.28	0.81	0.47
SVC	77.48	77.48	77.49	0.84	0.55	79.39	79.39	79.39	0.87	0.59

PCP based models

It computes the fraction of each standard physico-chemical property in the sequence. In our study using this feature for developing models, highest performance was of RF with an AUC of 0.87 on training and 0.90 on validation dataset.

Table 15: Performance of different ML based models using PCP

ML	Training Dataset					Validation Dataset				
	Sen	**Spec**	**Acc**	**AUC**	**MCC**	**Sen**	**Spec**	**Acc**	**AUC**	**MCC**
DT	69.47	69.28	69.37	0.76	0.39	77.10	69.47	73.28	0.78	0.47
RF	**78.44**	**79.77**	**79.10**	**0.87**	**0.58**	**80.92**	**80.15**	**80.53**	**0.90**	**0.61**
LR	79.96	79.77	79.87	0.88	0.60	83.21	79.39	81.30	0.89	0.63
XGB	78.82	78.63	78.72	0.85	0.57	83.21	82.44	82.82	0.89	0.66
KN	73.86	73.66	73.76	0.81	0.48	77.10	78.63	77.86	0.87	0.56

GNB	72.33	71.95	72.14	0.79	0.44	72.52	74.81	73.66	0.84	0.47
SVC	74.43	74.24	74.33	0.81	0.49	79.39	74.05	76.72	0.86	0.54

Discussion

Protective antigens are the one that invokes specific and enhanced adaptive immune response. They elicit a protective immune response thereby they are very important in vaccine preparation and drug design. One of the major challenges in vaccine development is the identification of the protective antigens that give rise to protective immune response. Since the laboratory methods of isolating an antigen from a microbial pathogen are very expensive and time consuming, computational methods are useful to predict the protective antigens. In this study, we have developed ML models to predict the protein as BPAgs. The positive and negative dataset is processed from Protegen database and AlgPred2 web server respectively and then different features of the protein are extracted from the Pfeature standalone server. We have utilized different classifiers like DT, SVC, KNN, MLP, RF, LR and XGB where for most of the features or descriptors, highest AUC is observed on RF based models. These identified BPAgs can then be used by researchers for the development of subunit and DNA based bacterial vaccines to prevent against bacterial diseases. We hope that this prediction method could be useful for the researchers working in the field of vaccine development.

Conclusion

Since bacterial protective antigens are of immense importance in the research area for developing subunit and DNA vaccines against different bacterial diseases, in this study we have made an attempt to develop ML prediction models to classify the protein as BPAgs. We have used different ML classifiers for the prediction like DT, SVC, KNN, MLP, RF, LR and XGB. The highest AUC achieved is 0.92 mostly seen in the case of RF based models. We hope that it could be beneficial for the researchers and the scientific community to develop vaccines based on these protective antigens.

Chapter 4

Summary

In this study, we have made an attempt to compile the bacterial vaccines for different human bacterial diseases in order to contribute in healthcare and to serve scientific community, students and general public to gather massive information about the vaccines that are used for preventing different bacterial diseases and they have made a huge positive impact regarding the health of the people. We have made available all this massive information about vaccines, their types, mechanism, routes etc. in the form of a web server BacVacDB which can be freely accessed from https://webs.iiitd.edu.in/raghava/bacvacdb/. This server provides the facility to allow users to perform basic and advanced search regarding the vaccines and to browse the vaccines with respect to different options like bacterial diseases, bacterial names, vaccine types and route of administration. We have manually curated the information for 167 approved and 204 clinical trial vaccines.

Also, we have developed a prediction method based on different ML classification techniques for predicting if a protein could be used as a BPAgs or not for the vaccine development against bacterial diseases. Here, we have collected the positive dataset from Protegen database and negative data from AlgPred2 server. We used 14 different features of proteins which were generated using Pfeature standalone tool and then applied different ML classification techniques on the same. We hope that this would be of great help to the scientific and academic community and also to general public.

Bibliography

[1] W.K. Funkhouser, Pathology: The Clinical Description of Human Disease, in: Mol. Pathol., Elsevier, 2009: pp. 197–207. https://doi.org/10.1016/B978-0-12-374419-7.00011-1.

[2] M. Jackson, L. Marks, G.H.W. May, J.B. Wilson, The genetic basis of disease, Essays Biochem. 62 (2018) 643–723. https://doi.org/10.1042/EBC20170053.

[3] M. Pigłowski, Pathogenic and Non-Pathogenic Microorganisms in the Rapid Alert System for Food and Feed., Int. J. Environ. Res. Public Health. 16 (2019). https://doi.org/10.3390/ijerph16030477.

[4] J.M. van Seventer, N.S. Hochberg, Principles of Infectious Diseases: Transmission, Diagnosis, Prevention, and Control, in: Int. Encycl. Public Heal., Elsevier, 2017: pp. 22–39. https://doi.org/10.1016/B978-0-12-803678-5.00516-6.

[5] S. Doron, S.L. Gorbach, Bacterial Infections: Overview, in: Int. Encycl. Public Heal., Elsevier, 2008: pp. 273–282. https://doi.org/10.1016/B978-012373960-5.00596-7.

[6] R. Ramanan, B.-H. Kim, D.-H. Cho, H.-M. Oh, H.-S. Kim, Algae–bacteria interactions: Evolution, ecology and emerging applications, Biotechnol. Adv. 34 (2016) 14–29. https://doi.org/10.1016/j.biotechadv.2015.12.003.

[7] Introduction to Bacteriology, 1996. http://www.ncbi.nlm.nih.gov/pubmed/21413299.

[8] M.R.J. Salton, K.-S. Kim, Structure, 1996. http://www.ncbi.nlm.nih.gov/pubmed/21413343.

[9] Y.-J. Zhang, S. Li, R.-Y. Gan, T. Zhou, D.-P. Xu, H.-B. Li, Impacts of gut bacteria on human health and diseases., Int. J. Mol. Sci. 16 (2015) 7493–519. https://doi.org/10.3390/ijms16047493.

[10] P. Del Giudice, Skin Infections Caused by Staphylococcus aureus., Acta Derm. Venereol. 100 (2020) adv00110. https://doi.org/10.2340/00015555-3466.

[11] C.B. Whitlow, Bacterial sexually transmitted diseases., Clin. Colon Rectal Surg. 17 (2004) 209–14. https://doi.org/10.1055/s-2004-836940.

[12] T. Bintsis, Foodborne pathogens., AIMS Microbiol. 3 (2017) 529–563. https://doi.org/10.3934/microbiol.2017.3.529.

[13] M.I. Hutchings, A.W. Truman, B. Wilkinson, Antibiotics: past, present and future, Curr. Opin. Microbiol. 51 (2019) 72–80. https://doi.org/10.1016/j.mib.2019.10.008.

[14] R.J. Fair, Y. Tor, Antibiotics and Bacterial Resistance in the 21st Century, Perspect. Medicin. Chem. 6 (2014) PMC.S14459. https://doi.org/10.4137/PMC.S14459.

[15] F. Prestinaci, P. Pezzotti, A. Pantosti, Antimicrobial resistance: a global multifaceted phenomenon., Pathog. Glob. Health. 109 (2015) 309–18. https://doi.org/10.1179/2047773215Y.0000000030.

[16] About Antibiotic Resistance | CDC, (n.d.). https://www.cdc.gov/drugresistance/about.html (accessed May 13, 2022).

[17] Antibiotic resistance, (n.d.). https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance (accessed May 13, 2022).

[18] R.P.N. Mishra, E. Oviedo-Orta, P. Prachi, R. Rappuoli, F. Bagnoli, Vaccines and antibiotic resistance., Curr. Opin. Microbiol. 15 (2012) 596–602. https://doi.org/10.1016/j.mib.2012.08.002.

[19] J.T. Poolman, Expanding the role of bacterial vaccines into life-course vaccination strategies and prevention of antimicrobial-resistant infections., NPJ Vaccines. 5 (2020) 84. https://doi.org/10.1038/s41541-020-00232-0.

[20] R.P. Mishra, E. Oviedo-Orta, P. Prachi, R. Rappuoli, F. Bagnoli, Vaccines and antibiotic resistance, Curr. Opin. Microbiol. 15 (2012) 596–602. https://doi.org/10.1016/j.mib.2012.08.002.

[21] I. Hajj Hussein, N. Chams, S. Chams, S. El Sayegh, R. Badran, M. Raad, A. Gerges-Geagea, A. Leone, A. Jurjus, Vaccines Through Centuries: Major Cornerstones of Global Health, Front. Public Heal. 3 (2015). https://doi.org/10.3389/fpubh.2015.00269.

[22] Vivotif Package Insert USA-Updated September 2013-Increase of Upper Specification Limit Vivotif® Typhoid Vaccine Live Oral Ty21a, (n.d.).

[23] S.A. Marathe, A. Lahiri, V.D. Negi, D. Chakravortty, Typhoid fever & vaccine development: a partially answered question., Indian J. Med. Res. 135 (2012) 161–9. http://www.ncbi.nlm.nih.gov/pubmed/22446857.

[24] A. Detmer, J. Glenting, Live bacterial vaccines--a review and identification of potential hazards., Microb. Cell Fact. 5 (2006) 23. https://doi.org/10.1186/1475-2859-5-23.

[25] TDVAX | FDA, (n.d.). https://www.fda.gov/vaccines-blood-biologics/vaccines/tdvax (accessed March 25, 2022).

[26] T.S.P. Tiwari, M. Wharton, Diphtheria toxoid, in: Vaccines, Elsevier, 2013: pp. 153–166. https://doi.org/10.1016/B978-1-4557-0090-5.00024-0.

[27] K.A. Bond, L.J. Franklin, B. Sutton, S.M. Firestone, Q-Vax Q fever vaccine failures, Victoria, Australia 1994-2013., Vaccine. 35 (2017) 7084–7087. https://doi.org/10.1016/j.vaccine.2017.10.088.

[28] Y.H. Khan, A. Saifullah, T.H. Mallhi, Bacterial Vaccines, in: Encycl. Infect. Immun., Elsevier, 2022: pp. 530–544. https://doi.org/10.1016/B978-0-12-818731-9.00170-1.

[29] L.E. Nigrovic, K.M. Thompson, The Lyme vaccine: a cautionary tale., Epidemiol. Infect. 135

(2007) 1–8. https://doi.org/10.1017/S0950268806007096.

[30] Pinkbook: Vaccine Administration | CDC, (n.d.). https://www.cdc.gov/vaccines/pubs/pinkbook/vac-admin.html (accessed May 20, 2022).

[31] C. Barranco, The first live attenuated vaccines, Nat. Res. 2021. (2020). https://www.nature.com/articles/d42859-020-00008-5 (accessed March 29, 2022).

[32] K.A. Smith, Louis pasteur, the father of immunology?, Front. Immunol. 3 (2012) 68. https://doi.org/10.3389/fimmu.2012.00068.

[33] S. Luca, T. Mihaescu, History of BCG Vaccine., Maedica (Buchar). 8 (2013) 53–8. http://www.ncbi.nlm.nih.gov/pubmed/24023600.

[34] B.J. Hawgood, Albert Calmette (1863-1933) and Camille Guérin (1872-1961): the C and G of BCG vaccine., J. Med. Biogr. 15 (2007) 139–46. https://doi.org/10.1258/j.jmb.2007.06-15.

[35] S.K. Verma, U. Tuteja, Plague Vaccine Development: Current Research and Future Trends., Front. Immunol. 7 (2016) 602. https://doi.org/10.3389/fimmu.2016.00602.

[36] I.J. Amanna, M.K. Slifka, Successful Vaccines., Curr. Top. Microbiol. Immunol. 428 (2020) 1–30. https://doi.org/10.1007/82_2018_102.

[37] WHO EMRO | Vaccine and vaccination | Meningococcal disease | Health topics, (n.d.). http://www.emro.who.int/health-topics/meningococcal-disease/vaccine-vaccination.html (accessed May 20, 2022).

[38] Pinkbook: Pneumococcal Disease | CDC, (n.d.). https://www.cdc.gov/vaccines/pubs/pinkbook/pneumo.html (accessed March 29, 2022).

[39] J.R. Gilsdorf, Hib Vaccines: Their Impact on Haemophilus influenzae Type b Disease., J. Infect. Dis. 224 (2021) S321–S330. https://doi.org/10.1093/infdis/jiaa537.

[40] R. Gasparini, D. Panatto, Meningococcal glycoconjugate vaccines., Hum. Vaccin. 7 (2011) 170–82. https://doi.org/10.4161/hv.7.2.13717.

[41] M.F. Tosi, Innate immune responses to infection., J. Allergy Clin. Immunol. 116 (2005) 241–9; quiz 250. https://doi.org/10.1016/j.jaci.2005.05.036.

[42] H.R. Mirzaei, Adaptive Immunity, in: Encycl. Infect. Immun., Elsevier, 2022: pp. 39–55. https://doi.org/10.1016/B978-0-12-818731-9.00028-8.

[43] L.M.F. Merlo, L. Mandik-Nayak, Adaptive Immunity, in: Cancer Immunother., Elsevier, 2013: pp. 25–40. https://doi.org/10.1016/B978-0-12-394296-8.00003-8.

[44] M.S. Rahman, M.K. Rahman, S. Saha, M. Kaykobad, M.S. Rahman, Antigenic: An improved prediction model of protective antigens, Artif. Intell. Med. 94 (2019) 28–41.

https://doi.org/10.1016/j.artmed.2018.12.010.

[45] B. Yang, S. Sayers, Z. Xiang, Y. He, Protegen: a web-based protective antigen database and analysis system., Nucleic Acids Res. 39 (2011) D1073-8. https://doi.org/10.1093/nar/gkq944.

[46] V. Vetter, G. Denizer, L.R. Friedland, J. Krishnan, M. Shapiro, Understanding modern-day vaccines: what you need to know, Ann. Med. 50 (2018) 110–120. https://doi.org/10.1080/07853890.2017.1407035.

[47] J.T. Poolman, Expanding the role of bacterial vaccines into life-course vaccination strategies and prevention of antimicrobial-resistant infections, Npj Vaccines. 5 (2020) 84. https://doi.org/10.1038/s41541-020-00232-0.

[48] K. Singh, S. Mehta, The clinical development process for a novel preventive vaccine: An overview, J. Postgrad. Med. 62 (2016) 4. https://doi.org/10.4103/0022-3859.173187.

[49] N.W. Baylor, Role of the national regulatory authority for vaccines, Int. J. Heal. Gov. 22 (2017) 128–137. https://doi.org/10.1108/IJHG-04-2017-0017.

[50] Y. He, Z. Xiang, Bioinformatics analysis of bacterial protective antigens in manually curated Protegen database, Procedia Vaccinol. 6 (2012) 3–9. https://doi.org/10.1016/j.provac.2012.04.002.

[51] B. Yang, S. Sayers, Z. Xiang, Y. He, Protegen: a web-based protective antigen database and analysis system, Nucleic Acids Res. 39 (2011) D1073–D1078. https://doi.org/10.1093/nar/gkq944.

[52] E. Ong, H. Wang, M.U. Wong, M. Seetharaman, N. Valdez, Y. He, Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens, Bioinformatics. 36 (2020) 3185–3191. https://doi.org/10.1093/bioinformatics/btaa119.

[53] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P.S. Raghava, AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes., Brief. Bioinform. 22 (2021). https://doi.org/10.1093/bib/bbaa294.

[54] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal-external, and external validation., J. Clin. Epidemiol. 69 (2016) 245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[55] I. Kononenko, M. Kukar, Machine Learning Basics, in: Mach. Learn. Data Min., Elsevier, 2007: pp. 59–105. https://doi.org/10.1533/9780857099440.59.