Manual Curation of Virus Based Vaccines and Prediction of B-Cell Epitopes using Machine Learning Techniques

Submitted by

Sadhana Tripathi (MT20214)

Under the guidance of

Prof. G.P.S Raghava

Head & Professor

in partial fulfilment 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 **"Manual Curation of Virus Based Vaccines and Prediction of B-Cell Epitopes using Machine Learning Techniques"** being submitted by **Sadhana Tripathi** 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.

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. G.P.S Raghava
Head & Professor (Department of Computational Biology)
Indraprastha Institute of Information Technology 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.

Sadhana Tripathi

M.Tech CB (MT20214)

Table of Contents

List of Abbreviations

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

HIV: Human immunodeficiency virus

HCV: Hepatitis C virus

SARS-CoV: SARS-associated coronavirus

MERS-CoV: Middle East respiratory syndrome coronavirus

SARS-CoV-2: Severe acute respiratory syndrome-related coronavirus-2

EMA: European Medicines Agency

WHO: World Health Organisation

FDA: Food and Drug Administration

CDC: Centers for Disease Control and Prevention

DRDO: Defence Research and Development Organisation

SAHPRA: South African Health Products Regulatory Authority

BPOM: Indonesia's drug and food authority

BLAST: Basic Local Alignment Search Tool

Blastp: Protein–protein BLAST

AAC: Amino Acid Composition

DPC: Dipeptide composition

TPC: Tripeptide composition

AAIndex: Amino Acid Index

DDOR: Distance distribution of residues

PSSM: Position specific scoring matrix

PAAC: Pseudo Amino Acid Composition

APAAC: Amphiphilic Pseudo Amino Acid Composition

SCON: Sequence Order Coupling Number

QSO: Quasi-sequence-order

RF: Random Forest

XGB: eXtreme Gradient Boosting

SVM: support vector machine

KNN: K-nearest neighbour

GNB: Gaussian Naive Bayes

LR: Logistic regression

DT: Decision tree

MHC: Major Histocompatibility Complex

List of Figures

List of Tables

Abstract

Unprecedented rise of newly emerging and re-emerging cases of infectious viral diseases have threatened humanity throughout. Vaccines stand as the safest and effective way to manage scale and spread of infectious diseases. First viral vaccine was developed in 1796 against smallpox, which successfully eradicated this deadly virus from the planet. Over the years, numerous vaccines have been developed against viral infections, which save millions of premature deaths every year. In this study we have made an attempt to integrate the resources associated with viral vaccines which are scattered across literature and websites. The manual curated repository of viral vaccines, ViralVacDB is web-based platform that provides coverage to different viral diseases and vaccines associated information. This repository provides information about 421 vaccines:139 approved vaccines and 282 vaccines under development. Various web-based interfaces have been configured in ViralVacDB to facilitate the scientific community to retrieve information in user-friendly manner. The database is available at https://webs.iiitd.edu.in/raghava/viralvacdb/.

B-cell are responsible for eliciting adaptive immunity response. Recognition of B-cell epitopes plays a crucial role in designing and development of synthetic vaccines. In past, various computational methods have been developed to predict either linear or conformational B-cell epitopes. In this study, we have made an attempt to classify linear and conformational B-cell epitopes and non-epitopes using different machine learning models based on various set of features.

Chapter 1

Introduction

Successful control and eradication of deadly infectious viruses can only be achieved and attainable by vaccination [1]. Over the internet, a number of resources are available on infectious viral diseases and vaccines to create awareness among people. Due to the ongoing pandemic scenario, substantial interest has been taken to make use of technological advancement and bridge the gap of predominant prevalent methods. Presenting the insight of vaccines associated with deadly viral diseases along with coverage on its type, phase, target strain, adjuvant, manufacturer information, dosage, targeted audience, approval status, clinical phase, administration route and clinical status along with product information is very limited. Due to the scatteredness of resources, searching information costs time and energy but its representation in a single platform in a simplistic way can lead to informed decision making, reducing the bridge of vaccine hesitation among the general public. To our knowledge, there is no single website that is dedicated to viral vaccines and giving a holistic approach to viral diseases and vaccines. Therefore, we have made an attempt to develop ViralVacDB which is a comprehensive repository of viral vaccines covering the vaccine approved and in clinical development against the viral diseases. We hope that the comprehensive information available in the ViralVacDB will certainly be very helpful to the policymakers, researchers and general public.

Vaccine development has been conventionally an experimental science, but in silico approaches are assisting in unravelling the mechanism for vaccination towards the design and development of vaccines against viral diseases. Predicting B-cell epitope is challenging but crucial for many immunological and biomedical applications including disease control, designing diagnostics kits and developing vaccine. Most studies work on the model to predict whether a given surface peptide is linear B-cell epitope or conformational B-cell epitope. In our study, we have also worked the model to predict whether the given peptide is epitope or non-epitope. It brings up the certainty with trained models integrating the composition-based feature, tagging along binary and evolutionary profile along with structural information to classify if the given patch of sequence is B-cell epitopes or not.

Chapter 2

Manual Curation of Virus Based Vaccines

Introduction

Past decades have been marked and characterised by the spread of new emerging viral diseases, increased infectivity, and prevalence of pre-existing pathogens. Viruses are the tiniest creature with the capability to decimate a population. They replicate only within a host cell [2]. They are capable of infecting all life forms, from bacteria, plants to animals and humans. When independent of the host, it consists of a viral genome contained within the capsid. Virus genome or genetic material is very diverse. It can be Deoxyribonucleic acid (DNA) or Ribonucleic acid(RNA), double or single stranded, linear or circular, monopartite or multipartite with their length as short as 2 kb or as long as 2500 kb [3]https://dx.doi.org/10.1007%2F978-981-15-0702-1_1. In some viruses, an additional layer or membrane, called envelope, is present around the protein shell. Viruses use cell host machinery to replicate with the host cell. It infects its host by attaching the host cell and penetrating within the cell. During penetration, protein is uncoated and only genetic material is transferred to the host cell. Its genetic material contains all the information needed to take control and infect host cells. Once inside, it replicates its genome and produces viral proteins to make new capsids[4]. This step is followed by virus assembly whereby the virus genome and capsid protein are assembled to form new viruses. The new viruses then burst out of the host cell devouring the host cell of its normal functionality. During this complete process of replication, some viruses tend to modify their genome by integrating the host genome and enhance their virulence pathogenicity [5][6]. Following its replication, viruses go on infecting new hosts trying to expand their host range [7].

In order to mark their existence in other species, they continuously evolve to adapt to new hosts and environments. They are known molecular manipulators [8] Their genetic diversity and variability is one of the contributing factors behind their spread. This is witnessed in rhinovirus (presence of 160 different serotypes) and dengue (four serotypes). Their continuous antigenic shift and drift as observed in case of influenza virus is another factor that contributes to its increasing virulence. Their high mutation rates, as reported in Human immunodeficiency virus (HIV) and Hepatitis C virus (HCV) is known to add variation to its genome and increases their chances of transmissibility [5]. Besides, mounting evidence suggests the continuous animal-human interaction has led to emergence and spread of zoonotic virus strain. This observation is witnessed in recent years in the case of Severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS) and Severe acute respiratory syndrome-related coronavirus-2 (SARS-CoV-2),

influenza H5N1, and H7N9 viruses [9]. This increased pathogenicity, genomic variability and transmissibility, stand as the cause of major apprehension and threat to humans [10].

To establish infection in humans, different viruses adapt different routes to gain entry in the human body. These include respiratory tract, gastrointestinal tract, genital tract, contact with skin, eyes or faecal. Some viruses are even capable enough to cross the placenta and infect the foetus. Respiratory tract is the most common portal of viral entry and major cause of respiratory infection. These consist of infection caused by adenovirus, influenza, SARS, and SARS-CoV-2 virus. Similarly, viruses following different portals of entry are responsible for causing a wide range of infectious diseases. It includes dengue, chicken pox, AIDS, common cold, and rabies [11] [12]. They are the reason behind many of the emerging infectious diseases and outbreaks. These includes SARS-CoV in 2003–2004, H1N1 based "swine flu" pandemic in 2009, prevalence of MERS-CoV since 2012, Zika virus in 2015–16, Ebola virus in 2013–2016, and the recent outbreak of SARS-CoV-2 2021[10][13][14][15]. As per World Health Organization (WHO) (https://www.who.int/) recent reports, hundreds of million people were affected by SARS-CoV-2 virus strain and approximately five million were recorded deceased Ref.

Over the course of human history, humans have adapted various approaches to control viral infection and prevent outbreaks. It includes avoidance of viral exposure, quarantine, control of vectors and elimination of non-human reservoirs[16] [7]. Also, to date wide range of therapeutic antiviral agents are now available in market, such as drugs against human immunodeficiency virus type 1 (HIV-1) infection and hepatitis C [16]. But there exist certain limiting factors to antiviral drugs. It is effective only when administered within a certain time frame before or after exposure and as long as the drug is being administered. That means it does not provide long-term immunity against the causative agent. Also, their misuse, overuse and antimicrobial resistance development have cost over thousands of lives [17].

On the other hand, vaccination stands as simple, safe, and effective medical intervention in response to infectious viral diseases [18]. They are known to prevent or mitigate infections by inducing a protective immune response in the body against the viruses represented in the vaccine. It develops long term immunological memory against the targeted pathogens and upon exposure, the immune system is able to mount a quick immune response compared to non-primed immune system[19] . A number of factors directly or indirectly influences host immune responses against the subjected viral vaccines. It depends on the type of vaccine (live- attenuated. Killed, vector

based, or recombinant) which in turn affects the magnitude and duration of immunity significantly [20]. Adjuvants such as aluminium salts (e.g. Hepatitis B vaccine) plays an important role in enhancing immunity and effectiveness of the vaccine [21]. Besides, the administration route of a vaccine also influences the immunogenicity of some vaccines [22]. Other important parameters associated with vaccines include, clinical phase of development, approving organisation, dosage, targeted strain etc, which are dispersed across web resources and literature [22]. All these parameters play a significant role in making informed decisions specially for the general public and clinicians. These details are not well documented which cost time and energy to capture the dynamics about different vaccines. This in-depth information when present in a single platform can benefit people working in the scientific community and policymakers to scroll, analyse, store and retrieve data at one go. To the best of our knowledge, there is no single platform that is dedicated to viral vaccines and covers associated information of vaccine and viral diseases extensively. Therefore, in this present we cover different aspects of viral vaccines like their types, clinical phase, administration route, mechanism of action and other details in-depth. The curated information is compiled in the form of a database named "ViralVacDB" and is made freely available to serve the society. We anticipate that this database will be highly beneficial for the researchers and people associated with pharma industry and immunoinformatics for the development of novel vaccine candidates.

Overview of viral vaccines

History

Humans have a long history of infectious diseases and an attempt at experimental trials to provide immunity against deadly diseases led to the discovery of vaccines. Using cowpox pustule inoculant by Edward Jenner, a successful attempt in 1796, to provide protection against smallpox marks the timeline in the history of Immunisation and lead to worldwide spread use of the technique to fight against infectious diseases. Later the up-gradation of technology for vaccine development led to the eradication of smallpox. The next successful attempt was by Louis Pasteur's rabies vaccine in the year 1885 which led to the discovery of bacteriology for the first time to elicit immune responses against the diseases. By 1930, successful attempts were made against diphtheria, tetanus, anthrax, cholera, plague, typhoid, and tuberculosis. The 20th century marks an important phase in

the history of vaccination as during this time period vaccines were created against several viral diseases like polio, measles, mumps, and rubella which otherwise have led to death in the history of mankind [23]. The timeline for vaccine development is represented in Figure 1.

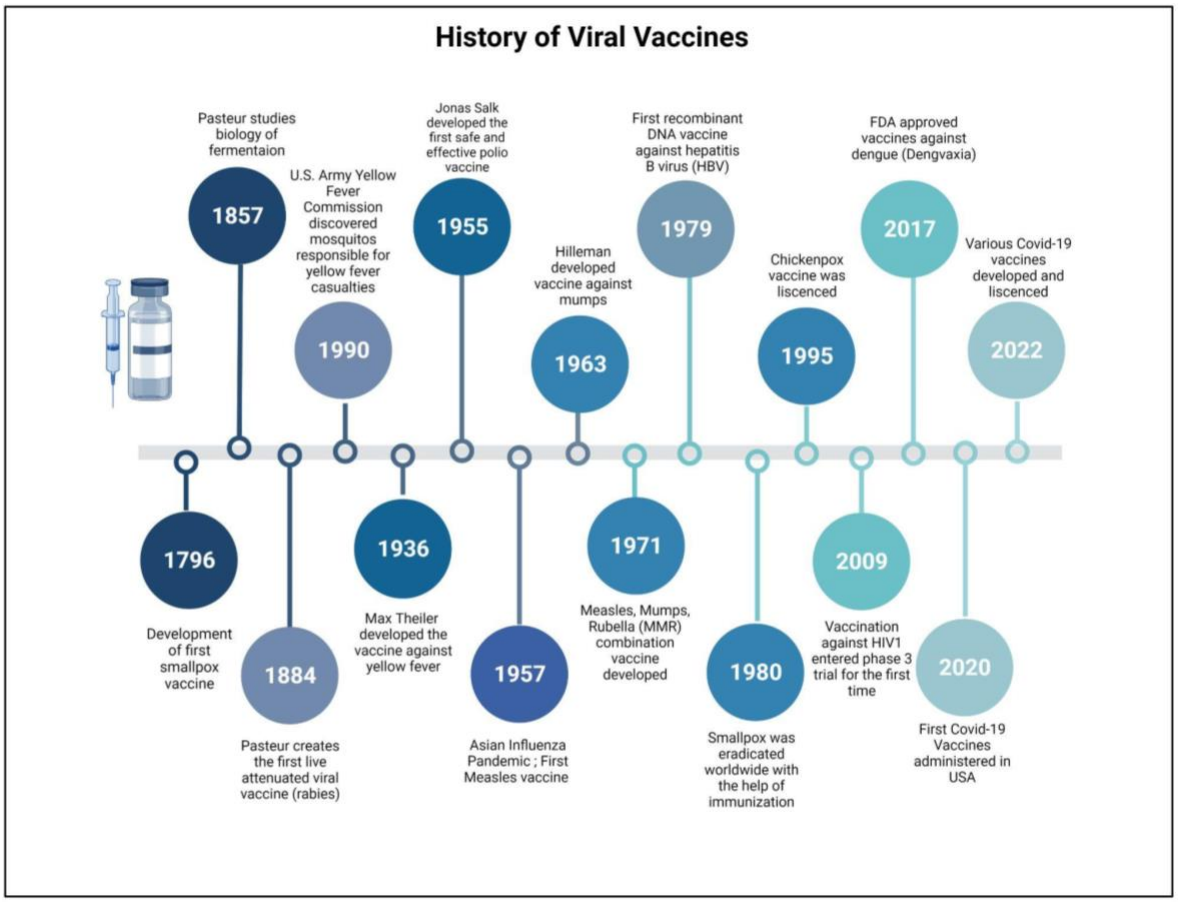

Figure 1: Timeline of viral vaccine development

Mechanism

Vaccines imitate an infection, stimulating the body to produce antibodies and defensive white blood cells in response, in order to facilitate development of host immune response [20]. It elicits an innate immune response which triggers antigen-dependent adaptive immune response. Innate immunity is the antigen independent first line of defence against pathogen entry whereas adaptive immunity is antigen dependent immunity and triggered at later stages. Vaccines induce cell-mediated adaptive immunity by triggering T lymphocytes and B lymphocytes [24]. Long-term immunity is activated by the adaptive immune system against the targeted pathogen as it establishes immunological memory and activates immune response when exposed to that pathogen[20].Covid-19 vaccine candidates Pfizer BNT162b2 are nucleoside-modified RNA

vaccines that targets full-length SARS-CoV-2 spike protein and provokes immunity on its exposure. Figure 2 presents the comparative graphical representation of infectivity induced by virus and induced immunity mechanism of vaccine.

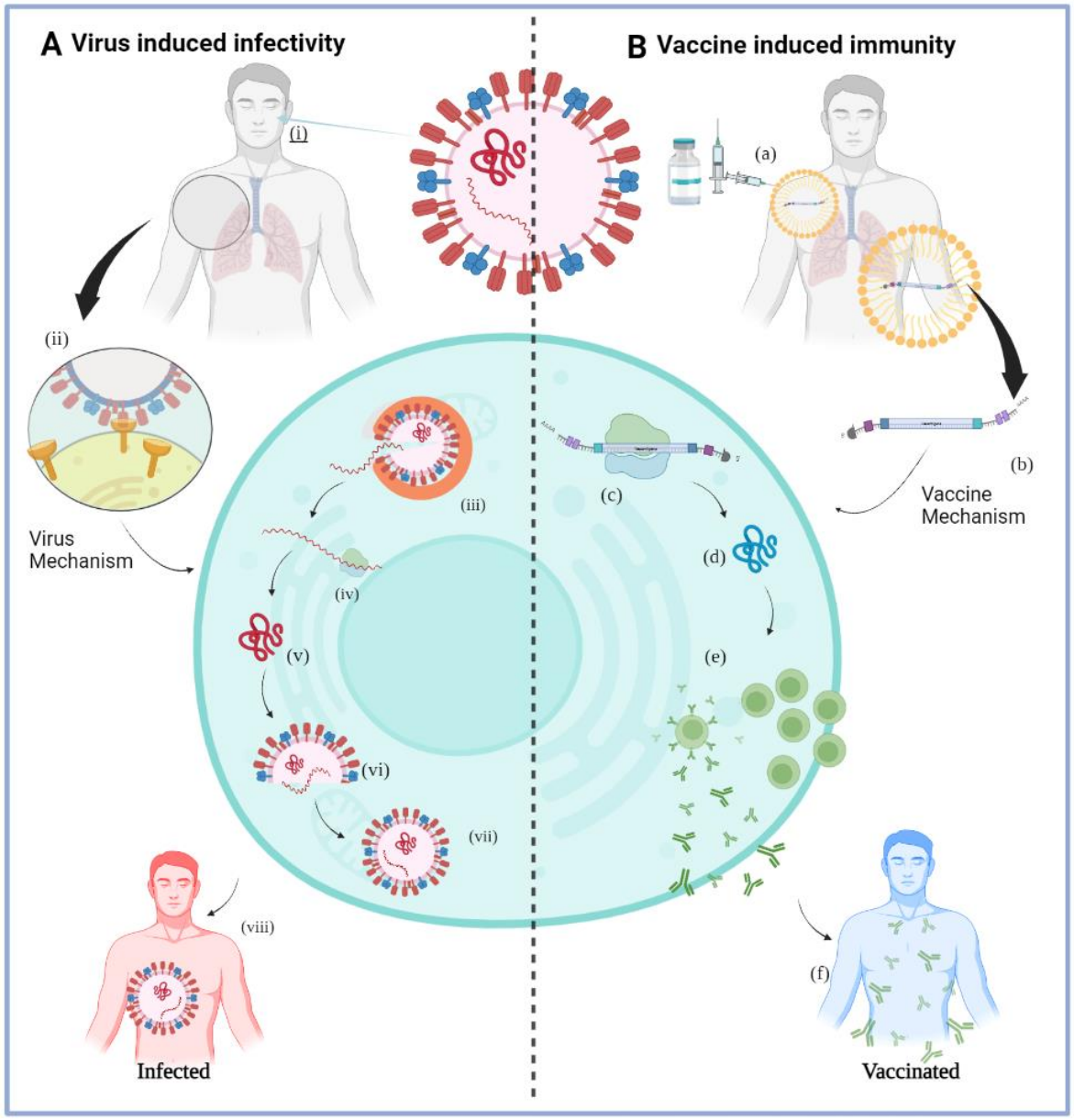

***Figure 2**: Comparative representation of comparative mechanism of action of virus and vaccine. A) Infectivity induced by virus: (i) Virus entry in host (ii) Attachment of virus to host cell (iii) genomic replication (iv) mRNA synthesis (v) Protein synthesis (vi) protein packaging (vii) protein assembly (viii) release. B) Immunity induced by vaccine: (a) vaccine injected (b) antigenic strain enters host cell (c) antigen recognition (d) Immune cell activates and begins protein production (e) antibody generation (f) Immune response*

Types of Vaccines

Pasteur's three Is paradigm i.e., isolate, inactivate, inject had been shifted towards an approach of rational design because of an improved understanding of pathogen-host interaction and mechanisms of the immune system. Targeted delivery of an antigenic material approach with minimalist composition has been adopted by the novel vaccine, revolutionising its dimensions with the benefit of safety, efficacy, and cost-reduction[25]. Targeted strains are either conventionally derived from the pathogen and are used in whole pathogen vaccine. It is produced synthetically which mimic the components of the pathogen as in case of nucleic-acid based vaccines [20]. Vaccine largely depends on the active part i.e., antigenic strain to elicit an immune response. Vaccines based on design and components used are classified as inactivated or live-attenuated or whole pathogens, DNA or RNA based nucleic acid vaccine, protein subunit, or viral vector (replicative or non-replicative) [26]. Conventional vaccines such as live attenuated and inactivated viral vaccines stand as the foundation of vaccine design with their application eradication of smallpox at global level [27]. Table 1 shows schematic representation of different types of vaccine, their advantages and disadvantages with viral vaccines associated with it.

Table 1*:* Representation of different types of vaccine, their advantages and disadvantages with viral vaccines associated with it

Vaccine Type	Subtype	Description	Advantage (Adv) & Disadvantage (Disadv)	Examples	Ref
Whole Pathogen Virus	**Live attenuated vaccine**	Contain live virus but weekend (attenuated) so that it is enough to replicate and activate immune response without causing disease.	Adv: Reflect natural infection; activates strong cell mediated and humoral immune response, lifelong immunity, good efficacy Disadv: Require adjuvant, booster to ensure efficacy and long-term effectiveness, occasionally incomplete inactivation, safety issues	Rubella (MMR)	[28], [29]
	Inactivated or dead vaccine	Contain virus which has been killed or altered such that they could not replicate but they require booster to elicit immune response	Adv: Stable; devoid of replication of inactivated pathogen; Safe for immunocompromised and pregnant ladies Disadv: Require adjuvant, booster to ensure efficacy and long-term effectiveness, occasionally incomplete inactivation, poor efficacy	Polio (IPV)	[24], [30], [29]

Viral vector	**Replicating viral vector vaccine**	Ability to actively invade host cell and replicate alongside delivering the vaccine antigen	Adv: Can be developed quickly and easily on large scale, cheaper; elicit response with one dose Disadv: Past exposure to the vector reduces effectiveness, chances of mild adverse side effects, relatively complex to manufacture	Ebola (Ervebo)	[31]
	Non-Replicating viral vector vaccine	Genetically engineered Replication-defective viral vector i.e. vectors are designed such that it can induce host immune responses but are not capable to perform replication inside the host cells	Adv: Non-infectious, reduced risk of adverse response Disadv: Effective as long as it is in host cell, require booster doses	Covid-19(Oxford - AstraZeneca)	[31]
Subunit	**Protein vaccine**	Protein fragments are extracted from surface and inserted into manufacturing cell	Adv: Non-infectious, Strong humoral response Disadv: Not capable enough to activate innate immunity, reduced immunogenicity, adjuvants are required	Covid-19(MVC-COV1901)	[24], [32]
	Recombinant Protein vaccine	Protein antigens that are produced in heterologous expression systems and capable to generate immunological response against incorporated antigenic proteins	Adv: Non-infectious, Strong humoral response Disadv: More Demanding manufacturing process, reduced immunogenicity, adjuvants are required	Human papilloma virus (HPV) (Gardasil)	[29], [33]
	Virus like particles (VLPs) vaccine	Virus derived multiprotein structure are capable of self-assembly, copying the organization and conformation of a virus particle but lack of genetic material doesn't instigate host infection	Adv: Non-infectious, can be manipulated to express multiple antigen; can also be used as adjuvants Disadv: Induction of anti-vector immunity, less stable	Dengue (MWNT-DENV3E)	[34], [35], [36]
Nucleic Acid	**DNA vaccine**	Antigen-encoding DNA plasmid induce an immunologic response	Adv: Non-infectious, egg and cell free, cost effective, mimic natural infection, activate both cell and humoral immunity, good stability Disadv: Could integrate with human genome and affect host cellular function	Covid-19 Covigenix VAX-001	[30], [34], [37]

			like cell growth, poor efficacy		
	RNA vaccine	RNA vaccine consists of an mRNA strand that codes for a disease-specific antigen	Adv: Non-infectious, produce sufficient antigen to stimulate an immune response, natural degradation, rapid and scalable production Disadv: Less stable, low immunogenicity	Covid-19 (Moderna)	[29] [34]

Administration of vaccine

Conventional route of administration for viral vaccines includes needle-mediated injections into the muscle (Intramuscular) and under or into the skin (Subcutaneous) or Intradermal respectively) [38]. With better understanding of immunological responses, mechanisms, design and development have led to the evolution of vaccine design adding several new categories like intranasal, oral, intravenous as another category for administering routes for vaccines. Intranasal influenza vaccine (FluMist, Fluenz), mimics the natural route of its causative agent that cause influenza infection and triggering an immune response directly in the respiratory tract[24].

Approving organisations

Each novel vaccine candidate goes through a detailed design and development process post discovery before being licensed for human use. Process checks for the safety, immunogenicity, and protective efficacy. To plan the developmental stages, regulatory agencies like European Medicines Agency (EMA) [https://www.ema.europa.eu/en], WHO, and the United States Food and Drug Administration (FDA)[https://www.fda.gov/] have issued guidelines and divided this development process is grouped into various categories with their defining parameters. Such as preclinical involves testing in animals and clinical involves clinical trials in human subjects stages [39]. The clinical steps are further categorised into clinical phase 1, phase 2, phase 3 and phase 4. There are various factors that are considered during their development such as type of vaccine (live/killed/subunit/DNA/peptide), administration route, dosage type, quantity, target population, disease epidemiology, and the availability of a pre-existing vaccine [40].

Material and Methods

Data collection

To collate the information on vaccines based against viral disease, a list of viral diseases was curated and queries were submitted in PubMed with keywords “viral vaccine”, “list of vaccines for Covid-19”,” Hepatitis A vaccines”,” viral vaccines”, “viral hepatitis vaccinations”,” Covid-19 vaccines in India”,” list of viral-based vaccines”. To search the approved list of viral vaccines, we also explored websites by FDA, WHO, Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/),Violinnet (https://www.violinet.org/). To further add the information of ongoing interventional studies against viral-based vaccines, we screened the clinicaltrial.gov platform.

After screening, information was extracted from websites, research articles, product insert sheets, and information sheets. The information incorporates different vaccine types, administration route, target strain, clinical phase, vaccine approval status, dosage, manufacturer information, viral nucleic acid content, age restriction, dosage type and adjuvant associated with viral vaccines. This extensive information is systematically catalogued in a tabulated manner. Consequently, 421 entries were collated in the ViralVacDB database from reliable information sources and consolidated in tabular form against 24 viral diseases.

Database architecture and web interface

The ViralVacDB database has been built using a standard platform based on the cross-platform Linux, Apache, MySQL and PHP (LAMP). MySQL (version 14.12) was used as the back-end for managing the data and Apache (version 2.2.17) as the HTTP server was used for designing the framework of this database. The front-end web interface was developed template responsive using bootstrap, a popular responsive development framework that includes HTML, CSS, and JavaScript. PHP programming languages were used to develop a common interface. Different functionality was integrated into ViralVacDB for compilation of data, retrieval of information, and data exploration with an attempt to make it compatible for mobiles, tablets and desktops use. The complete architecture depicting the information and tools embedded into ViralVacDB is represented in Figure 3.

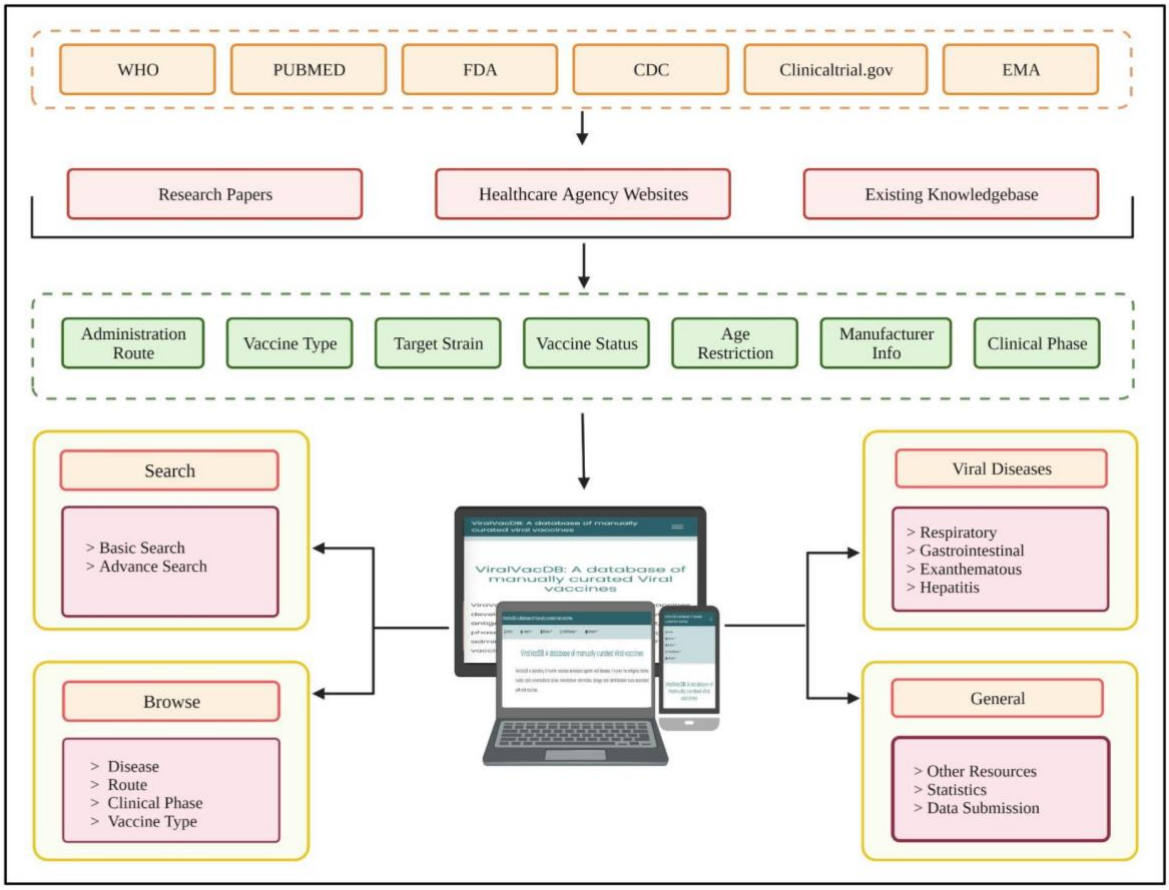

Figure 3: Schematic representation of ViralVacDB architecture and its modules

Organisation of database

Search Modules

Basic Search: The users are able to do query-based search within the database against the field such as disease name, vaccine name, vaccine type, administration route, clinical phase and clinical status. The interface allows users with the facility of a simple search tool to customise their search on the basis of desired category in a simplistic and time-efficient manner.

Advance Search: This facility allows users to perform advanced search by allowing users to do more complex searches on a wide range of fields available**.** Users can make different boolean operators such as AND, OR and NOT on different queries simultaneously and receive desired output accordingly. Users can retrieve, copy and download in the form of csv, excel. Each vaccine

has been organised into individual panels which store comprehensive “vaccine-cards” associated with its vaccine Id.

Browse Module

A simple and time-efficient browsing facility is facilitated on the ViralVacDB website to retrieve information from the database in a simple and effortless manner. Users can browse data on the basis of fields such as - 1) Name of disease 2) Administration route 3) Vaccine type 4) Classified disease category 5) Clinical phase 6) Vaccine status. This allows users to browse o different categories and retrieve information even if not familiar with desired information.

Viral Disease

This module consolidates information on viral disease and associated pathogens. Diseases are grouped into eight major viral diseases. Users can simply browse these categories which include- 1) Respiratory diseases 2) Gastrointestinal diseases 3) Exanthematous diseases 4) Hepatitis diseases 5) Cutaneous diseases 6) Haemorrhagic diseases 7) Neurologic Diseases and 8) Others. These categories are further sub categorised into different viral diseases and displayed in the form of cards. On clicking on the detail button on any of the diseases, a card will pop-open which covers information about disease and its causative agent comprehensively. It clubs’ information about disease symptoms, zoonotic evidence, associate virus structure, transmission, family, genus, proteomic, genomic profiling beside antigenic strains, and targeted host organisms. Genome and proteome reference link are also added which directs users to associated NCBI and UniProt information pages on single click.

Database statistics

ViralVacDB is a user-friendly web-based dedicated repository for viral vaccines. Information extracted from reliable resources and articles has been curated and catalogued in 24 different fields. The data is maintained in tabular format in which each record is comprehensively presented in the form of a card. The database holds 421 total entries with relevant vaccine information, of which 139 are licensed and are in use in different countries.

On analysis of data, it is noticed that of a total 421 vaccines covered in ViralVacDB, 87 records belong to inactivated vaccine category, 60 to live-attenuated, 46 to protein subunit, 57 belongs to

the nucleic acid type category (24, 17, 16 associated with DNA, RNA and mRNA category respectively) and similarly 106 associated with viral vector category which is further classified into recombinant (45), virus-like particle (350), replicating (8) and non-replicating vaccine (18) category.

As shown in the pie chart (Figure 4), major vaccine types belong to the whole pathogen vaccines (39.3%), followed by subunit vaccines (25.5%), and viral vectored vaccines (18.8%). Recent advancement shows the development of nucleic acid-based vaccines, covering 16.2 % of vaccines.

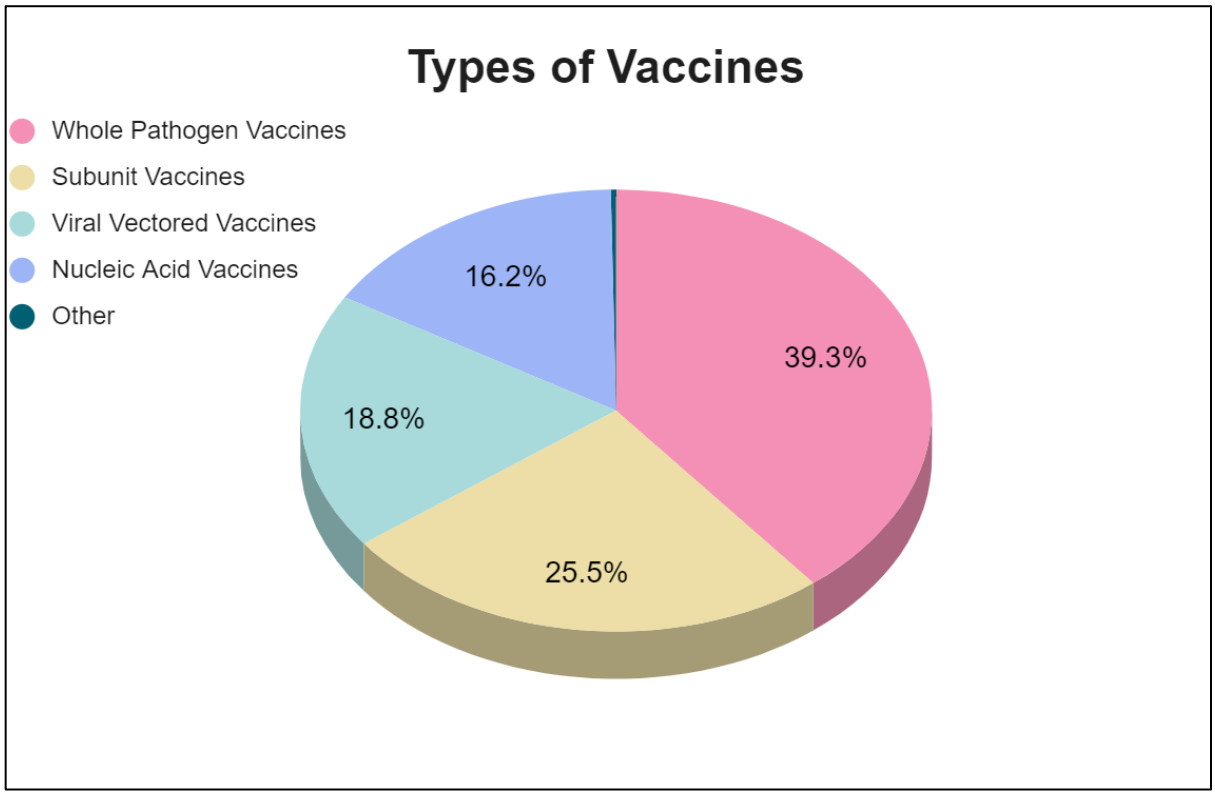

Figure 4: A schematic representation of distribution of viral vaccines based on types of vaccines.

Each vaccine is annotated with the administration route with 232 vaccines belonging to Intramuscular, 37 to Subcutaneous, 16 intranasal, 14 oral, 11 intradermal and 3 intravenous. Figure 5 shows distribution of the vaccine in different phases of development.

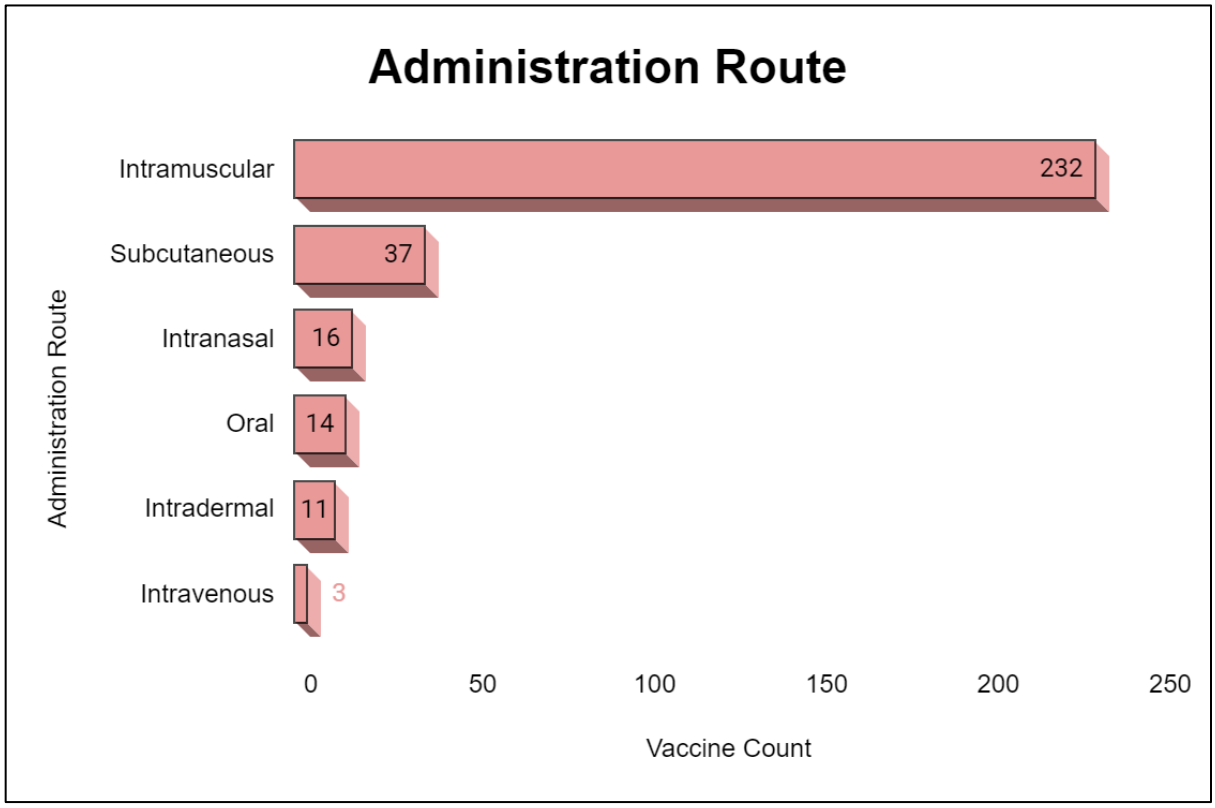

Figure 5: A schematic representation of distribution of viral vaccines based on administration route.

Information on the developmental stage and licensing are covered under the clinical phase and vaccine status category. On analysis, it is noticed that 42 vaccines are in the pre-clinical phase, 96 in phase 1, 59 in phase 2, 161 in phase 3 and 33 in phase 4. Vaccines in transition state- phase 1/ phase 2 or phase 2/ phase 3 are also represented in analysis with their counts summing to 11 and 6 respectively. A schematic representation of distribution of viral vaccines based on administration route is represented in Figure 6.

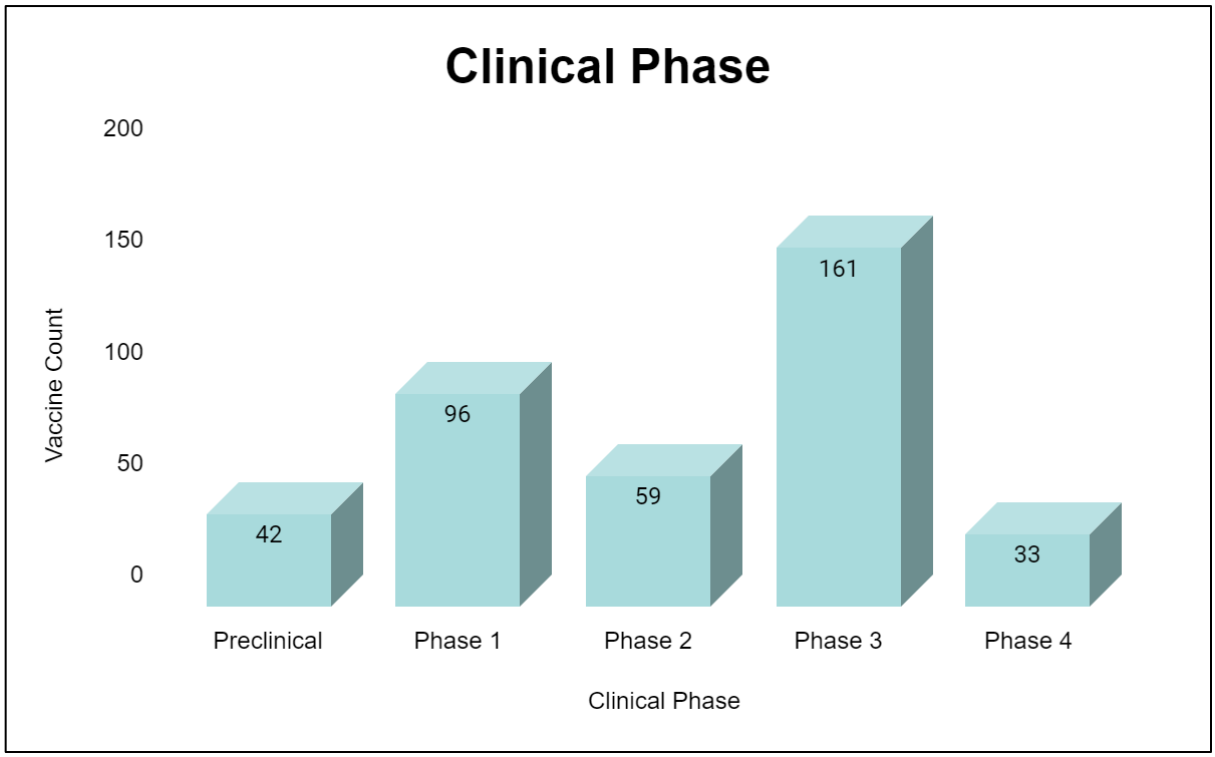

Figure 6: A schematic representation of distribution of viral vaccines based on Clinical Phase

Viral diseases are grouped into different categories depending on the mode of transmission and region affected. On that note, seven categories are presented namely respiratory with 246 vaccines, exanthematous with 77 vaccines, hepatitis with 28 vaccines, haemorrhagic with 23 vaccines, neurologic with 21 vaccines, gastrointestinal with 17 vaccines and cutaneous with 8 vaccines. This vaccine count on the basis of vaccine type, clinical phase, administration route, and classification is represented in Figure 7.

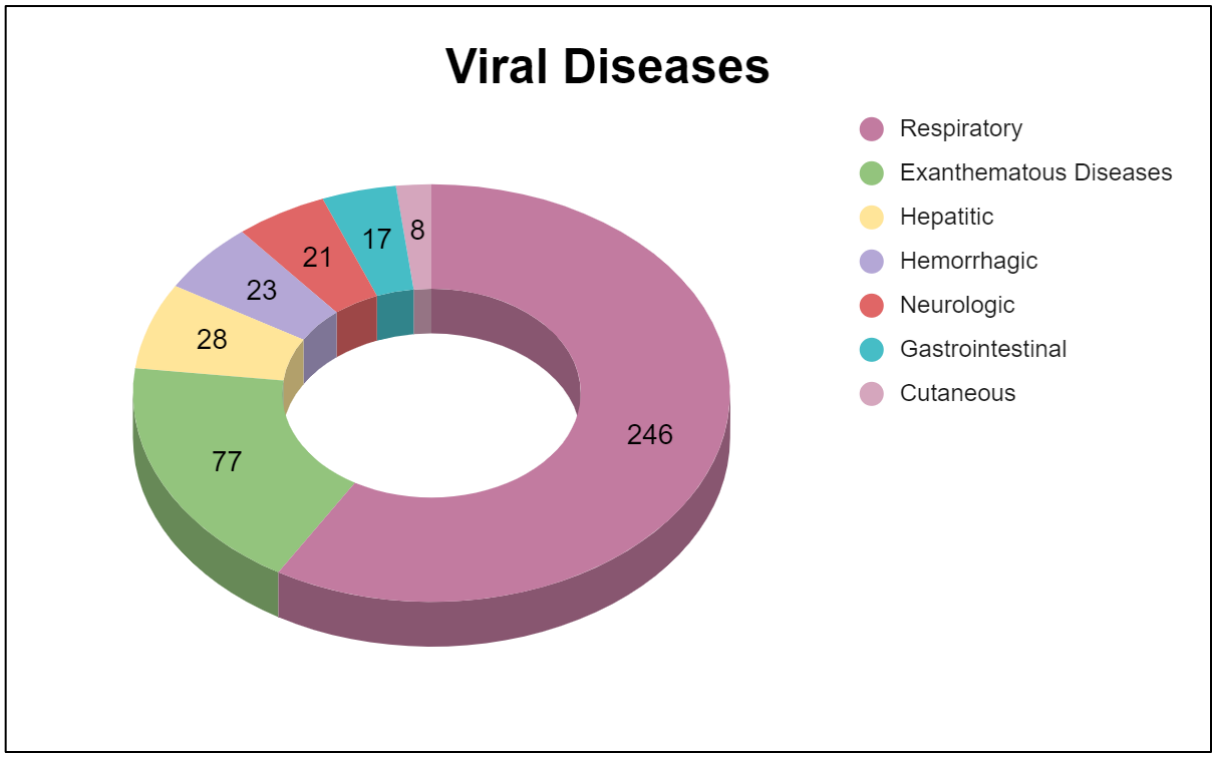

Figure 7: A schematic representation of distribution of viral vaccines based on administration route

To add coverage on viral diseases, information on various viral diseases and the responsible pathogen is well clubbed and presented under the separate tab of ViralVacDB. It covers information about zoonotic evidence, host, instances of occurrence, region affected globally besides adding supplementary details on transmission route, susceptible age group, symptoms, and co-infectivity associated with viral diseases. To present a better understanding of causal pathogens associated with diseases, information covering its family, genus, structural, genomic, proteome details with NCBI and UniProt link attached as genome and proteome reference is well-curated.

Comparison with existing databases

Databases associated with vaccines have evolved to compile the large volume of data related vaccines and integrate them into a single platform. ViralVacDB is the sole repository dedicated to viral vaccines for human use. It is developed to store, organise, and integrate the vaccine data that is scattered across diverse resources. It exclusively compiles 421 human vaccine records against 24 viral diseases which is an extensive vaccine record in itself for viral diseases so far. The existing database, Huvax (http://www.violinet.org), part of the VIOLIN database, focuses mainly on human vaccines licensed in the USA and Canada. In comparison to this database, ViralVacDB covers other countries as well and supports additional information like vaccine strains, adjuvants, dosage, country, year, and organisation of manufacturer along with information on licensing bodies. All this data is presented in a simple and user-friendly manner so that it is easy for users to understand, analyse and compare the immunological dynamics of different viral vaccines in a single platform. Besides, coverage of viral diseases is an added advantage to this database. There are existing knowledge bases like Vaccine Knowledge Project and The Immunisation Advisory Center (https://www.immune.org.nz/) which are limited to licensed vaccines in the UK and New Zealand respectively. They do not provide the vast fields associated with each vaccine record as covered in ViralVacDB. In comparison to these databases, ViralVacDB annotates massive information against each viral vaccine in a simplistic, easy-to-understand manner.

Utility of database

The Covid-19 pandemic has drawn attention to a bigger health challenge-one which is a comprehensive repository for viral diseases and information on vaccines associated with viral diseases. ViralVacDB is a database that covers information on viral vaccines approved and in the pipeline. Data available from research articles, verified websites are well-curated and integrated to provide information in a simpler form, with easy availability and accessibility. This initiative will allow the scientific community and general public to understand and search vaccines associated with different viral diseases and filter on the basis of age group, manufacturer, and demographics.

The major advantage is that VialVacDB provides a single platform to let users gather information on the list of approved vaccines by different recognized agencies like FDA, European Medicines

Agency (EMA) (https://www.ema.europa.eu/en), Defence Research and Development Organisation (DRDO) (https://www.drdo.gov.in/), South African Health Products Regulatory Authority (SAHPRA) (https://www.sahpra.org.za/). It also allows users with numerous services like (i) users to browse vaccine information by categories like viral disease, vaccine type, administration route, clinical phase, and vaccine status. (ii) users can retrieve information by selecting the information type required or can go for advanced search which is designed to facilitate users to search using multiple queries joined by logical operators like AND/OR/NOT. (iii) ViralVacDB allows users to sort and filter and download data as per the information required. For instance, users can easily download the selected data in the form csv. (iv) ViralVacDB covers basic facts including genomic, proteomic and epidemiological information on different types of viral diseases. Users can easily explore and retrieve information for the same.

ViralVacDB also provides a direct link associated with vaccine information. For instance, users can directly refer to the product information page of the vaccine. Users with a single click get the genomic and proteomic information of viral diseases. It will be very helpful in assisting academicians, researchers, and the general public. We hope that it will accomplish our main aim of "academic enhancement" and "knowledge enrichment" on a single platform.

Discussion

A virus is capable of causing mild infection to severe infections in humans with the potential of bringing about epidemics and pandemics. The occurrence of novel humans and the re-emergence of more virulent infectious pathogens such as observed in chikungunya and Zika, stands the major concern of the decade. Estimates indicate that approximately sixty percent of infectious diseases and seventy percent of emerging infections of humans are zoonotic in origin. [12].To add, the recent SARS-CoV-2 outbreak has generated much fear and sustained commercial, travel, and ecological disruption to systems at the global level [41].

Developing better ways to manage and anticipate the present ongoing scenario is critical to achieving the United Nations Sustainable Development Goals [42]. Vaccines have been proven to be safe and cost-effective healthcare interventions in humans since the discovery of the first vaccine against smallpox in the year 1798 [43]. The accelerated development of multiple vaccines during the Covid-19 pandemic is unprecedented, the process usually takes eight to fifteen years. The swift development of vaccines urges the need for a vaccine repository giving a complete

package to the understanding and knowledge base for the users of the approved and in pipeline vaccines in a lucid manner. Our database records comprehensively reflect human vaccines against twenty-four viral causative agents. This database will serve as a repository for human vaccines targeted against viral diseases which are licensed and are in use globally.

Plethora knowledge surrounding immune response and mechanism provides insights into the different vaccine development platforms. Vaccines containing weekend replicating strains of the targeted pathogen are live attenuated and include vaccines against measles, mumps, Covid-19, rubella, influenza, chikungunya, and yellow fever while vaccines containing the killed whole organism are termed inactivated and include vaccines against hepatitis A and Rabies. In addition to that, additional platforms are covered in ViralVacDB like nucleic acid and viral-vector based. For instance, INO-4800 is a DNA based vaccine targeted against the SARS-CoV-2 virus.

There exist various ways to administer these vaccines which include trivial subcutaneous or intramuscular injections to recent methods of intranasal and oral, which are designed to imitate the natural course of causative agents. For instance, vaccines like NanoVax developed against Covid-19 follow an intranasal route to mimic the natural route of respiratory infection and elicit an immune response against it. It is also an established fact that intradermal vaccination has been shown to be more immunogenic at much lower doses as compared to intramuscular vaccination in case of influenza, rabies, and HBV vaccines [20]. Similarly, vaccine dosage is another deciding factor that influences immune response. It is established during the early phases of clinical development based on target strain, disease, age group. For older adults, a higher dose is recommended as analysed in the case of the influenza vaccine to develop optimal safety and immunogenicity. Comparatively, for younger infants, 3–4 doses are administered by 12 months of age to build immunity [22].

There are several other aspects that are associated with vaccines such as target strains, adjuvant, clinical phase, manufacturer information, approving organisation, licensing countries, and clinical status. Such information is extensive and scattered across different platforms. To cover several aspects of viral vaccines, we developed ViralVacDB, which stores the information of 421 vaccines against pathogenic viral diseases. It is an attempt to provide comprehensive information on a single platform, which is freely available, and easy to search, scroll, compare and retrieve data in a time-efficient manner. We hope that it will accomplish our main aim of "academic enhancement" and "knowledge enrichment" of human-based viral vaccines on a single platform. This initiative will

allow the scientific community and general public to understand and search vaccines associated with different viral diseases and filter on the basis of age group, manufacturer, and demographics. We believe that this data will be helpful to the general public, clinicians, scientific community, and policymakers in analysing various viral vaccines and in developing an understanding of existing and emerging viral diseases in great detail and in a user-friendly manner.

Chapter 3

Prediction of B-cell Epitope using Machine Learning Techniques

Introduction

The immune system is a collective network of cells, and associated processes that provide protection to human body from foreign invaders, such as viruses, bacteria [44]. The innate immune system and the adaptive immune system are subsystems of the immune system. The innate immune system is the first line of defence against infection and does not retain immunological memory against infectious pathogens. It is a nonspecific defence mechanism that immediately induced post exposure of pathogen. Innate immunity consists of physical barriers such as the skin, cellular processes and humoral components. The adaptive immune system is triggered if a pathogen persists.

B-cells are an essential part of the adaptive immune system, as they are capable of recognizing and encountering foreign antigen with ability to provide long-term protection against harmful pathogens by generating immunological memory. The key components in this process are their immunoglobulins or antibodies. Antibodies recognize their targets through interactions between their binding site also known as paratope and a complementary region of the antigen known as epitope that is specific to paratope [49]. B-cell epitopes are grouped into two main categories- Linear (Continuous) and Conformational (Discontinuous). Linear B-cell epitopes are characterised by sequential peptides, whereas conformational B-cell epitopes discontinuous peptide that are arranged in closed spatial proximity by the protein folding. Ninety percent of B-cell epitopes are discontinuous in sequence [50] .

Recognition of B-cell epitopes is of high importance for immunological and biomedical applications including understanding disease aetiology, disease control, diagnostics, prevention and immune monitoring. It holds special importance in vaccine development in terms of designing different vaccines such as such as attenuated and subunit vaccines to provide a long-term protection toward desired pathogens [49]. B-cell epitope are identified using several experimental methods that including solving the three-dimensional structure of Ag-Ab complexes, peptide library screening of antibody binding region. The other methods include evaluation of functional assays in mutated antigens and study the interaction of antibody-antigen is evaluated using ELISA, protein crystallography and peptide-chip. These experimental techniques requires time, energy and is expensive and gives low-throughput, or have low accuracy [51].

To assist or substitute the experimental approaches, various in-silico computational methods have been developed for the prediction of linear and conformational B-cell epitope. Linear B-cell

epitopes are predicted using a primary sequence of peptides using a sequence-based method. Computational methods for the prediction of B-cell epitope were conventionally based on physicochemical features which includes hydrophilicity calculations, flexibility, surface accessibility, and β-turn propensity with inclusion of ML based method for improved performance. Currently available tools include PREDITOP [52], PEOPLE [53], BepiPred [54], ABCpred [55], LBtope [56] , BCPREDS [57] and SVMtrip [50] [58].

Although most the B-cell epitope is conformational yet their prediction is lagging behind linear B-cell epitope prediction. The reasons include prior knowledge of protein three-dimensional (3D) structure which is available for fraction of protein and isolating conformational B-cell epitopes from their protein structure is in itself a difficult task that requires suitable scaffolds for epitope grafting. So far, several attempts have been made to predict conformational method which includes predicting patches of solvent-exposed residues (CEP), amino acid statistics and spatial information (DiscoTope) , followed with inclusion of structure based method, mimotope, machine learning models and statistics based approach as used in MIMOPRO , EpiSearch , PepiTope , Mimox, PEACE , EpiPred , EPSVR , Epitopia , SEPPA , PEPITO , ElliPro , CBtope , and ABCpred [59]. In this study, we have made an attempt to classify the linear as well as conformational B-cell epitopes using various machine learning techniques.

Material and Methods

Dataset compilation and pre-processing

It is of high importance to select the right dataset for developing a prediction method. Evaluation matrix check for performance largely dependent on the dataset used for training model. In this study epitope data for linear and conformational B-cell epitopes was acquired from Kaggle based competition https://www.kaggle.com/competitions/epitope which was launched in January 2021 with the objective of classifying conformational and linear epitopes. Dataset was available with 6394 training data and 1598 validation data.

Pre-processing of data enhances quality of data and reduces redundancy. Removing duplicates and sequences with unnatural amino-acids resulted in 6307 epitopes. The whole dataset was further divided into 80% training as well as 20% testing dataset for internal validation. Various machine learning models has been developed to compute the performance and were finally evaluated on unseen validation dataset for external validation.

Protein features

We implemented Pfeature and Protlearn to generate various types of descriptors using sequential information of the peptides. The complete list of descriptors generated is provided in Table 2.

Composition-based features

We computed five types of composition-based features like amino acid composition, dipeptide composition, tripeptide composition, atomic composition, atom and bonds that are associated with peptide sequences.

Amino Acid based Composition (AAC)

It computes the frequency of each type of amino acid residue present in a protein sequence. The compositions of all 20 natural amino acids were calculated using the following equation 1:

$$AACi = Ri / L \quad (1)$$

where $AACi$ is the amino acid composition of residue type i; Ri and L number of residues of type i and length of sequence, respectively.

Dipeptide based Composition (DPC)

Amino acid composition provides only a frequency of different amino acid residues present, but lacks information regarding their order. Dipeptide composition is used to cover the global information associated with each sequence, which results in output of a fixed length vector of 400 (20*20). Dipeptide is made of two consecutive residues i.e., residue i and i+1. DPC is computed using following equation

$$DPC_j^{\,i} = D_j^{\,i} / (L - j) \quad (2)$$

Where DPC_J^i is the fraction of dipeptide of type i for jth order. D_J^i and L are the number of dipeptides of type i and length of a protein sequence, respectively.

Tripeptide based Composition (TPC)

In tripeptide composition, three consecutive amino acids are taken into consideration which provide information of arrangement order in addition to simple composition. Both previous and proceeding amino acid residues are used to form a tripeptide. There are total 8000 (20*20*20) possible tripeptides from 20 different types of natural amino acid residue.

$$TPC_i = T_i/(L - 2) \quad (3)$$

Where, TPC_i is tripeptide composition of tripeptide i, out of possible 8000 tripeptides. Ti and L are the number of tripeptides of type i and length of a protein sequence, respectively.

Atoms and Bonds

All amino acids are composed of atoms and bonds that bind them together. We have computed atom and bond composition present in each peptide sequence. Fractions of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms present in a protein sequence are captured by atomic composition whereas bond composition captures four bonds including aromatic, hydrogen bond, single bond and double bond.

$$ATC_i = A_i/N \quad (4)$$

$$BTC_i = B_i/N \quad (5)$$

Where, ATC_i is the atomic composition of atoms of type i, A_i is the number of atoms of type i and N is the number of atoms. Similarly, in BTC_i is bond composition for bonds of type i, Bi is number of atoms of type i and N is number of atoms in a protein sequence, respectively.

Physio-Chemical Properties

Amino Acid Index (AAIndex)

It is a set of 20 numerical values that represents various physiochemical and biochemical properties associated with amino acid residues. Following equation 6 is used for calculation of the amino acid index of each residue.

$$AAIC_i = AAI_i/L \quad (6)$$

Where, $AAIC_i$ is AA index composition of residue type i; AAIi and L are sum of AA index value of type i and length of sequence, respectively.

Shannon Entropy (SE)

Shannon entropy for a peptide sequence can be computed by equation 7.

$$H(X) = -\sum^{20}_{i=1} p_i log2 p_i \quad (7)$$

Where, i is the amino acid in the sequence (i=A, C, D, …, Y) and X is any protein/peptide sequence.

Structural Properties

It computes composition of advanced properties like secondary structure and surface accessibility, isoelectric point, instability index, gravy, aromaticity of a protein sequence

Binary Profiles (BP)

It computes a binary profile pattern in terms of 1 and 0 depending on presence and absence of each amino acid sequence. The presence is denoted by 1 whereas its absence is denoted by 0.

Evolutionary information-based features

Evolutionary information covers more information compared to a single sequence and it is computed in the form of a position specific scoring matrix (PSSM-400) wherein it generates a 20 × 20 matrix. It captures the arrangement and the order of occurrences of each type of 20 amino acids with respect to all the other amino acid residues present in the given protein sequence.

Miscellaneous

Other descriptors that we include in our study includes Conjoint Triad Descriptors (CTD), Composition Transition and Distribution, Pseudo Amino Acid Composition (PAAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), Quasi Sequence Order (QSO) and Sequence Order Coupling Number (SOCN). The complete list of descriptors generated is depicted in Table 2.

Table 2*:* List of various features along with their feature vector

Descriptors	Feature Vector
Amino Acid Composition (AAC)	20
Dipeptide composition (DPC)	400
Tripeptide composition (TPC)	8000
Atomic composition (ATC)	5
Bond composition (BTC)	3
Binary Profiles (BP)	500
Amino Acid Index (AAI)	553
Physio-Chemical Properties (PCP)	30
Distance Distribution of Residue (DDOR)	20
Conjoint triad descriptors (CTD)	343
Composition/Transition/Distribution - Composition	39
Composition/Transition/Distribution - Transition	39
Composition/Transition/Distribution - Distribution	195
Pseudo amino acid composition (PAAC)	23
Amphiphilic pseudo amino acid composition (APAAC)	26
Sequence-order-coupling number Grantham (SOCN)	3
Quasi-sequence-order (QSO)	23

Five-fold cross validation

In order to train machine learning and measure performance of our models on the training dataset, we applied in our study standard five-fold cross-validation technique. In five-fold cross validation, training data is divided into five folds wherein the training dataset was formed by combining four sets, while the corresponding test set contains the remaining one set. To make sure that the combination is used as a test set only once, five times repetition of the method is done. These

training and testing sets were then used for developing models [60]. To evaluate our model, we predicted label score in our validation dataset by submitting our predicted validation score in Kaggle.

ML-based classifiers

To develop a prediction model for classifying conformational and linear B-cell epitope we have used various classification models such as Random Forest (RF), eXtreme Gradient Boosting (XGB), support vector classifier (SVC), K-nearest neighbour (KNN), Gaussian Naive Bayes (GNB), logistic regression (LR) and Decision tree (DT). To build these machine learning prediction models, we made use of the Scikit-learn package of Python. Parameters were hyper tuned using GridSearchCV which resulted in optimization of hyper-parameters. Protein features were used for training and testing in classification models followed by five-fold cross-validation for the evaluating performance of the models [61][62]. The complete architecture is shown in Figure 8.

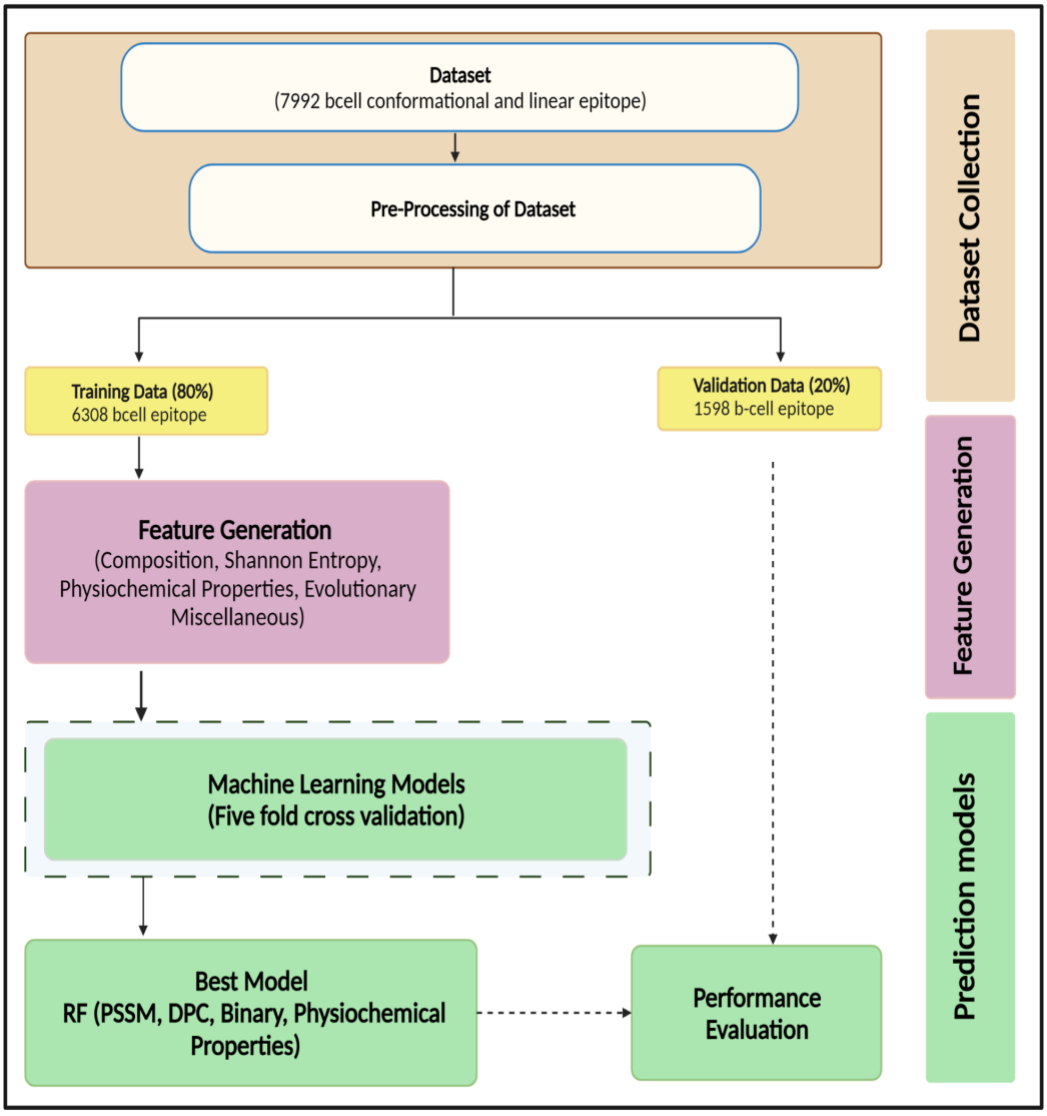

Figure 8: Schematic architecture of Model for prediction of Conformational and Linear B-cell epitopes

Performance evaluation parameters

To measure the performance of machine learning models' threshold-dependent parameters such as sensitivity (Sens), specificity (Spec), accuracy (Acc), Matthew's correlation coefficient (MCC) and area under receiver operating characteristic curve (AUC) were used. 'Sens' also known as the true positive rate (TPR) tells the ratio of true positives and total positives, whereas 'Spec' captures true negative rate (TNR). 'Acc' evaluated total correct predictions of both positive and negative values with respect to total positives and negatives whereas MCC calculates correlation coefficient between predicted and actual labels. The following standard formulae were used to calculate these parameters:

$$Sensitivity = \frac{T_P}{T_P + F_N} \qquad (i)$$

$$Specificity = \frac{T_N}{T_N + F_P} \qquad (ii)$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \qquad (iii)$$

$$MCC = \frac{(T_P * T_N) - (F_P * F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \qquad (iv)$$

Where, T_P, T_N, F_P and F_N stand for true positive, true negative, false positive and false negative, respectively.

Results

Performance of various ML models on different sets of features

Various machine learning models were generated using different set of computed features. The models were trained on the training dataset and then evaluated on testing dataset for internal validation. The final performance of the models is evaluated on unseen validation dataset.

Composition-based features

Five types of composition-based features like amino acid composition, dipeptide composition, tripeptide composition, atom and bonds composition were generated on training and testing dataset. These features were further used to develop ML-based model using different classifiers. These models were optimized by tuning the parameters at different threshold. The performance score of various ML models on composition-based features is tabulated in Table 3. It has been observed that RF-based model has performed better for AAC, DPC and TPC attaining the AUC of 0.81, 0.82 and 0.81 on testing dataset, respectively, whereas XGB-based model performed better for atoms and bonds composition as compared to other classifiers.

Table 3: The performance of machine learning models developed using composition-based features

Amino Acid Composition										
	Training					**Testing**				
Classifier	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	64.40	63.71	64.05	0.69	0.28	60.45	65.63	63.07	0.68	0.26
RF	**71.64**	**72.43**	**72.04**	**0.80**	**0.44**	**72.51**	**73.28**	**72.90**	**0.81**	**0.46**
LR	64.24	65.86	65.06	0.72	0.30	64.47	66.88	65.69	0.72	0.31
XGB	70.87	70.79	70.83	0.78	0.42	70.26	73.13	71.71	0.79	0.43
KNN	67.06	69.30	68.19	0.76	0.36	68.81	67.81	68.30	0.76	0.37
GNB	64.97	64.72	64.84	0.70	0.30	64.95	65.31	65.14	0.70	0.30
SVC	65.17	64.26	64.71	0.71	0.29	65.27	65.47	65.37	0.71	0.31
Dipeptide Composition										
DT	63.12	63.24	63.18	0.65	0.26	61.09	66.41	63.79	0.66	0.28
RF	**72.36**	**73.05**	**72.71**	**0.80**	**0.45**	**75.56**	**74.84**	**75.20**	**0.82**	**0.50**
LR	66.13	64.45	65.28	0.72	0.31	65.60	65.94	65.77	0.73	0.32
XGB	70.47	70.63	70.55	0.77	0.41	71.06	72.34	71.71	0.79	0.43
KNN	69.14	67.07	68.09	0.75	0.36	75.40	69.69	72.50	0.79	0.45
GNB	65.05	65.12	65.08	0.69	0.30	64.15	67.34	65.77	0.71	0.32
SVC	65.25	63.98	64.61	0.71	0.29	66.72	65.47	66.09	0.72	0.32
Tripeptide Composition										

DT	62.27	47.32	54.70	0.58	0.10	94.70	19.69	56.66	0.63	0.22
RF	**67.74**	**73.33**	**70.57**	**0.77**	**0.41**	**68.17**	**77.97**	**73.14**	**0.81**	**0.46**
LR	63.96	65.43	64.71	0.71	0.29	68.49	68.28	68.38	0.75	0.37
XGB	59.78	63.98	61.91	0.68	0.24	49.36	80.16	64.98	0.72	0.31
KNN	67.02	65.59	66.29	0.73	0.33	71.38	66.25	68.78	0.76	0.38
GNB	71.60	49.90	60.60	0.61	0.22	74.44	55.63	64.90	0.65	0.31
SVC	39.29	81.93	60.90	0.66	0.24	0.00	100.00	50.71	0.50	0.00
Atoms										
DT	59.26	59.33	59.29	0.64	0.19	66.08	59.69	62.84	0.67	0.26
RF	61.51	61.67	61.59	0.67	0.23	65.43	65.16	65.29	0.70	0.31
LR	61.07	60.66	60.86	0.65	0.22	63.34	64.38	63.87	0.68	0.28
XGB	**61.27**	**63.16**	**62.23**	**0.68**	**0.24**	**62.06**	**67.03**	**64.58**	**0.70**	**0.29**
KNN	58.94	62.34	60.66	0.65	0.21	61.74	63.75	62.76	0.68	0.26
GNB	60.23	60.58	60.40	0.65	0.21	62.38	64.84	63.63	0.68	0.27
SVC	50.62	60.46	55.61	0.60	0.11	0.00	100.00	50.71	0.50	0.00
Bonds										
DT	61.59	63.08	62.35	0.66	0.25	64.63	63.91	64.26	0.68	0.29
RF	63.36	62.69	63.02	0.67	0.26	60.29	66.41	63.39	0.68	0.27
LR	61.31	62.96	62.15	0.67	0.24	65.11	66.88	66.01	0.71	0.32
XGB	**63.96**	**64.53**	**64.25**	**0.68**	**0.29**	**62.06**	**69.06**	**65.61**	**0.71**	**0.31**
KNN	63.24	64.14	63.69	0.68	0.27	62.86	66.56	64.74	0.70	0.29
GNB	61.79	61.71	61.75	0.64	0.24	64.31	65.00	64.66	0.68	0.29
SVC	63.52	63.28	63.40	0.67	0.27	67.69	64.53	66.09	0.70	0.32

Repeats and Distribution

Features like single-spaced amino acid pair composition (kaa), 2-spaced amino acid pair composition (kaa2) were computed for repeats and distribution category. It is clearly shown in the Table 4 that SVC-based models for both the features is performing well when compared to other classifiers. The model for kaa and kaa2 has attained the AUC of 0.80 and 0.77 for testing dataset, respectively.

Table 4: The performance of machine learning models developed using repeats and distribution

Single-spaced amino acid pair composition										
	Training					Testing				
Classifier	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	57.73	59.13	58.44	0.61	0.17	61.58	61.25	61.41	0.62	0.23
RF	68.42	71.33	69.90	0.77	0.40	71.87	72.03	71.95	0.78	0.44
LR	66.98	67.85	67.42	0.74	0.35	67.52	71.72	69.65	0.77	0.39
XGB	67.70	67.58	67.64	0.73	0.35	70.26	67.34	68.78	0.75	0.38
KNN	63.16	62.30	62.72	0.68	0.26	64.47	59.38	61.89	0.69	0.24
GNB	64.64	64.06	64.35	0.70	0.29	64.31	66.56	65.45	0.71	0.31
SVC	**69.02**	**70.28**	**69.66**	**0.77**	**0.39**	**73.15**	**73.59**	**73.38**	**0.80**	**0.47**
2-spaced amino acid pair composition										
DT	63.76	54.99	59.31	0.61	0.19	64.31	52.34	58.24	0.59	0.17
RF	69.87	67.35	68.59	0.76	0.37	69.94	67.97	68.94	0.75	0.38
LR	67.58	67.62	67.60	0.74	0.35	68.01	69.84	68.94	0.74	0.38
XGB	66.41	67.23	66.83	0.73	0.34	65.60	67.66	66.64	0.73	0.33
KNN	63.84	60.42	62.11	0.67	0.24	68.01	53.44	60.62	0.66	0.22
GNB	68.62	63.67	66.11	0.70	0.32	68.33	63.59	65.93	0.71	0.32
SVC	**69.67**	**70.20**	**69.94**	**0.76**	**0.40**	**71.70**	**70.47**	**71.08**	**0.77**	**0.42**

Models developed on Physio-Chemical Properties, Shannon Entropy and Binary Profile

Several ML-based were developed based on physicochemical properties, shannon entropy, and binary profiles of the sequences. The performance of all three features is shown in Table 5. It can be seen that XGB-based model attained maximum AUC of 0.76 and 0.66 on testing dataset for PCP and SE, whereas RF-based model has achieved the maximum AUC and MCC with balanced sensitivity and specificity in case of BP.

Table 5: The performance of machine learning models developed using physio-chemical properties, shannon entropy and binary profiles

Physio--Chemical Properties

Training						Testing				
Classifier	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	60.43	60.23	60.33	0.64	0.21	64.63	63.28	63.95	0.69	0.28
RF	67.10	67.85	67.48	0.74	0.35	68.33	67.97	68.15	0.75	0.36
LR	65.81	63.90	64.84	0.72	0.30	65.43	64.69	65.06	0.72	0.30
XGB	**68.66**	**68.99**	**68.83**	**0.75**	**0.38**	**68.49**	**67.34**	**67.91**	**0.76**	**0.36**
KNN	64.44	63.90	64.17	0.71	0.28	66.72	62.19	64.42	0.72	0.29
GNB	64.44	63.36	63.89	0.68	0.28	65.43	65.00	65.21	0.69	0.30
SVC	65.45	64.10	64.76	0.71	0.30	65.76	65.16	65.45	0.72	0.31
Shannon Entropy										
DT	61.83	61.09	61.46	0.65	0.23	59.00	65.78	62.44	0.66	0.25
RF	62.48	60.62	61.53	0.66	0.23	63.18	61.09	62.12	0.67	0.24
LR	77.66	34.42	55.75	0.58	0.13	74.44	37.19	55.55	0.58	0.13
XGB	**61.79**	**61.83**	**61.81**	**0.66**	**0.24**	**61.58**	**63.59**	**62.60**	**0.66**	**0.25**
KNN	61.15	61.60	61.38	0.64	0.23	61.09	60.94	61.01	0.65	0.22
GNB	57.05	53.66	55.33	0.58	0.11	57.40	54.69	56.02	0.58	0.12
SVC	53.40	56.51	54.97	0.58	0.10	57.24	55.00	56.10	0.58	0.12
Binary Profiles										
DT	59.62	58.98	59.29	0.61	0.19	55.63	59.06	57.37	0.60	0.15
RF	**68.90**	**70.90**	**69.92**	**0.77**	**0.40**	**67.04**	**71.41**	**69.26**	**0.77**	**0.39**
LR	69.39	70.75	70.08	0.76	0.40	66.24	70.00	68.15	0.75	0.36
XGB	68.78	69.38	69.08	0.76	0.38	68.33	67.66	67.99	0.75	0.36
KNN	60.91	60.31	60.60	0.65	0.21	66.56	51.09	58.72	0.63	0.18
GNB	86.38	37.90	61.81	0.63	0.28	87.94	38.13	62.68	0.64	0.30
SVC	71.15	69.50	70.31	0.77	0.41	71.22	70.00	70.60	0.76	0.41

Miscellaneous

Features like APAAC, PAAC, CTD, QSO and SOCN based features were grouped under miscellaneous category. The performance all these features is tabulated in Table 6.

Table 6: The performance of machine learning models developed using miscellaneous features

Amphiphilic Pseudo Amino Acid Composition

	Training					Testing				
Classifier	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	61.83	61.75	61.79	0.66	0.24	62.70	65.78	64.26	0.69	0.29
RF	70.71	69.73	70.21	0.77	0.40	70.74	70.47	70.60	0.78	0.41
LR	68.42	69.85	69.14	0.76	0.38	69.13	72.03	70.60	0.78	0.41
XGB	**68.34**	**67.89**	**68.11**	**0.75**	**0.36**	**70.10**	**68.44**	**69.26**	**0.77**	**0.39**
KNN	69.23	67.35	68.27	0.75	0.37	69.61	67.34	68.46	0.75	0.37
GNB	66.53	66.60	66.57	0.73	0.33	64.31	67.03	65.69	0.73	0.31
SVC	69.79	71.06	70.43	0.77	0.41	71.87	73.28	72.58	0.79	0.45
Pseudo Amino Acid Composition										
DT	60.79	62.57	61.69	0.65	0.23	58.20	64.38	61.33	0.64	0.23
RF	69.55	68.75	69.14	0.76	0.38	71.87	69.69	70.76	0.78	0.42
LR	69.02	69.69	69.36	0.76	0.39	70.26	72.19	71.24	0.78	0.43
XGB	68.26	67.50	67.88	0.75	0.36	71.87	69.38	70.60	0.77	0.41
KNN	67.22	68.40	67.82	0.75	0.36	67.69	68.75	68.23	0.76	0.36
GNB	66.29	66.45	66.37	0.72	0.33	63.99	66.41	65.21	0.73	0.30
SVC	**70.23**	**70.08**	**70.16**	**0.77**	**0.40**	**73.47**	**71.09**	**72.27**	**0.78**	**0.45**
Conjoint Triad Descriptors										
DT	62.52	63.67	63.10	0.68	0.26	61.09	67.03	64.11	0.69	0.28
RF	69.75	69.77	69.76	0.76	0.40	69.61	69.22	69.41	0.77	0.39
LR	65.29	64.57	64.92	0.72	0.30	65.76	63.75	64.74	0.72	0.30
XGB	**69.67**	**70.24**	**69.96**	**0.77**	**0.40**	**71.54**	**69.53**	**70.52**	**0.78**	**0.41**
KNN	65.13	66.29	65.72	0.72	0.31	67.69	64.22	65.93	0.73	0.32
GNB	63.88	63.90	63.89	0.70	0.28	65.60	63.59	64.58	0.71	0.29
SVC	64.97	65.62	65.30	0.72	0.31	65.76	65.94	65.85	0.72	0.32
Quasi Sequence Order										
DT	60.19	59.45	59.81	0.64	0.20	63.67	57.66	60.62	0.64	0.21
RF	70.19	69.97	70.08	0.77	0.40	72.19	70.31	71.24	0.79	0.43
LR	68.86	69.81	69.34	0.76	0.39	70.10	71.88	71.00	0.78	0.42
XGB	68.34	69.18	68.77	0.75	0.38	71.22	72.03	71.63	0.78	0.43
KNN	**67.50**	**67.03**	**67.26**	**0.74**	**0.35**	**69.78**	**66.56**	**68.15**	**0.75**	**0.36**

GNB	64.81	63.90	64.35	0.70	0.29	61.42	64.84	63.15	0.69	0.26
SVC	69.02	69.03	69.03	0.75	0.38	70.58	70.78	70.68	0.77	0.41
Sequence Order Coupling Number										
DT	60.95	57.41	59.16	0.61	0.18	66.24	55.47	60.78	0.62	0.22
RF	59.02	58.98	59.00	0.63	0.18	60.13	62.66	61.41	0.64	0.23
LR	59.74	60.81	60.29	0.64	0.21	61.09	62.50	61.81	0.66	0.24
XGB	57.37	60.89	59.16	0.63	0.18	59.33	62.66	61.01	0.64	0.22
KNN	60.67	58.86	59.75	0.63	0.20	60.77	59.69	60.22	0.65	0.21
GNB	60.59	59.45	60.01	0.64	0.20	62.70	59.53	61.09	0.66	0.22
SVC	**58.66**	**61.99**	**60.35**	**0.64**	**0.21**	**63.99**	**59.22**	**61.57**	**0.66**	**0.23**

Performance of ML-based models on validation dataset

Combined features

In order to improve the performance of the model, we have combined different sets of features and evaluated the performance on validation dataset. The performance RF-based model with combined features is shown in Table 7. It has been observed that combination of different feature sets have improved the performance of the model as compared to individual feature.

Table 7: The performance of RF-based model with combined features on training and validation dataset

Combined Features	Training Dataset	Validation Dataset
AAC+DPC	82.30	77.71
AAC+DPC+BP	81.57	77.60
AAC+DPC+APAAC	81.80	77.28
AAC+ DPC + C	80.20	77.17
AAC+BP	80.10	77.16
AAC+BP +APAAC+ DPC	81.41	77.12
AAC+BP +APAAC	80.20	76.40
AAC+DPC+TPC	82.80	76.20
AAC+APAAC	80.30	76.17
AAC+ C	78.30	75.70

AAC+AAI+APAAC	75.04	72.88
AAC+DPC+AAI	75.14	72.84
AAC+BP +AAI	74.60	72.60
AAC+AAI	74.60	72.40

Performance of RF-based model on best feature set

The best performance of the model is obtained after combining evolutionary information-based features (PSSM profiles), physicochemical properties, binary profile along with dipeptide composition. We have tune RF-based model with different parameters to attain the maximum performance in terms of AUC. After combining these features, it has been observed that the model achieved the highest AUC of 79.6 on validation dataset, which is higher than individual feature. The performance of RF-based on best feature set with different parameters is shown in Table 8.

Table 8: The performance of RF-based on best feature set with different parameters

Best Model (PSSM + PCP + BP + DPC)		
ML Model	**Training Dataset**	**Validation dataset**
Random Forest (class_weight='balanced', criterion='entropy', max_depth=100, max_features='log2', n_estimators=8000)	79.1	78.5
Random Forest (n_estimators=8000, n_jobs=-1, random_state=13)	79.4	79.6

Discussion

The immune system is subdivided into innate and adaptive immune systems. Within hours after a foreign pathogen appearance in the body, innate immunity comes into play that involves nonspecific defence mechanisms. All multicellular beings have innate immunity. In contrast, adaptive immunity is only present in vertebrates. The adaptive immune system is able to destroy invaders on their own. The adaptive immune system is capable of immunological memory and tend to remember the pathogens that once infect host body and enables stronger attacks each time that pathogen is encountered [63] B-cell epitopes are regions on the surface of antigen, which antibodies identify with greater specificity, binds to and elicit immune response. This binding interaction between epitope (antigen) and paratope (antibodies) is core of the adaptive immunity.

Linear and Discontinuous B-cell epitope prediction has been a growing interest ever since the first method developed around 1981. Recognition of B-cell epitope with the help of accurate in-silico based prediction approach can aid the faster and cheaper vaccine design process which has become a necessity since the rise in cases of infectious diseases and pandemic over the last decade. Moreover, the experimental approach costs animal life and is a laborious and time-costing approach. Henceforth, predicting B-cell epitope using in silico approach is of great importance in the field of synthetic vaccine design, diagnostic test and immunotherapeutic [59]. Prediction of linear and discontinuous B-cell epitopes are divided into sequence based and structural based approaches. Various prediction methods are developed that are targeted on prediction of either linear or discontinuous B-cell epitope using the as either linear or structural based information [49].

In our study we developed approach to classify and identify B-cell epitope from non-epitopes by applying various approaches of feature generation and applying ML model prediction. We have generated composition-based features, physicochemical features, binary profile, repeats and distribution-based features, structural features as well as miscellaneous features. We tried different combinations of these features and applied feature selection techniques. Different machine learning models were trained on these features but best results were captured by a RF-based model on combining AAC and DPC (AUC = 77.71) as well on combining AAC, Binary and DPC features (AUC= 77.6). Combining DPC with BP, PCP and PSSM and applying RF-based model have further improved the performance and attained AUC of 79.6. Combining evolutionary PSSM profile and structural properties like aromaticity index, isometric index with composition-based feature and binary feature have successfully predicted with higher sensitivity and specificity. We got higher performance in both training, testing as well as validation dataset.

Considering the present scenario of prevalent infectious diseases and the occurrence of a couple of pandemics in the last two decades, it is of utmost importance to develop tools that can classify discontinuous epitope from linear epitope. The ability to identify these antibody binding sites in antigenic sequence is primarily important in the field of medical, immunological and biological applications. We believe that our method would aid in the more accurate recognition of conformational epitope and thereby bring a significant improvement in the field of synthetic vaccine design and development.

Chapter 4

Summary

Vaccination is the safest and most effective way for preventive diseases. Over the years the world has experienced a wave of pandemic costing millions of lives. Since the introduction of Edward's vaccinia immunisation to ameliorate the smallpox epidemic, a large number of safe and effective vaccines have been developed for use in humans against viral pathogens. Vaccine induced protection is understood at best when compiled with information of viral pathogenesis and vaccine associated information such as vaccine type, age group, adjuvant, administration route. We tried to cover such information in a single platform and provided users the ability to access, search and download the information in a single platform in a user-friendly manner.

Conventional vaccine development has been dependent on empirical strategy which cost an average of fifteen years to develop one involving repetitious trial and error. To overcome this cumbersome and inefficient process, in silico vaccine design offers new and important addition to the evaluation of candidate's vaccine. B-cell epitope prediction helps in identifying potential candidates that are effective in eliciting immune response, reducing the cost and time associated with conventional vaccine design and development.

Future Objectives

As vaccine targeted preventable infectious diseases continue to slow down, people have become increasingly concerned about the risks associated with vaccines, especially on child and pregnant ladies' health. In our database, we have covered vaccines approved or in development for age groups. There is scope of covering the consequences and components of vaccines that have direct effect. Our model for prediction B-cell epitope is based on sequential peptide regions. There is scope of adding epitope associated immunological responses information such as Major Histocompatibility Complex (MHC) to further classify B-cell epitopes into linear and conformational B-cell epitopes for effective design of synthetic vaccines.

Bibliography

[1] A.J. Pollard, E.M. Bijker, Publisher Correction: A guide to vaccinology: from basic principles to new developments., Nat. Rev. Immunol. 21 (2021) 129. https://doi.org/10.1038/s41577-020-00497-5.

[2] V. Kaminskyy, B. Zhivotovsky, To kill or be killed: how viruses interact with the cell death machinery., J. Intern. Med. 267 (2010) 473–82. https://doi.org/10.1111/j.1365-2796.2010.02222.x.

[3] K. V. Chaitanya, Structure and Organization of Virus Genomes, in: Genome and Genomics, Springer Singapore, Singapore, 2019: pp. 1–30. https://doi.org/10.1007/978-981-15-0702-1_1.

[4] P.E. Pellett, S. Mitra, T.C. Holland, Basics of virology., Handb. Clin. Neurol. 123 (2014) 45–66. https://doi.org/10.1016/B978-0-444-53488-0.00002-X.

[5] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation., Cell. Mol. Life Sci. 73 (2016) 4433–4448. https://doi.org/10.1007/s00018-016-2299-6.

[6] P.-Y. Lozach, Early Virus-Host Cell Interactions., J. Mol. Biol. 430 (2018) 2555–2556. https://doi.org/10.1016/j.jmb.2018.06.049.

[7] S.A. Plotkin, Vaccines for epidemic infections and the role of CEPI., Hum. Vaccin. Immunother. 13 (2017) 2755–2762. https://doi.org/10.1080/21645515.2017.1306615.

[8] M. Kvansakul, Viral Infection and Apoptosis., Viruses. 9 (2017). https://doi.org/10.3390/v9120356.

[9] B. Yang, K.D. Yang, Immunopathogenesis of Different Emerging Viral Infections: Evasion, Fatal Mechanism, and Prevention., Front. Immunol. 12 (2021) 690976. https://doi.org/10.3389/fimmu.2021.690976.

[10] V. Bernasconi, P.A. Kristiansen, M. Whelan, R.G. Román, A. Bettis, S.A. Yimer, C. Gurry, S.R. Andersen, D. Yeskey, H. Mandi, A. Kumar, J. Holst, C. Clark, J.P. Cramer, J.-A. Røttingen, R. Hatchett, M. Saville, G. Norheim, Developing vaccines against epidemic-prone emerging infectious diseases., Bundesgesundheitsblatt. Gesundheitsforschung. Gesundheitsschutz. 63 (2020) 65–73. https://doi.org/10.1007/s00103-019-03061-2.

[11] J. Louten, Virus Transmission and Epidemiology, in: Essent. Hum. Virol., Elsevier, 2016: pp. 71–92. https://doi.org/10.1016/B978-0-12-800947-5.00005-3.

[12] D.T. Mourya, P.D. Yadav, P.T. Ullas, S.D. Bhardwaj, R.R. Sahay, M.S. Chadha, A.M. Shete, S. Jadhav, N. Gupta, R.R. Gangakhedkar, P. Khasnobis, S.K. Singh, Emerging/re-emerging viral diseases & new viruses on the Indian horizon., Indian J. Med. Res. 149 (2019) 447–467. https://doi.org/10.4103/ijmr.IJMR_1239_18.

[13] D.M. Morens, A.S. Fauci, Emerging Pandemic Diseases: How We Got to COVID-19., Cell. 182 (2020) 1077–1092. https://doi.org/10.1016/j.cell.2020.08.021.

[14] M. Trovato, R. Sartorius, L. D'Apice, R. Manco, P. De Berardinis, Viral Emerging Diseases: Challenges in Developing Vaccination Strategies., Front. Immunol. 11 (2020) 2130. https://doi.org/10.3389/fimmu.2020.02130.

[15] J.F. Bale, Emerging viral infections., Semin. Pediatr. Neurol. 19 (2012) 152–7. https://doi.org/10.1016/j.spen.2012.02.001.

[16] G. Andrei, Vaccines and Antivirals: Grand Challenges and Great Opportunities, Front. Virol. 1 (2021). https://doi.org/10.3389/fviro.2021.666548.

[17] A. Tagliabue, R. Rappuoli, Changing Priorities in Vaccinology: Antibiotic Resistance Moving to the Top., Front. Immunol. 9 (2018) 1068. https://doi.org/10.3389/fimmu.2018.01068.

[18] No Title, (n.d.).

[19] S. Payne, Viral Vaccines, in: Viruses, Elsevier, 2017: pp. 73–79. https://doi.org/10.1016/B978-0-12-803109-4.00007-6.

[20] A.J. Pollard, E.M. Bijker, A guide to vaccinology: from basic principles to new developments., Nat. Rev. Immunol. 21 (2021) 83–100. https://doi.org/10.1038/s41577-020-00479-7.

[21] K.L. Goldenthal, K. Midthun, K.C. Zoon, Control of Viral Infections and Diseases, 1996. http://www.ncbi.nlm.nih.gov/pubmed/21413344.

[22] P. Zimmermann, N. Curtis, Factors That Influence the Immune Response to Vaccination., Clin. Microbiol. Rev. 32 (2019). https://doi.org/10.1128/CMR.00084-18.

[23] S. Plotkin, History of vaccination., Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 12283–7. https://doi.org/10.1073/pnas.1400472111.

[24] 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.

[25] J. Wallis, D.P. Shenton, R.C. Carlisle, Novel approaches for the design, delivery and administration of vaccine technologies., Clin. Exp. Immunol. 196 (2019) 189–204. https://doi.org/10.1111/cei.13287.

[26] M. Brisse, S.M. Vrba, N. Kirk, Y. Liang, H. Ly, Emerging Concepts and Technologies in Vaccine Development., Front. Immunol. 11 (2020) 583077. https://doi.org/10.3389/fimmu.2020.583077.

[27] P.D. Minor, Live attenuated vaccines: Historical successes and current challenges., Virology. 479–480 (2015) 379–92. https://doi.org/10.1016/j.virol.2015.03.032.

[28] C. D'Amico, F. Fontana, R. Cheng, H.A. Santos, Development of vaccine formulations: past, present, and future., Drug Deliv. Transl. Res. 11 (2021) 353–372. https://doi.org/10.1007/s13346-021-00924-7.

[29] L. Versteeg, M.M. Almutairi, P.J. Hotez, J. Pollet, Enlisting the mRNA Vaccine Platform to Combat Parasitic Infections., Vaccines. 7 (2019). https://doi.org/10.3390/vaccines7040122.

[30] M.J. Francis, Recent Advances in Vaccine Technologies., Vet. Clin. North Am. Small Anim. Pract. 48 (2018) 231–241. https://doi.org/10.1016/j.cvsm.2017.10.002.

[31] M. Robert-Guroff, Replicating and non-replicating viral vectors for vaccine development., Curr. Opin. Biotechnol. 18 (2007) 546–56. https://doi.org/10.1016/j.copbio.2007.10.010.

[32] R. Cid, J. Bolívar, Platforms for Production of Protein-Based Vaccines: From Classical to Next-Generation Strategies., Biomolecules. 11 (2021). https://doi.org/10.3390/biom11081072.

[33] J. Pollet, W.-H. Chen, U. Strych, Recombinant protein vaccines, a proven approach against coronavirus pandemics., Adv. Drug Deliv. Rev. 170 (2021) 71–82. https://doi.org/10.1016/j.addr.2021.01.001.

[34] C. Zhang, G. Maruggi, H. Shan, J. Li, Advances in mRNA Vaccines for Infectious Diseases., Front. Immunol. 10 (2019) 594. https://doi.org/10.3389/fimmu.2019.00594.

[35] A. Roldão, M.C.M. Mellado, L.R. Castilho, M.J.T. Carrondo, P.M. Alves, Virus-like particles in vaccine development., Expert Rev. Vaccines. 9 (2010) 1149–76. https://doi.org/10.1586/erv.10.115.

[36] S.D. Jazayeri, C.L. Poh, Development of Universal Influenza Vaccines Targeting Conserved Viral Proteins., Vaccines. 7 (2019). https://doi.org/10.3390/vaccines7040169.

[37] K.H. Khan, DNA vaccines: roles against diseases., Germs. 3 (2013) 26–35. https://doi.org/10.11599/germs.2013.1034.

[38] T. Kramps, J. Probst, Messenger RNA-based vaccines: progress, challenges, applications., Wiley Interdiscip. Rev. RNA. 4 (n.d.) 737–49. https://doi.org/10.1002/wrna.1189.

[39] K. Singh, S. Mehta, The clinical development process for a novel preventive vaccine: An overview., J. Postgrad. Med. 62 (n.d.) 4–11. https://doi.org/10.4103/0022-3859.173187.

[40] A.L. Cunningham, N. Garçon, O. Leo, L.R. Friedland, R. Strugnell, B. Laupèze, M. Doherty, P. Stern, Vaccine development: From concept to early clinical testing., Vaccine. 34 (2016) 6655–6664. https://doi.org/10.1016/j.vaccine.2016.10.016.

[41] A. Giraud-Gatineau, P. Colson, M.-T. Jimeno, C. Zandotti, L. Ninove, C. Boschi, J.-C. Lagier, B. La Scola, H. Chaudet, D. Raoult, Comparison of mortality associated with respiratory viral infections between December 2019 and March 2020 with that of the previous year in Southeastern France., Int. J. Infect. Dis. 96 (2020) 154–156. https://doi.org/10.1016/j.ijid.2020.05.001.

[42] B.S. Graham, N.J. Sullivan, Emerging viral diseases from a vaccinology perspective: preparing for the next pandemic., Nat. Immunol. 19 (2018) 20–28. https://doi.org/10.1038/s41590-017-0007-9.

[43] S. Rauch, E. Jasny, K.E. Schmidt, B. Petsch, New Vaccine Technologies to Combat Outbreak Situations., Front. Immunol. 9 (2018) 1963. https://doi.org/10.3389/fimmu.2018.01963.

[44] L.B. Nicholson, The immune system., Essays Biochem. 60 (2016) 275–301. https://doi.org/10.1042/EBC20160017.

[45] D.D. Chaplin, Overview of the immune response., J. Allergy Clin. Immunol. 125 (2010) S3-23. https://doi.org/10.1016/j.jaci.2009.12.980.

[46] J.S. Marshall, R. Warrington, W. Watson, H.L. Kim, An introduction to immunology and immunopathology., Allergy Asthma. Clin. Immunol. 14 (2018) 49. https://doi.org/10.1186/s13223-018-0278-1.

[47] N.C. Smith, M.L. Rise, S.L. Christian, A Comparison of the Innate and Adaptive Immune Systems in Cartilaginous Fish, Ray-Finned Fish, and Lobe-Finned Fish., Front. Immunol. 10 (2019) 2292. https://doi.org/10.3389/fimmu.2019.02292.

[48] F.A. Bonilla, H.C. Oettgen, Adaptive immunity., J. Allergy Clin. Immunol. 125 (2010)

S33-40. https://doi.org/10.1016/j.jaci.2009.09.017.

[49] M.C. Jespersen, S. Mahajan, B. Peters, M. Nielsen, P. Marcatili, Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes., Front. Immunol. 10 (2019) 298. https://doi.org/10.3389/fimmu.2019.00298.

[50] L. Potocnakova, M. Bhide, L.B. Pulzova, An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction., J. Immunol. Res. 2016 (2016) 6760830. https://doi.org/10.1155/2016/6760830.

[51] P. Reche, D.R. Flower, M. Fridkis-Hareli, Y. Hoshino, Peptide-Based Immunotherapeutics and Vaccines 2017., J. Immunol. Res. 2018 (n.d.) 4568239. https://doi.org/10.1155/2018/4568239.

[52] J.L. Pellequer, E. Westhof, PREDITOP: a program for antigenicity prediction., J. Mol. Graph. 11 (1993) 204–10, 191–2. https://doi.org/10.1016/0263-7855(93)80074-2.

[53] A.J. Alix, Predictive estimation of protein linear epitopes by using the program PEOPLE., Vaccine. 18 (1999) 311–4. https://doi.org/10.1016/s0264-410x(99)00329-1.

[54] M.C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes., Nucleic Acids Res. 45 (2017) W24–W29. https://doi.org/10.1093/nar/gkx346.

[55] S. Saha, G.P.S. Raghava, Prediction of continuous B-cell epitopes in an antigen using recurrent neural network., Proteins. 65 (2006) 40–8. https://doi.org/10.1002/prot.21078.

[56] H. Singh, H.R. Ansari, G.P.S. Raghava, Improved method for linear B-cell epitope prediction using antigen's primary sequence., PLoS One. 8 (2013) e62216. https://doi.org/10.1371/journal.pone.0062216.

[57] Y. El-Manzalawy, D. Dobbs, V. Honavar, Predicting linear B-cell epitopes using string kernels., J. Mol. Recognit. 21 (n.d.) 243–55. https://doi.org/10.1002/jmr.893.

[58] Y. El-Manzalawy, V. Honavar, Recent advances in B-cell epitope prediction methods., Immunome Res. 6 Suppl 2 (2010) S2. https://doi.org/10.1186/1745-7580-6-S2-S2.

[59] J.L. Sanchez-Trincado, M. Gomez-Perosanz, P.A. Reche, Fundamentals and Methods for T- and B-Cell Epitope Prediction., J. Immunol. Res. 2017 (2017) 2680160. https://doi.org/10.1155/2017/2680160.

[60] P. Agrawal, S. Bhalla, K. Chaudhary, R. Kumar, M. Sharma, G.P.S. Raghava, In Silico Approach for Prediction of Antifungal Peptides., Front. Microbiol. 9 (2018) 323.

https://doi.org/10.3389/fmicb.2018.00323.

[61] J. Bac, E.M. Mirkes, A.N. Gorban, I. Tyukin, A. Zinovyev, Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation., Entropy (Basel). 23 (2021). https://doi.org/10.3390/e23101368.

[62] S.J. Rigatti, Random Forest., J. Insur. Med. 47 (n.d.) 31–39. https://doi.org/10.17849/insm-47-01-31-39.1.

[63] J. Parkin, B. Cohen, An overview of the immune system., Lancet (London, England). 357 (2001) 1777–89. https://doi.org/10.1016/S0140-6736(00)04904-7.