

**A High Throughput *In Silico* Pipeline to Discover
Potential Therapeutic Targets in Prokaryotic
Pathogens**



By

MUHAMMAD RIZWAN

NUST201260268MRCMS64012F

Supervisor

Dr. Jamil Ahmed

**RESEARCH CENTRE FOR MODELING & SIMULATION
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY**

2015

**A High Throughput *In silico* Pipeline to Discover
Potential Therapeutic Targets in Prokaryotic
Pathogens**

MUHAMMAD RIZWAN

Research Centre for Modeling & Simulation

A thesis submitted to the National University of Sciences & Technology
in partial fulfillment of the requirement for the degree of
Masters of Sciences

2015

STATEMENT OF ORIGINALITY

I Muhammad Rizwan, hereby declare that the work embodied in this dissertation is the result of work produced by me and has not been submitted for a higher degree to any other institution.

Date

MUHAMMAD
RIZWAN

Dedication

This dissertation is dedicated to my
PARENTS, TEACHERS, MENTORS, and
the destitute and poor patients from whom the current prevailing economic
system has snatched the right of treatment even made food difficult.

Acknowledgements

“Has there not been over man a period of time, when he was not a thing worth mentioning?(1) Verily, We have created man from drops of mixed semen (Nutfah), in order to try him, so We made him hearer and seer(2) Verily, We showed him the way, whether he be grateful or ungrateful(3).”

The QURAN, Surah Al-Insan 76/1:3

هَلْ أَتَى عَلَى الْإِنْسَانِ حِينٌ مِّنَ الدَّهْرِ لَمْ يَكُن شَيْئًا مَّذْكُورًا (١) إِنَّا خَلَقْنَا
الْإِنْسَانَ مِنْ نُطْفَةٍ أَمْشَاجٍ نَّبْتَلِيهِ فَجَعَلْنَاهُ سَمِيعًا بَصِيرًا (٢) إِنَّا هَدَيْنَاهُ السَّبِيلَ
إِمَّا شَاكِرًا وَإِمَّا كَفُورًا (٣)

“All the praises and thanks be to Allah, Who has guided us to this, and never could we have found guidance, were it not that Allah had guided us!”

The QURAN, Surah Al-A'raf 7/43

الْحَمْدُ لِلَّهِ الَّذِي هَدَانَا لِهَذَا وَمَا كُنَّا لِنَهْتَدِيَ لَوْلَا أَنْ هَدَانَا اللَّهُ

First of all, praise is due to almighty **ALLAH** with His compassion and mercifulness for giving me the strength to complete this endeavor.

This dissertation appears in its current form due to the assistance and guidance of several persons. I would therefore like to offer my sincere thanks to all of them. I would like to express my gratitude to my supervisor **Dr. Jamil Ahmed** (H.O.D, Department of Computational Sciences, Research Centre for Modeling & Simulation, RCMS, NUST) and co-supervisor **Dr. Amjad Ali** (Assistant Professor, Atta-ur-Rahman School of Applied Biosciences, ASAB, NUST) who taught me how to perform in a field which I never knew before.

I would like to extend my genuine appreciation to my committee members **Dr. Ishrat Jabeen** (Assistant Professor, Department of Computational Sciences, RCMS, NUST)

and **Mr. Tariq Saeed** (Assistant Professor and PhD candidate, Department of Computational Sciences, RCMS, NUST) and acknowledge their help and support.

I would like to convey special thanks to **Miss. Anam Naz** (a PhD scholar at Atta-ur-Rahman School of Applied Biosciences, ASAB, NUST) for her guidance and assistance and review. I am also thankful to **Dr. Afreenish Hassan** and **Miss. Kanwal Naz** (PhD scholars at Atta-ur-Rahman School of Applied Biosciences, ASAB, NUST) for reviewing this draft and provide valuable suggestions.

I highly appreciate the contribution of my sister **Tamsila Parveen** (MS Bioinformatics from COMSATS Institute of Information Technology, Islamabad) who helped me in learning genomics and bioinformatics.

I additionally appreciate the help and support of all persons who were directly or indirectly involved. I express my cordial gratitude to my parents and family members for their prayers, love, patience and unconditional support during the accomplishment of this degree and research period. I express my heartiest gratitude to my class mates and friends especially the partners of all fun and stress **Mr. Ehsaan Aadeeb** (now a PhD scholar at Kyungpook National University, Daegu, South Korea), **Mr. Ahsan** (Mobile and HPC Software architect) and **Dr. Raheel Khan** (Pharmacist at Pakistan Institute of Medical Sciences, PIMS, Islamabad) for their motivation. Finally, I would like to express special thanks & acknowledge the guidance and mentor support of my colleagues **Mr. Dawood Liaquat** (Manager Software Development at Techlogix, Lahore, Pakistan) and **Mr. Nauman Faridi** (VP Software Development at DPL, Islamabad, Pakistan).

I would also like to say thanks to **Mr. Nishant Gandhi** (M.Tech. in Computer Science & Engineering, Dept. of Computer Science & Engineering, Indian Institute of Technology, Patna, India) for his guidance about parallel processing in Java library JOMP (Java OpenMP).

I humbly extend my thanks to all concerned persons who co-operated with me in any capacity.

Abstract

Infectious diseases are emerging rapidly throughout the globe, and antibiotic resistant bacterial strains lead to the therapeutic failure; ultimately causing the high risk of cost and re-infection. Vaccination is considered as one of the most effective mechanisms for the treatment of a particular disease; and conventional vaccinology approaches have rendered certain limitations against some pathogens. Reverse vaccinology and subtractive proteomics are novel competent computational approaches to identify putative therapeutic targets against the infectious agents. In current study we have developed an *in silico*, multi-threaded, configurable and scalable pipeline employing subtractive-reverse vaccinology analysis technique and named it **VacSol** (<https://sourceforge.net/projects/vacsol/>). The principle objective of the VacSol development is to screen out genes/proteins from microbial genome/proteome that could be employed as potential therapeutic targets against them and furthermore be exploited for *in vitro* and *in vivo* laboratory evaluation.

VacSol is multi-mode approach working as a standalone tool as well as a pipeline or a package of tools. It has two major features, first is to prioritize proteins followed by determination of the major histocompatibility complex MHC I & II for the prioritized proteins. VacSol evaluates whole proteome (composed of thousands of protein sequences) of a microbe and prioritize the proteins being essential, virulent, non-host homolog and immunogenic simultaneously and then additionally perform epitope analysis of prioritized proteins. .

A study was conducted in Atta-ur-Rahman School of Applied Biosciences ASAB, NUST based on reverse vaccinology using online web based tools and evaluated the whole proteome of *Helicobacter Pylori* 26695. The approach took more than five months to prioritize proteins as potential therapeutic targets and their immunogenic peptides conserved among all *H. pylori* strains. VacSol was employed to repeat the same analysis and it evaluated *Helicobacter Pylori* proteome containing 1,576 proteins within one and a half hour on four core machine and provided ten protein sequences as

candidates for vaccine targets which can be further subjected for laboratory evaluation by domain experts. VacSol efficiently reduces the laborious and hectic effort of months by generating results within a few hours by integrating all the required tools. VacSol results can also be accelerated in proportion to number of cores available in the machine. By default it uses maximum number of available cores, but this feature is governable. VacSol was deployed and tested at Ubuntu 12.04 64 bit machine. Its results do not depend on any universal set of rules that may vary based on the provided input and the version of tools that have integrated in this pipeline. VacSol is an efficient, cost and time effective tool that eliminates the false candidates and recognizes the potential therapeutic targets for further laboratory evaluation.

List of abbreviations & acronyms

ABCpred	Artificial neural network based B-cell epitope prediction server
BCPREDS	B-cell epitope prediction server
BLAST	Basic Local Alignment Search Tool is an algorithm for comparing primary biological sequence information
BLASTp	Protein BLAST
CELLO	A subcellular localization predictor using SVM
CELLO2GO	A web server for protein subcellular localization prediction with functional Gene Ontology annotation
DA-PLS	Primary lateral sclerosis (PLS) is a kind of disorder and DA-PLS stands for Discriminant Analysis Primary lateral sclerosis
DEG	Database of Essential Genes
FASTA	A text-based format for representing either nucleotide sequences or peptide sequences
HMMTOP	Prediction tool of transmembrane helices and topology of proteins
IEDB	Immune Epitope Database and Analysis Resource
Jenner-Predict	Prediction server of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions
MHC	Major histocompatibility complex
MHCPred	A quantitative T-cell epitope prediction server of peptide MHC binding
MvirDB	A microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defense applications
MySQL	An open-source relational database management system
NERVE	New Enhanced Reverse Vaccinology Environment
OrthoMCL	Tool for identification of ortholog groups for eukaryotic genomes
PATRIC	PAThosystem Resource Integration Center.
PERL	Practical Extraction and Report Language
Phyre	Protein homology/analogy recognition engine
PPI	Protein Protein Interaction

PROFtmb	A web server for predicting bacterial transmembrane beta barrel proteins
ProPred1	Predictor of promiscuous MHC Class-I binding site
ProPredII	MHC Class-II binding peptide prediction server
PSORTb	Protein subcellular localization predictor with refined localization subcategories and predictive capabilities for all prokaryotes
PVC	Potential Vaccine Candidates
RV	Reverse Vaccinology
SPAAN	A software program for prediction of adhesins and adhesin-like proteins using neural networks
SVM	Support Vector Machine
TargetP	Predictor for subcellular location of eukaryotic proteins
TMHMM	Markov Model Prediction tool of TransMembrane Helices in proteins
TVC	Therapeutic Vaccine Candidates
UniProt	Universal Protein resource
Vacceed	A high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology
Vaxign	The first web-based Vaccine design program for Reverse Vaccinology and Applications for Vaccine development
VaxiJen	Prediction of Protective Antigens and Subunit Vaccines.
VFDB	Virulence Factors Database of pathogenic Bacteria
Victor	Manually curated virulence factor database
WoLf PSORT	WoLF PSORT is an extension of the PSORT program for protein subcellular localization prediction
PPI	Protein-protein Interaction

Table of Contents

1	Introduction.....	1
1.1	Background.....	1
1.2	Classification of Vaccinology	2
1.2.1	Conventional Vaccinology	2
1.2.1.1	Limitations of Conventional	3
1.2.2	Reverse Vaccinology.....	4
1.2.3	Comparison of Traditional and Computational Vaccinology	6
1.3	Limitations.....	6
1.4	Motivation	7
1.5	Aims and Objectives.....	9
2	Literature Review	10
2.1	NERVE.....	10
2.1.1	NERVE Features	10
2.1.2	NERVE Limitations	11
2.2	VaxiJen.....	11
2.2.1	VaxiJen Limitations	11
2.3	Vaxign	12
2.3.1	Vaxign Features.....	12
2.3.2	Vaxign Limitations.....	12
2.4	Vacceed	13
2.4.1	Vacceed Features.....	13
2.4.2	Vacceed Limitations.....	13
2.5	Jenner-predict	13
2.5.1	Jenner-predict Features	14
2.5.2	Jenner-Predict Limitations	14
2.6	Summary.....	15
3	Methodology	16
	Stepwise Subtractive Genomics Reverse Vaccinology Procedure	16
3.1	Functionality Identification	17

3.1.1	Localization Prediction	18
3.1.2	Screen out Essential Genes	18
3.1.3	Identification of Non-host Homologous	18
3.1.4	Screen out Virulent Factors	18
3.1.5	Screen out Transmembrane Helices	19
3.1.6	Functional Annotation	19
3.2	Tools Selection	19
3.2.1	Selection Criteria	20
3.2.2	Tools Used	20
3.3	Design Interface	23
3.4	Architecture Design	24
3.4.1	High Level Architecture Diagram	24
3.5	Implementation	30
3.5.1	Packages Designed for Organized Pipeline Development	31
4	Results	33
4.1	Localizer Results for Subcellular Localization Prediction	34
4.2	List of VacSol Screened out Proteins	35
4.3	Online Cross Checking of VacSol Generated Results	36
4.3.1	Sub-cellular Prediction Results	36
4.3.2	Essential Genes Prediction	36
4.3.3	Virulence Factor Detection Results	37
4.3.4	Number of Transmembrane Helices Prediction	37
4.4	Protein Annotation Results	38
4.5	Protein-Protein Interaction Results	39
5	Discussion	40
5.1	Future Perspective	44
1	Appendix A: Installation Guide	45
1.1	Pre Requisite Tools/Languages	45
1.1.1	PSORTb Installation	46
1.1.2	Perl (5.6.X or higher)	46
1.1.3	Bioperl (1.2.X or higher)	46

1.1.4	Stand Alone NCBI Blast	46
1.1.5	PFTOOLS.....	46
1.1.6	OSDDlinux Installation.....	50
1.1.7	Databases.....	51
1.1.8	VacSol	52
1.1.9	Configurations	52
2	Appendix B: User Guide	57
2.1	GUI Mode.....	59
2.1.1	Submission Method.....	60
2.1.2	Threshold: Homologous with Human	65
2.1.3	Threshold: Virulence Factors	66
2.1.4	Threshold: Localization Filter	67
2.1.5	Threshold: Essential Genes	68
2.1.6	Threshold: Epitope Maps	68
2.2	STANDALONE Mode	69
	Bibliography	71

LIST OF FIGURES

<i>Figure 1.1: Diagrammatic representation of Conventional Vaccinology.....</i>	<i>3</i>
<i>Figure 1.2: Diagrammatic representation of Reverse Vaccinology.....</i>	<i>4</i>
<i>Figure 1.3: Visualization of Conventional Reverse Vaccinology approach.</i>	<i>5</i>
<i>Figure 1.4: Visualization of Pan Genome or Comparative Genome approach.....</i>	<i>5</i>
<i>Figure 1.5: Visualization of Subtractive Genome Analys.....</i>	<i>6</i>
<i>Figure 1.6: Schematic View of H-Pylori Secretome and Exoproteome Analyysi.....</i>	<i>7</i>
<i>Figure 3.1: Schematic View of Pipeline Functionality.....</i>	<i>17</i>
<i>Figure 3.2: Modular-based VacSol Interface Design</i>	<i>23</i>
<i>Figure 3.3: Working modes of VacSo.....</i>	<i>24</i>
<i>Figure 3.4: Schematic view of VacSol Main Processing.....</i>	<i>26</i>
<i>Figure 3.5: Elaboration of in-depth processing of prioritized protein thread.....</i>	<i>27</i>
<i>Figure 3.6: Schematic Viw of Epitop Mapping through VacSol.</i>	<i>30</i>
<i>Figure 3.7: Overview of Package Diagram.....</i>	<i>31</i>
<i>Figure 4.1: Graphical Representation of Localizer Results.....</i>	<i>34</i>
<i>Figure 4.2: CELLO2GO Results.</i>	<i>36</i>
<i>Figure 4.3: Interactive Depiction of Protein-Protein Interaction Network.</i>	<i>39</i>

LIST OF TABLES

<i>Table 3.1: Tools used in VacSol. Describing the currently integrated tools in VacSol.</i>	20
<i>Table 3.2: Comparison of VacSol with already available PVCs prediction pipelines.</i>	21
<i>Table 3.3: Java Libraries used for VacSol.</i>	31
<i>Table 4.1: VacSol Input Parameters.</i>	33
<i>Table 4.2: Therapeutic/Putative Vaccine Candidates (TVCs/PVCs).</i>	34
<i>Table 4.3: VacSol Screened out Putative Vaccine Candidates.</i>	35
<i>Table 4.4: Database of Essential Genes (DEG) Online Results.</i>	36
<i>Table 4.5: TMHMM Results for Transmembrane Helices.</i>	37
<i>Table 4.6: Functional Annotation of Prioritized proteins.</i>	37

CHAPTER 1: Introduction

1 Introduction

Disease and infections contribute as one of the major factor predominantly affecting the human health over the years. infectious diseases are highlighted to be the main cause of disabilities and human deaths worldwide [1]. Eradication or elimination of such diseases always remain the focus of scientific research, and as a major requirement of human being since the beginning of life. Diseases can be classified on the basis of their causes, symptoms, and mode of mechanisms [2]. Most diseases and infections are caused by different types of micro-organisms including protozoa, fungi, virus and bacteria etc. generally termed as pathogenic micro-organisms or pathogens [3, 4]. Infections caused by pathogens are treated by either of two different ways; antibacterial drugs or vaccination.

Though reported for many successful cases antibacterial drugs do not work in every case as many of the resistant bacterial strains lead to fail this treatment regime [5]. These resistant strains can also be the cause of high risk rate of re-infection along with high cost and some deleterious side effects [6]. Vaccinations are considered as one of the most effective mechanisms for the treatment of particular bacterial diseases in last couple of centuries [7]. It played a key role in the elimination of infectious and threatening diseases, as well as served to significantly decrease mortality rate in some diseases which were considered incurable in the past. Vaccination helps to improve the immunity rate against particular diseases as it takes advantage of body's natural ability to discover how to eliminate almost any type of pathogenic agent or microbe.

1.1 Background

The earliest possible documented method regarding vaccination was variolation practiced in India and China, where inoculation with powdered scabs from people affected with smallpox was used to protect against the disease [8].

Vaccination is a practice in which patient is injected with attenuated or inactive form of pathogens to immune to further infections. The first vaccine was discovered in 17th century, in which infected material was isolated from cow and used for the treatment of smallpox [9]. Edward Jenner first time coined the terminology “vaccine” in 1796 on the basis of its discovery from a virus (cowpox) affecting cows and ‘vacca’ is a Latin word for cow [10, 11]. This vaccination practice was introduced into Western medicine formally, and then this practice was spread worldwide. Rational development of vaccine was started by Louis Pasteur, who formalized the process of vaccinology after recognizing microbes as infectious agents. Louis Pasteur established the basic rules of vaccine development based on “isolation, inactivation and injecting the attenuated microbes”. During last few centuries, various approaches were applied for the advancement of vaccine. Most of the 20th century developed vaccines were relied on the Pasteur’s principles of vaccinology which is not based on genomic sequences [12].

1.2 Classification of Vaccinology

Vaccinology mainly classify into two major classes traditional/conventional vaccinology and computational/reverse vaccinology.

1.2.1 Conventional Vaccinology

Vaccine development based on Pasteur’s vaccinology is usually referred as conventional vaccinology that is mentioned in Figure 1.1. These conventional approaches required *in-vitro* growth of pathogens which is not possible in all cases [13]. In spite of producing the fruitful results for many cases in history, this approach is time consuming and failed in most of the cases especially for microbes that could not be cultivated *in-vitro* [13].

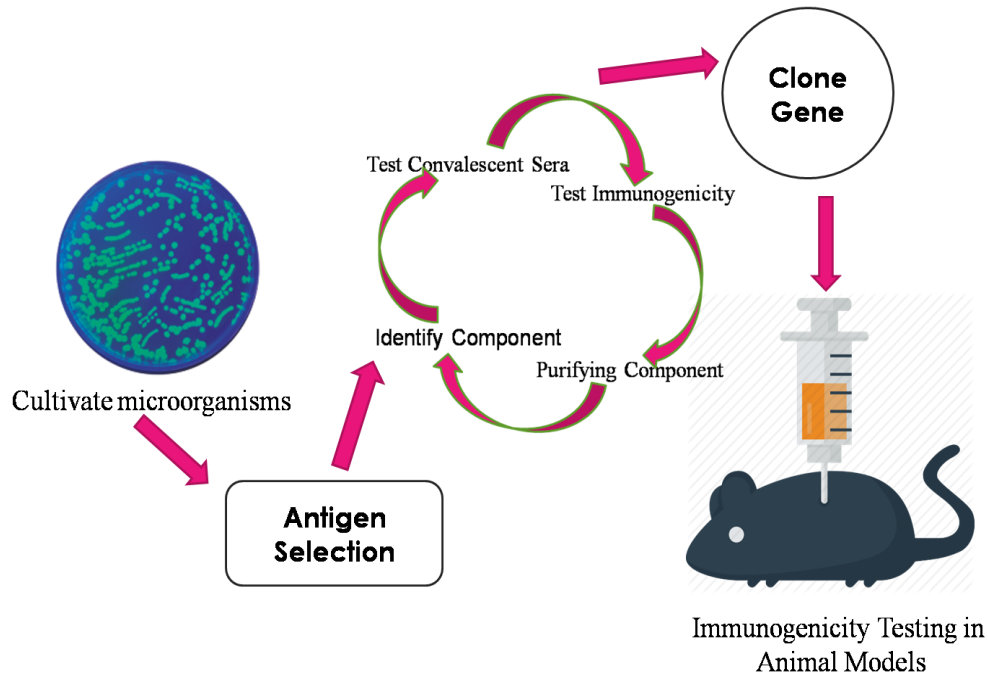


Figure 1.1: Diagrammatic representation of Conventional Vaccinology. Conventional preparation of vaccines is comprised of different stages; Firstly, it requires the cultivation of microorganisms and different selection materials are needed for the isolation and selection of antigens. Selected antigens are subjected for purifying and further immunogenicity testing. Different sera are required for preserving the antigens. Finally these screened antigens are conducted for cloning and the efficacy is checked on animal models. After clinical trials the vaccine is approved for use.

1.2.1.1 Limitations of Conventional Vaccinology

Few limitations which usually encountered in applying conventional vaccinology approach are mentioned below:

- Maintenance of specific cell culture and safety procedure.
- Impossible to grow all pathogens.
- Costly and time consuming procedure.
- Insufficient attenuation or killing may lead to revert virulence in final vaccine.

1.2.2 Reverse Vaccinology

With the advantage of bioinformatics approaches and high throughput sequencing techniques, Rino Rappouli has revolutionized the previous Pasteurs' vaccinology approach that is in use since decades [14]. Rino Rappouli elaborated a new era in this field with a complete paradigm shift [15, 16]. This new computational approach is termed as "Reverse Vaccinology" [14]. Reverse vaccinology is an advanced *in-silico* approach for the preparation of vaccines using genomic information and bioinformatics tools [11].

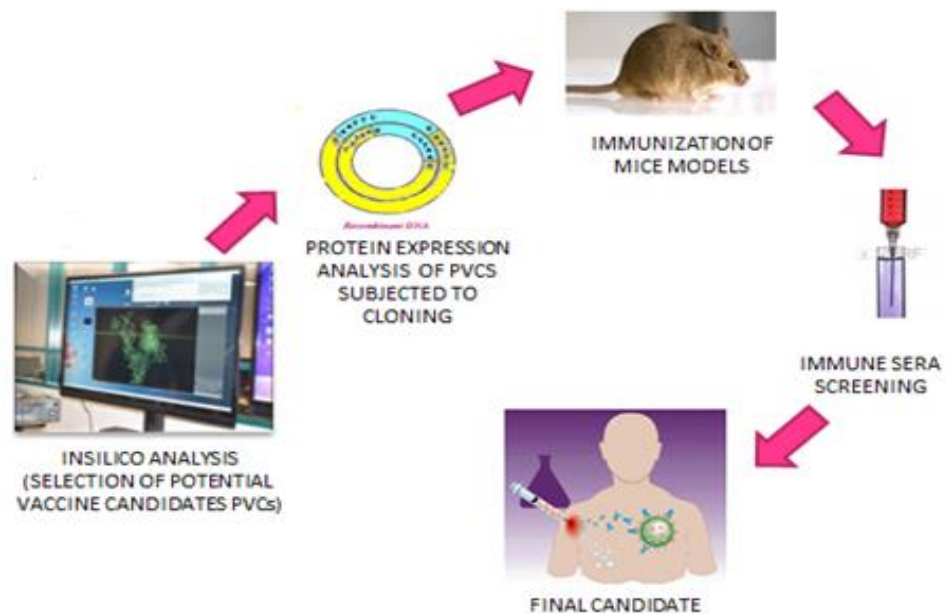


Figure 1.2: Diagrammatic representation of Reverse Vaccinology. This approach is based on the computational technique using the genome and proteome sequence. In this approach, complete microbial genome has screened for detecting potential vaccine candidates (PVCs)

1.2.2.1 Advantages of Reverse Vaccinology

The main advantage of reverse vaccinology is that it reduces the need of microorganism's cultivation, thus saves both time and cost for the identification of candidate therapeutic targets [12]. Vaccines developed through reverse vaccinology are safe and effective for human use, and vaccine against serogroup B meningococcus is one of the successful example of vaccine development through reverse vaccinology [16].

1.2.2.2 Stages of Reverse Vaccinology

Reverse vaccinology can be divided into following three stages [17].

First Phase: Conventional Reverse Vaccinology

Classical reverse vaccinology approach is efficient than conventional/traditional approach, and is applied for the identification of putative antigens which are surface exposed can be identified by mining genome sequence that can be used as vaccine candidates [18].

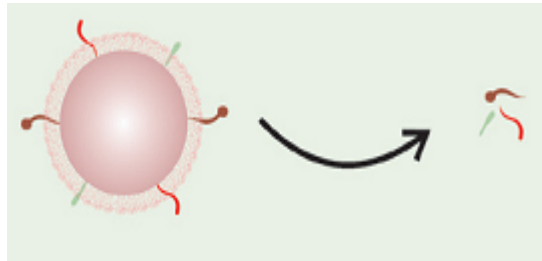


Figure 1.3: Visualization of Conventional Reverse Vaccinology approach. [1]

Second Phase: Pan Genome or Comparative Genome Analysis

Comparative or pan genome reverse vaccinology is a comparison based approach that compares varied genome sequences from the different microbial strains using computational analysis. This approach is based on estimation of core genome which is conserved and mutually shared by all genomes. It is an extensive approach and due to the variable antigenic behaviour, it leads to avoid the microbial escape [19, 20].

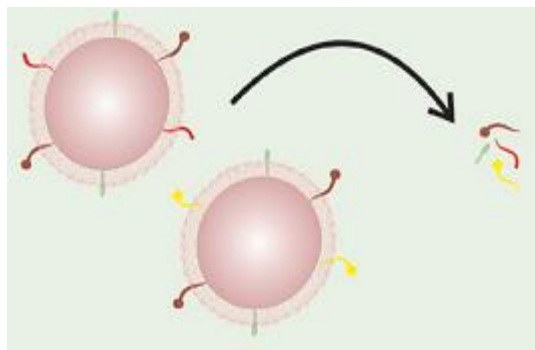


Figure 1.4: Visualization of Pan Genome or Comparative Genome approach. [1]

Third Phase: Subtractive Reverse Vaccinology

The subtractive reverse vaccinology is mainly the comparison between non-pathogenic and pathogenic microbial genome. This approach helps in the selection of antigens which take part in pathogenesis and virulence of microbe [17].

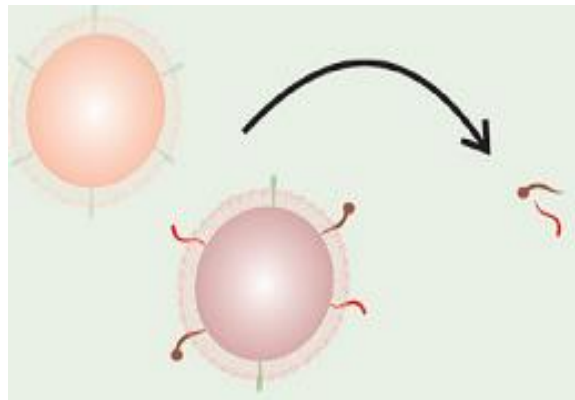


Figure 1.5: Visualization of Subtractive Genome Analysis. It represents pathogenically involved antigens [1]

1.2.3 Comparison of Traditional and Computational Vaccinology

Traditional vaccinology technique is an expensive and time consuming technique that requires 5-15 years. Advances in bioinformatics and high-throughput next generation sequencing remodel the slow growing vaccinology field to identify the potential therapeutic targets due the complete genome accessibility. This computational technique requires 1-3 years to obtain immunogenic target, and besides many other advantages, does not need the pathogenic cultivation [18].

1.3 Limitations

In silico tools show constraint in explaining the hosts behaviour to a protein or group of proteins [18] so it cannot replace the authenticity, acquired after laboratory experiment. But it will aid to reduce the test sample which will definitely help to cut down the time and cost [21]. This approach includes lack of algorithms for clearly indicating antigen and protective immune response correlation, and also lacks in providing entire information regarding putative candidates [5].

1.4 Motivation

A study was conducted in Atta-ur-Rahman School of Applied Biosciences (ASAB), NUST for the analysis of human gastric pathogen *H. pylori* for the identification of therapeutic drug targets against *H. pylori* pathogen. During that study vaccine candidates have identified and characterized by providing the detailed insights of *H. pylori* pathogen using reverse vaccinology [22]. ASAB students put lot of efforts for the completion of this study and required more than five months to analyse *H. pylori* 26695 proteome of 1,576 proteins manually.

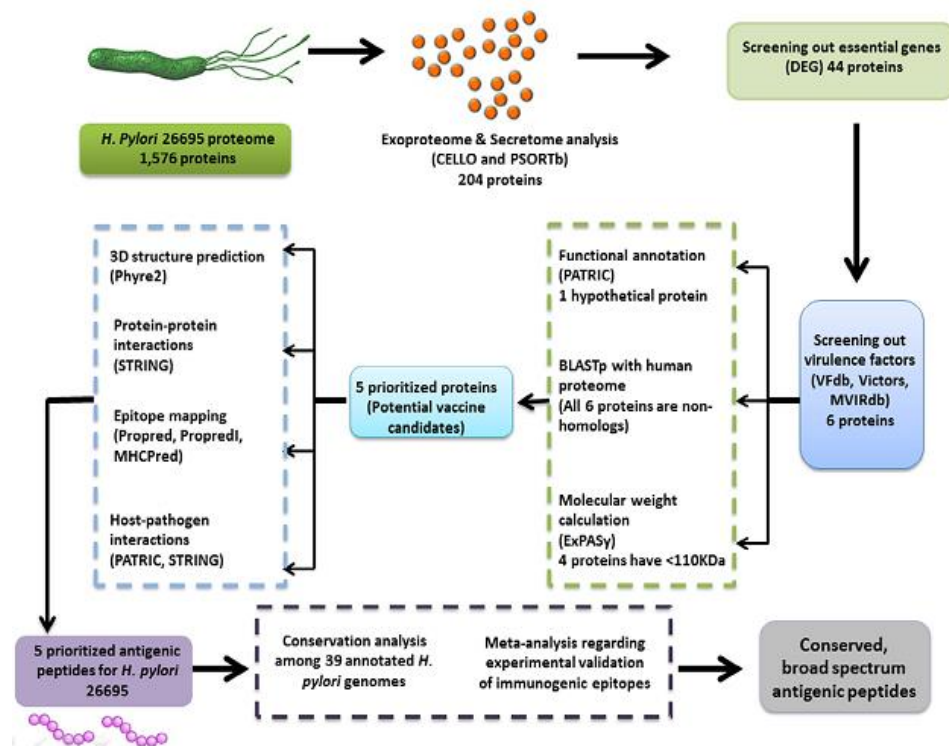


Figure 1.6: Schematic flow chart of H-Pylori Secretome and Exoproteome Analysis. As a result of secretome and exoproteome analysis of *H-pylori* 26695 carried out by using PSORTb and CELLO, only 204 proteins out of 1576 proteins were selected for further evaluation. Essential genes have been screened out for identifying the therapeutic drug targets based on virulent and essential proteins by using DEG. Victor, VFDB, and MvirDB were searched out for detecting the virulence proteins. These virulent and essential genes were subjected to further functional attribution annotation. Proteins exhibited molecular weight of Less than 110 kDa were selected for screening out the targeted vaccines. Only five protein sequences were prioritized and considered as potential vaccine candidate for further epitope and PPI interaction. All the above information has gained from the study "Identification of putative vaccine candidates against

Helicobacter pylori exploiting exoproteome and secretome: A reverse vaccinology based approach" by Naz *et al.*, [5]

Figure 1.6 depicted the step wise methodology used by the previous group (Naz *et al.*,) for obtaining the detailed inside information regarding gastric pathogen *H. pylori* [2]. This study was conducted on the basis of web based tools, databases and demanded laborious manual efforts which is the main limitation of this strategy. For example, an online web based tool PSortb (used for protein sub-cellular localization prediction) does not accept file of size more than 0.1 Megabytes. Similarly, CELLO2GO (a tool to predict localization of proteins within a proteome) is also limited in the perspective of file size and does not parse a file having more than 400 sequences [23, 24]. These limitations enforce the user to break file into different segments and then combine their results manually. Furthermore, these web based tools require excessive time in generating and delivering results.

In current genomic era, host and pathogen genome sequences are available in many authentic and curated databases. Availability of huge data sets make it easy to identify the therapeutic targets using prediction tools. It is difficult to acquire information of each and every tool with their limitations, and understand functionality and input parameters of these tools. Furthermore, different tools require different input file formats and parsing the first tool results into the format of another tool is also the major challenge for the researcher [25]. There are various web based or standalone tools, available for the identification of therapeutic targets but still there is a need of a comprehensive, scalable, and configurable pipeline which can integrate all results to predict some potential therapeutic targets.

VacSol is an *in-silico*, configurable, expandable, and scalable pipeline that base on the strategy presented in “*In-silico* subtractive genomics for target identification in human bacterial pathogens” [26] and “A novel strategy of epitope design in *Neisseria gonorrhoeae*” [27] with few modifications.

1.5 Aims and Objectives

This study has been designed to develop a comprehensive, configurable and scalable pipeline to predict some novel potential therapeutic targets against bacterial pathogens that have developed resistance against designed vaccines or drugs. The purpose of this pipeline is to reduce the manual effort as well as providing the computational aid to domain experts for their further in-vivo analysis. Main objectives of this pipeline are:

- Prediction of essential, virulent and non-host homologous targets against pathogen
- Prediction of surface exposed immunogenic epitopes having ability to bind most of the MHC alleles
- Annotation of targeted proteins
- Development of comprehensive and integrated pipeline

CHAPTER 2: Literature Review

2 Literature Review

Recent studies suggested that in the current era of computer-aided drug designing, *in silico* approaches for therapeutic agents identification are employed to cope with various illnesses [28-30]. Various computational vaccine target identification tools and databases have mainly developed to speed up the vaccine design [27]. New high throughput genomic sequencing and functional annotation data enable researchers to design reliable, predictive and analytical tools [31] which are usually available online [32-35], but only a handful of software [25] and pipelines [21, 31] are available for identifying the therapeutic bioactive targets using reverse vaccinology. On contrary, various immuno-informatics tools are available which only work for B-cells and T-cells immune epitope predictions [32-35].

A few pipelines already available for vaccine designs have been discussed below. Each has its own strengths and limitations based on their input/output parameters.

2.1 NERVE

NERVE stands for “New Enhanced Reverse Vaccinology Environment”. NERVE is Perl based modular pipeline for *in-silico* identification of potential vaccine candidates. NERVE generates results through text interface configuration. It is comprised of eight Perl scripts, and the whole script is divided into two parts for managing the whole process of vaccine candidate identification. Data is produced and stored in first part whereas second one refers to the selection of restricted data [31].

2.1.1 NERVE Features

NERVE uses PSORTb 2.0 [23] for the prediction of protein subcellular locations. SPAAN [36] software is used for the prediction of adhesion property of proteins. HMMTOP [37] algorithm implemented in PROFtmb [38] to calculate number of transmembrane helices. It is detected by using BLAST algorithm to compare human proteome against pathogen as query sequence. Essential genes are detected using BLASTp algorithm by selecting UniProt [39] database. Epitope binding prediction is not

clearly defined in NERVE, and here is considered as non-essential and just focused on exogenous proteins.

Acquisition and accuracy of this pipeline is tested on various bacterial proteome data (e.g. *Bacillus anthracis*, *Yersinia pestis*, *Pseudomonas aeruginosa*, *Streptococcus agalactiae*) by following above mentioned six different interpretive steps.

NERVE employs protein FASTA sequence as input generally in two ways, (1) direct pasting of sequence into window and (2) uploading .fasta file. Output is generated in MySQL table and .txt files. MySQL table further ranks the data into user unfriendly html table, and the final results are displayed in table and text format.

2.1.2 NERVE Limitations

Even though as an efficient modular based and standalone software for the identification of potential vaccine candidates, NERVE software is limited on extracellular proteins by extracting the cytosolic and inner membrane proteins [31]. This software/tool emphasized only on adhesive proteins presuming that adhesion proteins are potential vaccine candidates [21, 31]. Nevertheless, several non-adhesin putative functional proteins can also participate in host-pathogen interactions (i.e. porin, flagellin, invasin, etc.), and most of them are pathogenic as well as antigenic [40].

2.2 VaxiJen

VaxiJen is the first protective antigen prediction server designed on alignment-independent strategy. In this server, antigens are detected and classified on protein physicochemical properties without applying sequence alignment. For determining the whole protein antigenicity, VaxiJen used three different databases, bacterial, viral, and tumor with 70-89% prediction accuracy [41]. This server is applied on only 200 observations (100 non-antigens and 100 known antigens) from each dataset, and is designed on different criteria based on “Discriminant analysis by partial least square”. VaxiJen server can be used solely or with alignment dependent approaches.

2.2.1 VaxiJen Limitations

VaxiJen server is designed and based on discriminant analysis and partial least square (DA-PLS) methods by using 100 known (positive) protective antigens and 100 non-

antigens (negative) datasets for predicting PVCs. Surprisingly, from applied whole bacterial proteome, this server with default parameters predicts more than half of proteins as protective antigens making it almost impractical [31].

2.3 Vaxign

Vaxign is the first freely web-based, genome-wide targeted vaccine prediction program, based on reverse vaccinology and immune informatics strategy for predicting MHC class I and II epitopes [21].

2.3.1 Vaxign Features

Vaxign uses PSORTb 2.0 with 96% precision for the prediction of protein subcellular localization. It uses HMMTOP algorithm that is implemented in PROFTmb with 86% accuracy for calculating number of transmembrane helices. Unfortunately not all proteins through Vaxign can be analyzed for transmembrane helices due to the slow processing of PROFTmb (very time consuming). Adhesin probability is predicted using SPAAN with a cut-off of 0.51 [36]. OrthoMCL [42] is used for the detection of host-pathogens non-homology. IEDB immune epitope database [43] is used for epitope binding, and it detects only T-cells of either MHC class I or II epitope binding.

Vaxign includes adhesion probability, epitope binding to MHC class I and class II, protein subcellular location, human/mouse protein conservation, and transmembrane helices. Vaxign utility was checked on uropathogenic *Escherichia coli* (UPEC) and compared with several experimental studies.

Input is accepted in only two ways, (1) directly paste the sequence and (2) upload .fasta file with upto 500 sequences. It generates the results which can be exported to MS Excel file.

2.3.2 Vaxign Limitations

Even though Vaxign is very good for vaccine prediction, but it exhibits a few of limitations, such as: limited number of input ways and provides overall result which can be exported to Excel (limited file formats number). Conversely, it is web based and slower process and not all proteins are filtered for transmembrane domain analysis [21]. It requires more time for generating results.

2.4 Vacceed

Vacceed is a highly configurable architecture designed to perform high throughput *in silico* identification of eukaryotic therapeutic vaccine candidates. *In silico* vaccine candidate determination is validated or works accurately only in a laboratory. Vacceed potentially reduces the false vaccine candidates that are selected for laboratory validation thus saving time and money. Like NERVE, Vacceed is also a modular, scalable program designed only on command-line and Perl based syntax [25].

2.4.1 Vacceed Features

WoLf PSORT[44] and TargetP [45] are used for subcellular localization prediction. TMHMM is applied [46] for determining number of transmembrane helices but without setting <2 number of helices. Proteins comprised of >2 transmembrane helices are not considered as good for PVC prediction, because these transmembrane helices regions fails to clone [5]. Vacceed only determines T-cell binding prediction sites, and MHC I-Binding and MHC II-Binding prediction algorithms have been used for this purpose. BLASTp is used for determining non homologous host-pathogen detection. Essential genes and pathogenicity are not clearly defined that which tools have been used for this purpose.

Vacceed takes input of 'thousands of protein sequences'. A machine learning algorithm has been used to obtain output. A sequential protein candidates list in text file (send through e-mail) is its main output. Output is also generated in log file which is sent through e-mail.

2.4.2 Vacceed Limitations

Highly efficient, scalable, and configurable, Vacceed provided limited information on pathogenicity and putative functional genes which are the main parameters for detecting the potential vaccine candidates.

2.5 Jenner-predict

A web server, Jenner-predict, has been designed for predicting potential vaccine candidates with prior base methodology from proteome. The web server Jenner-Predict outperformed the methods of VaxiJen, Vaxign and NERVE. This web server consider

both adhesive and non-adhesive known-functional domains as PVCs in the prediction of non-cytosolic proteins (such as: porin, adhesin, choline-binding, flagellin, virulence, invasin, colonization, fibronectin-binding, toxin, penicillin-binding, solute-binding and transferring-binding). These domains are involved in host-pathogen interactions for the web server prediction accuracy. Jenner-predict thus out performed VaxiJen, Vaxign, and NERVE due to their non-consideration of un-adhesive proteins as PVCs. The web-server has identified maximum known PVCs successfully belongs to various functional classes [40].

2.5.1 Jenner-predict Features

Jenner-predict uses PSORTb 3.0 for predicting protein subcellular localization [47]. HMMTOP 2.0 is applied for determining number of transmembrane helices [37]. Immune Epitope Database (IEDB) is used to determine B-cells and T-cells epitope binding by checking PVCs against IEDB epitopes [43]. BLAST is used for predicting non homologous host-pathogen. Jenner-predict also focused on essential and un-adhesive proteins in PVC prediction.

Jenner-predict is designed on html and Perl scripts and its utility is tested on reported known vaccine candidates from *Escherichia coli* and *Streptococcus pneumoniae* proteomes. It takes proteome/protein sequences in FASTA format. Perl is used at the backend, posts the user provided sequence to the main program in queue. User may bookmark the queued job for tracking and obtaining the results. After the completion of job, a tabular form output is represented.

2.5.2 Jenner-Predict Limitations

It showed better performance than Vaxign and VaxiJen, Jenner-Predict even though efficient server, but elapse with time limitation problems. As a web based server, need internet connection and book-mark queued job may require some time for generating results.

2.6 Summary

All the available pipelines are designed for the prediction of therapeutic/potential vaccine candidates, but all have some limitations. Still not a single pipeline is designed which covers all the aspects necessary for efficiently and quickly predicting PVCs required to eliminate/eradicate infectious disease.

The advantage of web based pipelines (Vaxign, VaxiJen, and Jenner-predict) is that they are ready to use. Along with this ace, these also have file size and other limitations issues.

Stand-alone, modular, and configurable pipelines (NERVE and Vacceed) are designed on command-line interface. Which is not easily understandable by an immunological researcher. NERVE software/tool emphasized just on adhesive and extracellular proteins presuming that adhesin likeliness of extracellular proteins are potential vaccine candidates [21, 31]. But several non-adhesin putative functional proteins can also participate in host-pathogen interactions (i.e., porin, flaggelin invasin, etc.), and most of them are pathogenic as well as antigenic [40]. Whereas Vacceed provided limited information on pathogenicity and putative functional genes which are the main parameters for detecting the potential vaccine candidate [25].

Thus, still there is a demand of designing such a platform/pipeline which covers all aspects including user friendly, stand-alone, modular, multi-mode, multi-functional, scalable, configurable, extendable, time and cost efficient.

CHAPTER 3: Methodology

3 Methodology

In silico subtractive reverse vaccinology approach plays a key role in determining therapeutic bioactive agents for drug discovery process. This approach is usually applied for the therapeutic target identification [48].

The purpose of this study is based on *in-silico* subtractive genomics and reverse vaccinology strategy for designing a modular, configurable, and scalable pipeline for therapeutic/potential vaccine candidates (TVCs/PVCs) prediction; the strategy was already used by Barh [26] and Naz [27] in their research work. This technique was also followed in another study by Wilson *et al* [5] with some modifications.

Subtractive reverse vaccinology is a stepwise filtration process which starts from whole genome sequences and at the end, a handful significant proteins as targeted vaccine candidates are achieved.

Stepwise Subtractive Genomics Reverse Vaccinology Procedure

1. Screen out proteins of selected sub cellular localizations
2. Screen out the potentially functional/essential genes
3. Identify the non-host homologues proteins
4. Identify the virulence factors and screen out virulent proteins
5. Screen out the proteins having Transmembrane helices < 2
6. Perform protein functional annotation
7. Screen out proteins having molecular weight < 110 kDa
8. Perform epitope analysis and verify antigenic behaviour of peptides

The purposed methodology for developing a modular, configurable, and scalable pipeline to efficiently predict TVCs/PVCs is divided into 6 main sections.

3.1 Functionality Identification

VacSol functionality flow diagram is elaborated in the following schematic flow diagram.

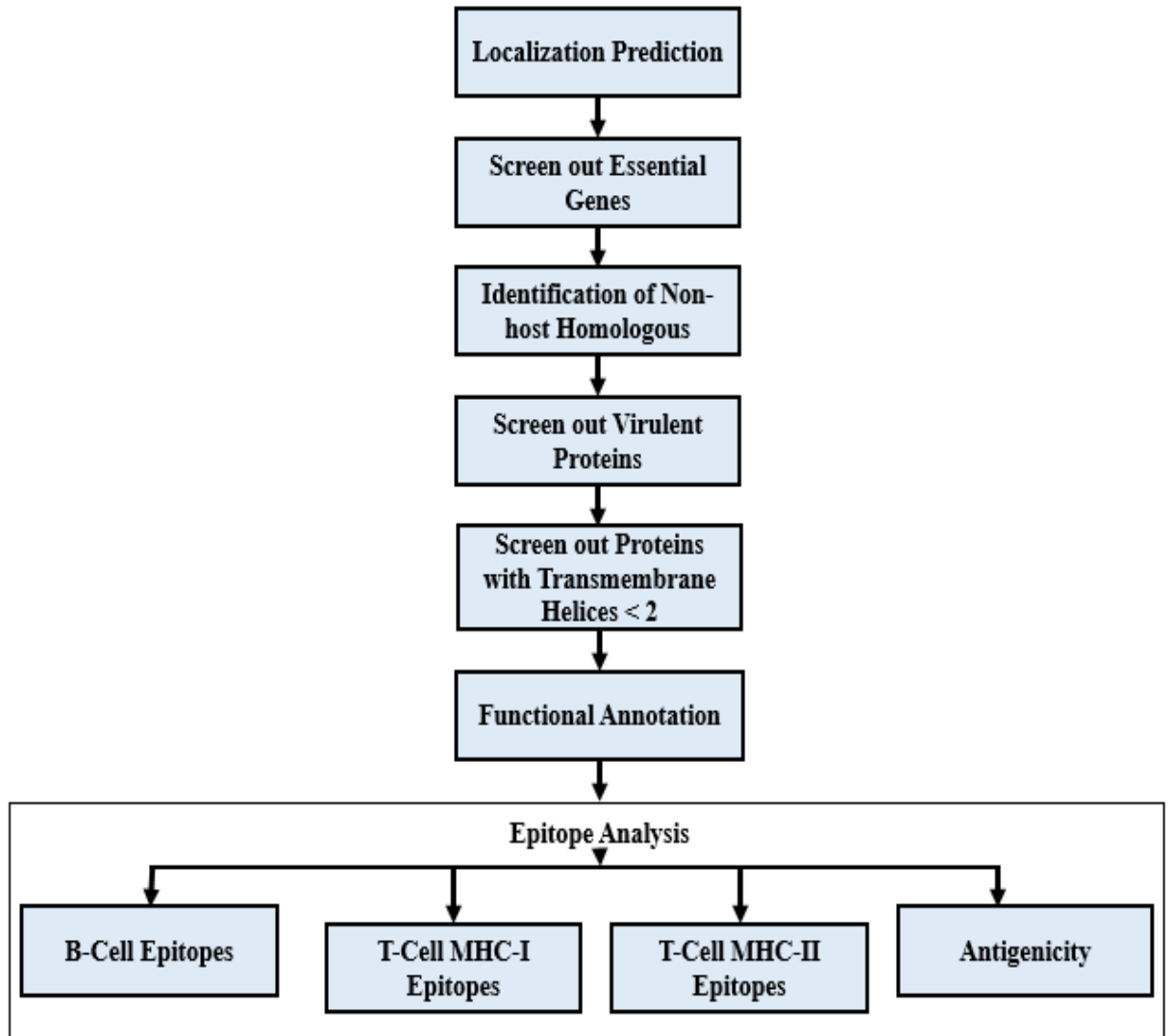


Figure 3.1: Schematic flowchart of Pipeline Functionality. Stepwise demonstration of proposed pipeline to prioritize the proteins as putative vaccine candidate (PVC). The approach include identification of sub-cellular location, essential genes, non-host homology, virulence factors, transmembrane helices and epitope binding.

3.1.1 Localization Prediction

Localization prediction refers to identification of protein sub-cellular localization (SCL) which is a computational method to predict the localization of functionally distinct compartments of membrane bounds. Protein localization prediction is important to understand the functionality as well as position of protein in cell. It is necessarily considered for the prediction of PVCs [5]. In identification of vaccine drug targets, the protein subcellular prediction on the cell surface is of particular interest, because these are mainly involved in pathogenesis. Extracellular or secreted proteins are readily accessible to antibody as compared to intracellular proteins and therefore represent ideal vaccine candidates [49]. Thus, localization prediction substantially contributes to enhance the PVCs identification process [50].

3.1.2 Screen out Essential Genes

Essential/potentially functional genes are considered as necessary for the survival of an organism. Different organisms have different essential genes. For vaccine/drug target identification against particular bacteria, the essential genes prediction of those bacteria is an important parameter. Thus, essential gene detection is an important part in furnishing novel TVCs/PVCs targets for the therapy against infectious diseases [51]. Screening out of essential genes reduces the set of target proteins which definitely impact on the cost and time of TVCs identification by separating out the non-essential genes from given set of proteins for further investigation [5]. Database of Essential Genes (DEG) was used for the prediction of essential genes [52].

3.1.3 Identification of Non-host Homologous

One of the important point which should kept in mind for TVCs/PVCs prediction is that vaccine targets should not be host (human) homolog; homology with host is considered to avoid the autoimmunity. These targets should not be homolog to non-pathogenic bacteria to protect human gut flora [5]. For this purpose, BLASTp was used [53].

3.1.4 Screen out Virulent Factors

Molecules which are produced by pathogens and are capable enough to produce pathogenicity of the organism and colonialize in the host organism cell are playing a vital role in causing diseases [54]. Identification of such virulent factors is very important and

necessary in the drug discovery process. Mostly pathogenic essential genes are virulent, thus this essentiality and virulence check is key factor to predict pathogens target proteins that lead to prioritize the proteins for vaccine candidates prediction [5].

Essentially validated genes were further analysed for confirming their virulence role through the use of MvirDB [55] and VFDB [56].

3.1.5 Screen out Transmembrane Helices

Transmembrane segment is actually a thermodynamically stable three-dimensional (3D) protein structure in membrane, which includes a single alpha helix, beta barrel, or any other structure. Transmembrane helices length is usually of 20 amino acids.

Proteins Transmembrane helices prediction step have importance in vaccine target identification process for predicting TVCs. Proteins with transmembrane helices less than 2 are considered as best candidates for therapeutic vaccine target identification, because more than one transmembrane helices in a protein make it difficult to colonize and express, as well as multiple transmembrane helices fail to purify recombinant proteins for vaccine development [5]. More than one transmembrane spanning regions presence was the main cause to fail 250 out of 600 vaccine candidates from cloning and expression [16]. HMMTOP was applied for determine transmembrane helices by using its parameters default values, HMMTOP is based on hidden Markov model (HMM) algorithm [37, 57].

3.1.6 Functional Annotation

Functional annotation is the step to characterize the protein functionality based on its sequence or similarity with other proteins. UniProt is used for the functional annotation of proteins [39]. This feature will help to understand that which functionality of pathogen can be ceased by targeting the putative proteins.

3.2 Tools Selection

Various well performed and currently available tools were selected and integrated in VacSol pipeline depending on some criteria for achieving its best performance and to outperform the previous pipelines. Previous pipelines used some of these but older versioned tools, and not any previous pipeline used all the tools that are integrated in

VacSol pipeline. Comparison of VacSol with already available PVCs prediction pipelines is mentioned in Table 3.2.

3.2.1 Selection Criteria

1. A criterion for tools selection was to develop pipeline based on the academically accepted tools which were well cited by research community and have proven authenticity.
2. The second criteria for tool selection were their standalone version availability and feasible integrate-ability along with other tools.

3.2.2 Tools Used

Table 3.1: Tools used in VacSol. Describing the currently integrated tools in VacSol

Name	Version	Function
BLAST+	2.2.27	New command line sequence alignment application which had been developed on NCBI C++ toolkit [58]
Pftools	2.3	Package of programs supporting search method of generalized profile format[59]
PSORTb	3.0	Protein subcellular localization prediction tool [47]
HMMTOP	2.0	Transmembrane topology prediction tool[37]
DEG	10.0	Database of essential genes[52]
MvirDB	N/A	Database of protein toxins, virulence factors and antibiotic resistance genes [55]
VFDB	N/A	Virulence factors database [56]
FastaValidator	1.0	Open-source Java library to parse and validate FASTA formatted sequences[60]
ABCPred	N/A	B-Cell epitope prediction tool [61]
Propred-I	N/A	Prediction of promiscuous MHC Class-I binding sites[62]
Propred	N/A	Predict MHC Class-II binding regions in an antigen sequence[63]
UniProt-SwissProt	N/A	Manually annotated protein sequences database with information extracted from literature and curator-evaluated computational [39, 59, 64]

Table 3.2: Comparison of VacSol with already available PVCs prediction pipelines.

Features	Tools					
	VacSol	NERVE	Vacceed	Vaxign	VaxiJen	Jenner-predict
Subcellular Localization	✓	✓	✓	✓	This is basically designed for antigen prediction, so only focuses on B-Cells and T-cells detection.	✓
Host-Pathogen non-homology Detection	✓	✓	✓	✓		✓
Essential Gene Prediction	✓	✓	Not clearly defined	✓		✓
Virulence Detection (of adhesive and non-adhesive proteins)	✓	Only for adhesive proteins	×	Mentioned for adhesive proteins		✓
Transmembrane Helix Prediction < 2	✓	✓	Without setting criteria<2	Not all proteins can be analyzed		✓
Functional Annotation	✓	×	×	✓		×

Epitope Mapping prediction (Both B-Cells & T-Cells binding with both MHC-I & II classes)	✓	Not clearly defined	Only T-cells binding	Only T-cells binding		✓
Input Ways	4 ways	2 ways	< 4 ways	2 ways		Provide sequence to the main program in queue
Output Ways	Output in 5 different formats	MySQLt able and textfiles	Text file and log file	MSExce l file		Tabular form output
Internet Dependent/ Web based	✗	✗	✗	✓	✓	✓

Detailed information regarding different tools.

3.3 Design Interface

Interface was designed for the preparation of the pipeline. The purpose of this Interface was to develop a clear outline and describe how these tools are connected with each other. Here each tool has its unique importance and takes input in specific format, and generate output on the basis of its specific input format and different from other tools input/output format. This was mandatory to identify and understand the input formats of each tool, so that the output of one tool can be converted in such way that the upcoming tool easily accepts that output. Because unordered stepwise subtractive genomics reverse vaccinology procedure require connection among different tools to prioritize the proteins for TVCs/PVCs prediction, and this connection could only be achieved through parsing of one format to other format for generating final results. This interface has been designed on four different modules tha have mentioned in Figure 3.2.

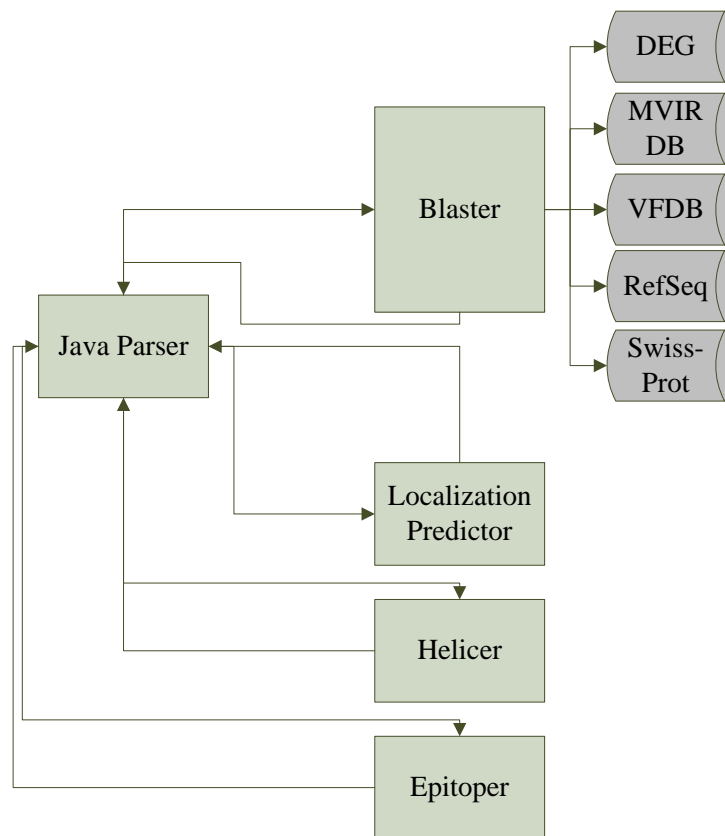


Figure 3.2: Modular-based VacSol Interface Design. VacSol is basically designed on four different modules. (i) Blaster: a module for predicting homology using Blastp, (ii) Localization Predictor: for the prediction of subcellular location, (iii) Helicer: to predict transmembrane helices, and (iv) Epitoper: a module designed to predict epitope mapping.

3.4 Architecture Design

3.4.1 High Level Architecture Diagram

High-level architecture diagram of purposed pipeline is given below which depicts the modes of working and the high-level communication process of VacSol pipeline. This diagram does not explain internal functionality of each process.

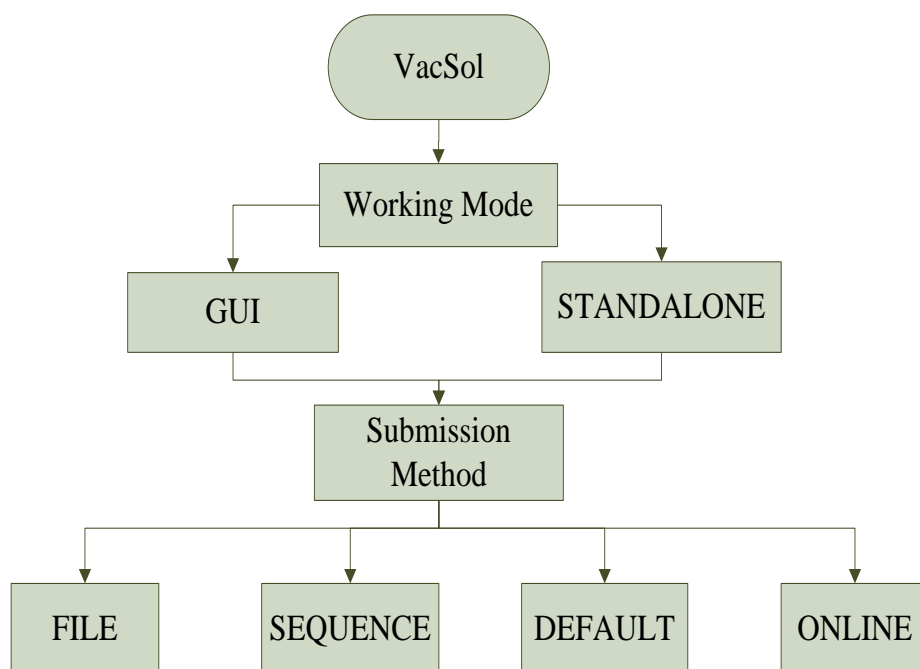


Figure 3.3: Working modes of VacSol. VacSol pipeline is composed of two working modes (1) Graphical User Interface (GUI) and (2) Standalone mode. This pipeline is of flexible nature as it accepts input in four different ways.

This architecture diagram clearly describes that pipeline is working in two different modes. One is GUI mode and other is STANDALONE mode. If user selects GUI mode, then pipeline opens a window offering interface to the user to interact with the pipeline, and requires input from user. But when user selects STANDALONE mode, it will require input in an XML template format. This mode is very helpful for further integration of pipeline with any other tool. Our purposed pipeline will be very beneficial for a

biological/immunological researcher because it can be dually used as a single tool or as pipeline to predict PVCs.

In spite of working mechanism of pipeline, whether it is a GUI or STANDALONE mode, it offers user to provide input in FASTA format in four different ways

- File
- Sequence
- Default
- Online

Pipeline validates the input format through Fasta Validator, and validated input is subjected to main process for the prediction of vaccine targets. Main process is comprised of threading options, and one can run as many threads as maximum number of cores available in the system, in general, this pipeline can work in parallel way. Single protein sequence is processed by a single thread to identify the prioritized sequences based on the provided input parameters. Sequence prioritizing process is performed in a number of steps to prioritize the input sequences that is clearly elaborated in figure 3.5.

After processing all the sequences in input file, prioritized sequences are subjected for further processing of epitope mapping which is used to determine the antigenicity of proteins. All the prioritized sequences after checking the epitope mapping, again sent to threading pool processing for generating the final results. Final results are generated in five different formats (FASTA, XML, JSon, HTML, and PDF format), ensuring the expandability and scalability of the aimed pipeline.

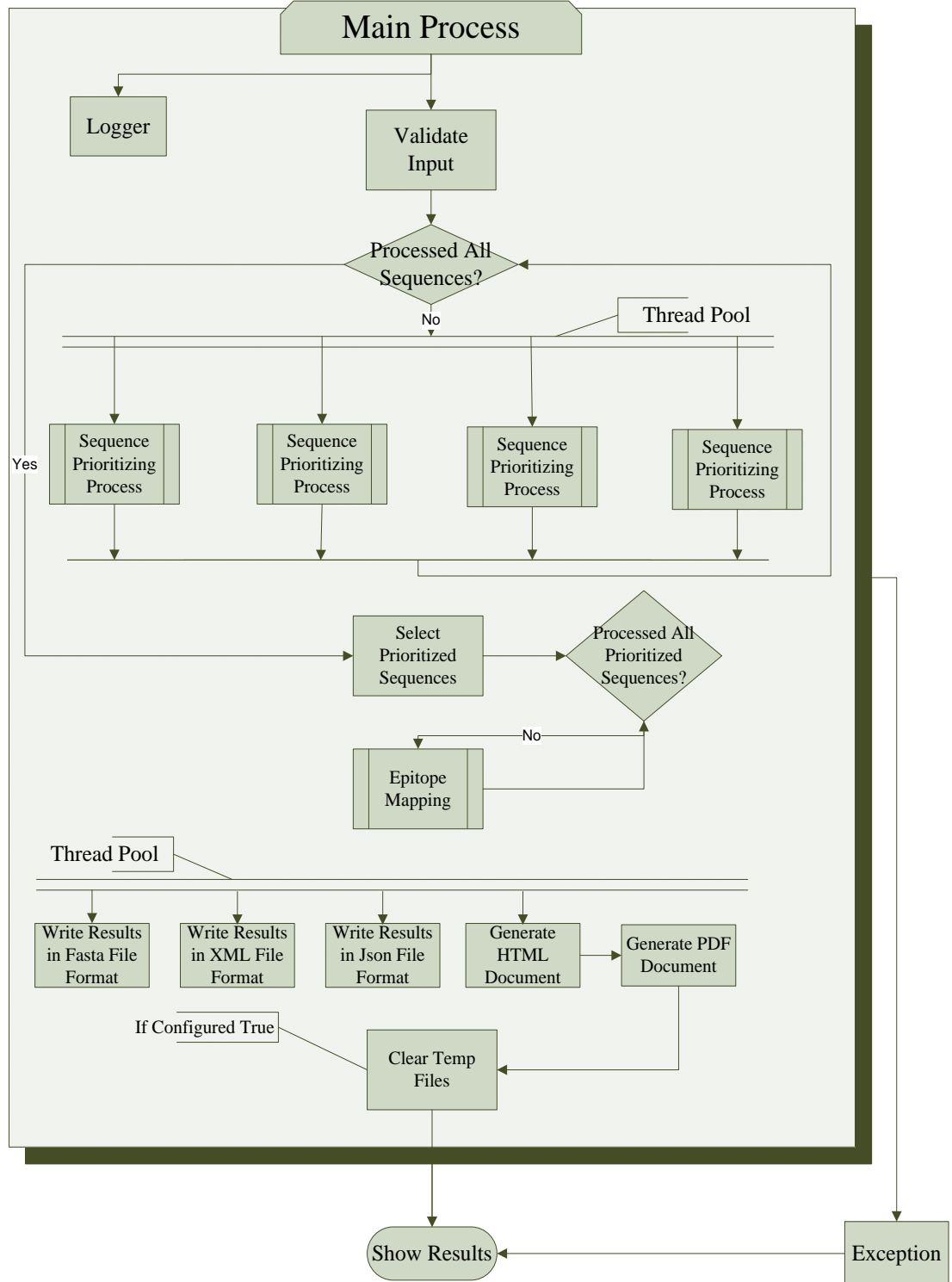


Figure 3.4: Schematic view of VacSol Main Processing. Inside processing procedure of VacSol is demonstrated in this flow diagram.

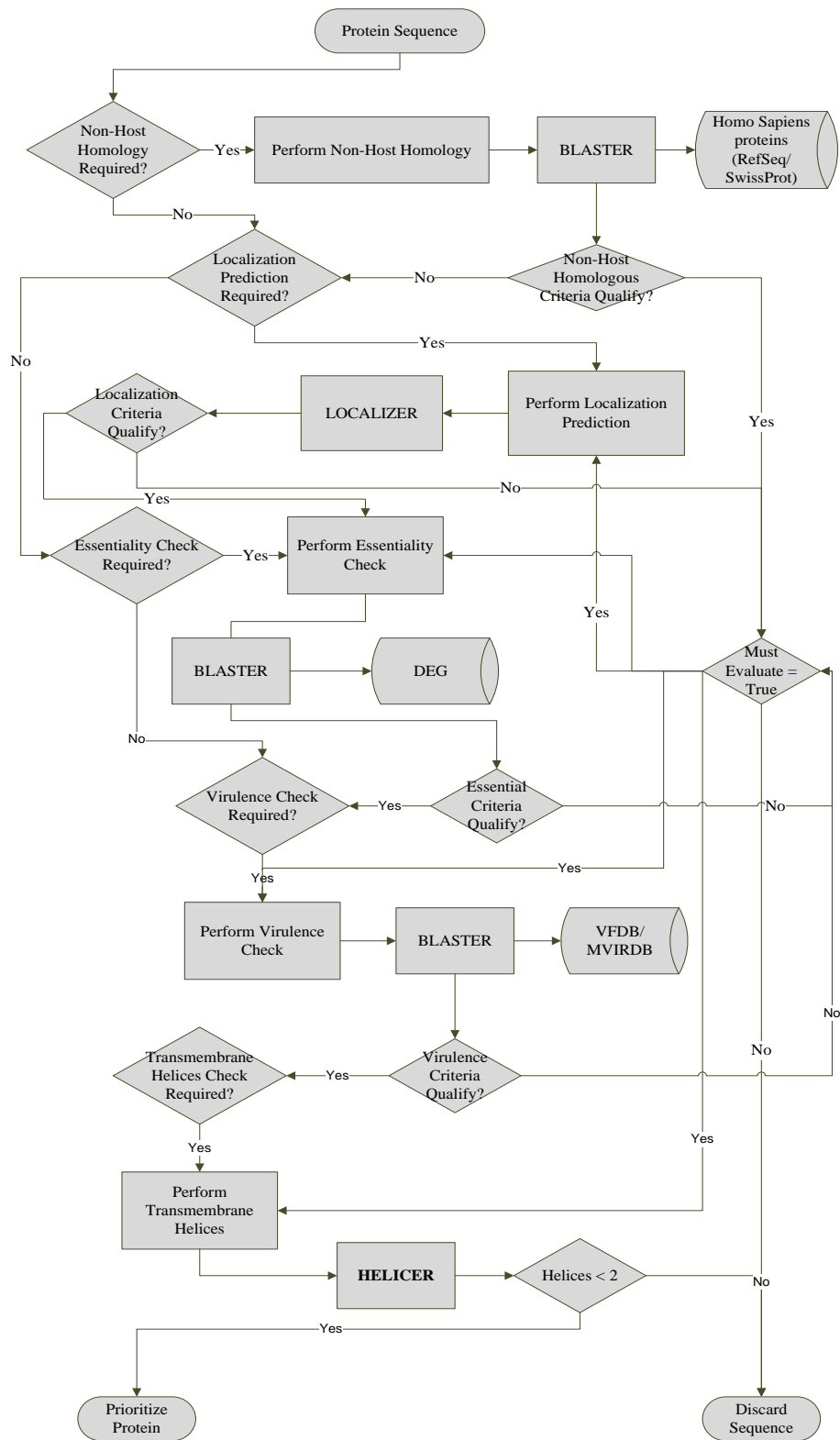


Figure 3.5: Elaboration of in-depth processing of prioritized protein thread. This clearly demonstrates the procedure of prioritizing process.

The purpose of VacSol is to offer its users a multi-dimensional pipeline through which they can use separately Blaster, Helicer, Localization predictor, and Epitoper modules according to their requirement, as well as they can run all the tools simultaneously for screening out the prioritized proteins. Figure 3.5 demonstrates the entire sequence prioritizing procedure during threading pool processing, and describes how to separately utilize different tools.

Different checks have been decided for screening out the proteins as vaccine targets. According to the above flow diagram, only those proteins will be considered as priority which crosses all the checks by following the stepwise reverse vaccinology procedure. User provided protein sequences are assessed by performing BLAST against RefSeq and SwissProt databases by selecting Bit Score > 100, E-Value < $1.0 e^{-5}$, and percentage identity > 35% which is the normal criteria for the sequences to lie in homology safe zone [65]. Basically at this point, blaster module is used to find out the matched sequences between pathogen and host, homologous sequences are not considered and the result is generated on the basis of remaining (un-matches/non-homologous) sequences. If non-host homology criteria do not qualify, the results will not be generated; but VacSol offers users either to use other tools or not. After qualifying the non-host homology criteria, user must evaluate the tools for further processing. Proteins sequences were subjected further to the next module (localization predictor).

If localization prediction criteria qualified (means matches the sequences with PSORTb) through localizer and blaster modules, the matched sequences are delivered to next step (Essentiality check) for qualifying, and fulfilling the prioritizing protein requirements. If localizer does not qualify the localization prediction criteria (none of the sequences obtained in extracellular and cell surface location), again VacSol offers the users to use other modules by checking the 'Must Evaluate' check box. If 'Must Evaluate' check is checked then the pipeline must evaluate the selected steps whether previous step pass the criteria or not; but if 'Must Evaluate' check is not checked then subsequent step will not proceed if its preceding step fails to meet the input criteria. For example, if a protein is not essential then VacSol skip steps of virulence, transmembrane helices etc.

Essential proteins are searched out by performing the BLAST search of the sequences against DEG database through the use of Blaster module. Only those sequences are subjected to determine virulence factor which qualify the essentiality criteria. If none of the sequence passes this (essentially criteria) check, and even user wants to check the virulence factor, then 'Must Evaluate' criteria should be checked by user. Similarly virulence factor is determined by using VFDB and MvirDB through the Blaster module, and sequences with virulence factors are further processed for determining the number of transmembrane helices. If none of the sequence is obtained with virulence factor, again user is offered to either find number of transmembrane helices or not. If user wants to determine the number of helices without the detection of virulence factor, he/she must check 'Must Evaluate'. Those proteins having number of transmembrane helices < 2 as well as fulfill all the previous criteria are screened out and considered as the prioritized proteins.

Here discarded sequence referred to those sequences which do not pass all steps criteria, and only those are screened out and displayed which pass all the steps of applied strategy.

Epitoper Processing

When prioritized sequences are screened out, then pipeline passes these sequences to epitope mapping process which is carried out by performing a number of steps. Epitope mapping is further used to identify the antigenicity of protein epitopes. Figure 3.6 display the inside VacSol processing for epitope mapping, which describes that prioritised proteins are subjected to calculate number of B-cell epitopes through the processing of Epitoper module (named on the basis of epitope mapping). All the prioritised proteins after calculating number of B-cells are subjected to predict T-cell epitopes. Those T-cells epitopes which bind with maximum number of both MHC-I and MHC-II class molecules are extracted. VacSol also provides the option to check antigenicity with the availability of internet connection, and top probable antigenic epitopes are displayed in final result. Final results are generated in five different formats which are explained in Figure 3.4.

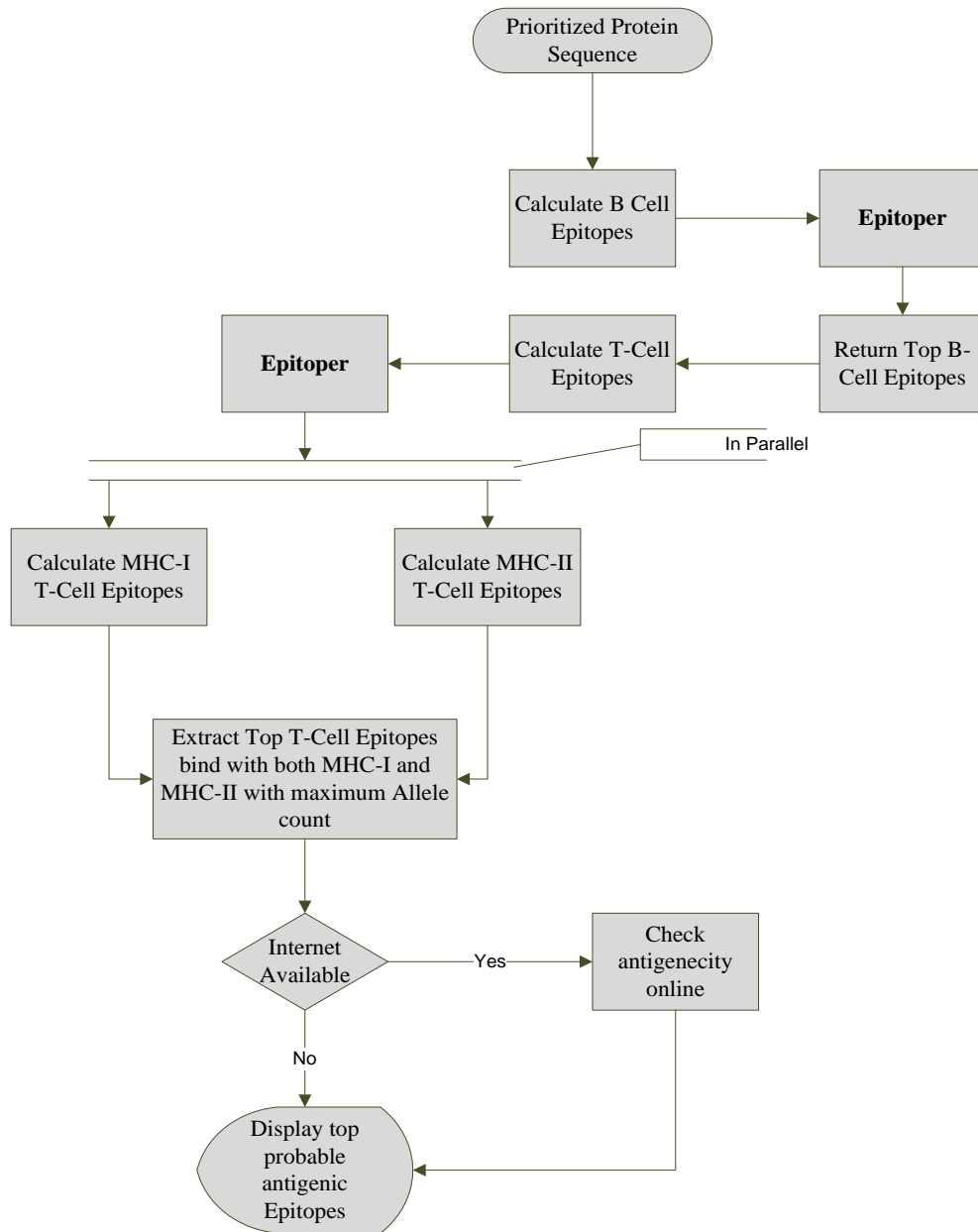


Figure 3.6: Schematic View of Epitope Mapping through VacSol.

3.5 Implementation

Pipeline is implemented in Java 1.7 in Eclipse IDE (Integrated Development Environment). Following libraries mentioned in Table 3.3 were required for the integration of tools mentioned in Table 3.1 for performing different functional requirements to screen out vaccine candidate targets. Besides these libraries, a few other

were also included in the project to resolve the dependency issue of integrated functionally required libraries.

Table 3.3: Java Libraries used for VacSol.

Libraries	Version	Purpose
FastaValidator.jar	1.0	Validate Fasta file format
gson-2.3.1.jar	2.3.1	For Java to JSon conversion
xstream-1.4.7.jar	1.4.7	From XML to Java conversion
itextpdf-5.4.2.jar	5.4.2	To generate Pdf files
jsoup-1.8.1.jar	1.8.1	To generate and manipulate HTML templates
log4j-1.2.17.jar	1.2.17	Logging

3.5.1 Packages Designed for Organized Pipeline Development

Different packages have been designed for the separation of concerned and to organize pipeline code.

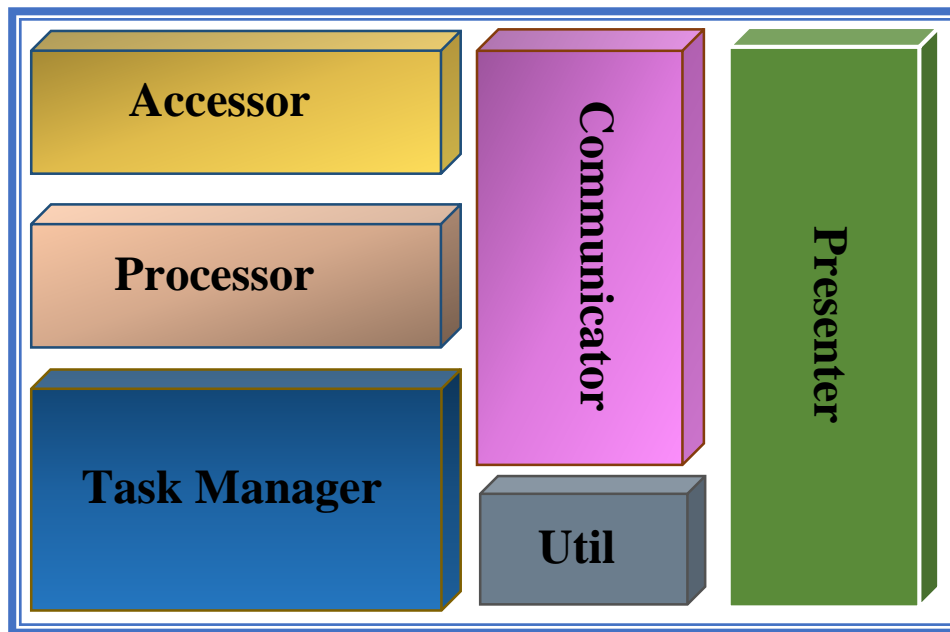


Figure 3.7: Overview of Package Diagram. VacSol is comprised of six packages that have been designed to accomplish its developmental task.

Figure 3.7 displays the VacSol packages that have been described in following:

3.5.1.1 Accessor

Accessor package is responsible for controlling the incoming and outgoing access of pipeline such as: performing I/O operations and dealing with file reading and writing, as well as handling *http* request and response objects. It is designed to accept user input and delivers it to processor, and takes processor output to generate and display the results through presenter.

3.5.1.2 Communicator

Communicator is responsible to deal with data transfer objects (DTO) [66]. This package contains classes that named with the same scientific/logical words such as: Genome, Allele *etc.* The purpose of these classes are just to transfer data objects between two different layers or packages, and behaving as a communication bridge between these packages for functionality.

3.5.1.3 Presenter

This layer/package contains the classes for designing the graphical user interface (GUI) offering the user to interact with inside pipeline functionality.

3.5.1.4 Util

This package handles the utility classes. These classes performs functionality for configurations, logging handling, exceptions handling, contains list of values *etc.*

3.5.1.5 Task Manager

This package is logically merged with processor package. The classes in task manager are just to organize and initiate thread pools, means task manager is responsible for managing thread pool functionality.

3.5.1.6 Processor

This is the main package containing process classes, which is actually related to processing functionality. Invoke tools integrated in this pipeline, parse their results, perform processing and controls the functionality of other tools

CHAPTER 4: Results

4 Results

After the implementation of proposed pipeline for the prediction of therapeutic/putative vaccine candidates, VacSol was applied on *H. pylori* proteome for utility testing. *H. pylori* strain 26695 (RefSeq accession No. NC_000915) is comprised of 1,576 proteins or coding regions [5, 67]. VacSol screened/searched out only 10 protein sequences as therapeutic/putative vaccine candidates from this gastric pathogen. Each tool implemented in VacSol pipeline requires specific input parameters mentioned in Table 4.1, for determining antigenic proteins.

Table 4.1: VacSol Input Parameters.

Function	Values
Submission Method	File
Must Evaluate	True
Homologous with Human	Bit Score > 100, E-Value < 1.0 E ⁻⁵ Percentage Identity > 35% Blastp with RefSeq and Swissprot (Homo-Sapiens protein sequences) database
Virulence Factors	Bit Score > 100, E-Value < 1.0 E ⁻⁵ Percentage Identity > 35% Blastp with (MVIRD and VFDB) database
Essential Genes	Bit Score > 100, E-Value < 1.0 E ⁻⁵ Percentage Identity > 35% Blastp with DEG database
Localization Filter	Extracellular and Outer Membrane
Epitope Maps	B-Cell Threshold: 0.4 Windows Length: 20 T-Cell Threshold: 0.4 Top Peptides: 4

Table 4.2: Therapeutic/Putative Vaccine Candidates (TVCs/PVCs). 10 out of 1576 protein

Total Sequences	Non-Homologous with Human	Subcellular Localization	Essential	Virulent	Transmembrane Helices < 2	Prioritized Protein
1576	1452	65	667	267	1254	10

sequences were extracted as TVCs/PVCs through prioritized procedure by VacSol.

H. pylori 1,576 proteins were subjected for PVCs prediction through VacSol, which screened out 1,452 non-human homologous proteins. 65 proteins extracted out as secretome and exoprotome (lies on cell surface and extracellular proteins). 667 results were obtained for essential genes prediction, 267 proteins were screened as virulent, and 1,254 proteins were obtained with transmembrane helices < 2. Finally 10 out of 1,576 proteins were screened as prioritized proteins (which was the main goal of VacSol development) that were considered as TVCs/PVCs.

4.1 Localizer Results for Subcellular Localization Prediction

Localizer classify subcellular protein localizations and display in a pi-chart as shown in Figure 4.1

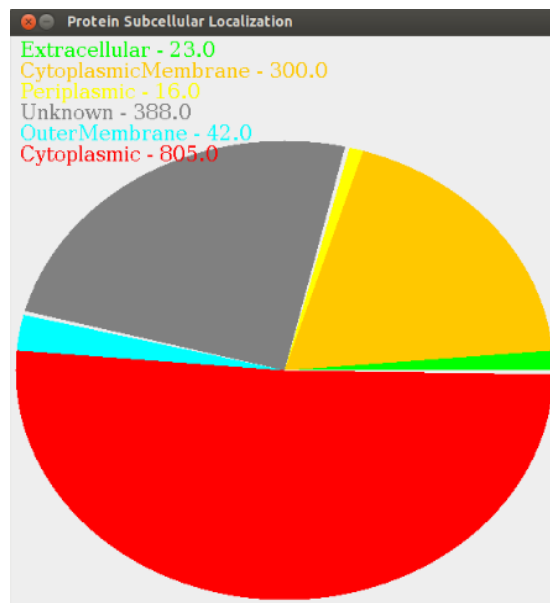


Figure 4.1: Graphical Representation of Localizer Results. Localizer module generated results that are demonstrated in graphical style. Different protein locations were coded with different colors. Red color describes cytoplasmic proteins. Grey color represents unknown/ hypothetical proteins; Yellow indicates the proteins that reside in periplasmic region. Cyan color represents outer membrane proteins, and green is used for extracellular proteins.

According to applied epitope parameters. The 20 mers B-Cell epitopes of these prioritized proteins, and the respective top four T-Cell peptides (with both MHC-I and MHC-II allele counts) calculated by VacSol pipeline, and sorted based on the calculated log scores. According to VaxiJen all these peptides are probable antigen.

4.2 List of VacSol Screened out Proteins

VacSol screened out ten proteins as mentioned in Table 4.3.

Table 4.3: VacSol Screened out Putative Vaccine Candidates. All the screened out prioritized proteins as putative vaccine candidates are mentioned in this table.

Sr. No	Protein Name	Locus Name/ Gene Symbol
1	fecA	HP1400
2	flagellin A (flaA)	HP0601
3	Putative beta-lactamase	HP1098
4	Iron(III) dicitrate transport protein (FecA)	HP0807
5	Flagellin B (flaB)	HP0115
6	toxin-like outer membrane protein	HP0289
7	Toxin-like outer membrane protein	HP0922
8	Beta-lactamase HcpA	HP0211

9	toxin-like outer membrane protein	HP0610
10	Iron(III) dicitrate transport protein (FecA)	HP0686

4.3 Online Cross Checking of VacSol Generated Results

VacSol prioritized protein results were further cross checked through available online tools to validate the VacSol utility, and to verify that this pipeline really fulfills the subtractive genomic reverse vaccinology criteria.

4.3.1 Sub-cellular Prediction Results

Screened out proteins by VacSol are also qualified for selected localizations using CELLO2GO and snapshot of the result is mentioned below in Figure 4.2.

Localization Prediction	Localization	Amount
	Extracellular	7
	Outer membrane	3
	Periplasmic	0
	Inner membrane	0
	Cytoplasmic	0

Figure 4.2: CELLO2GO Results. Protein subcellular localization prediction online server (CELLO2GO) results indicated that VacSol screened out prioritized proteins can be the putative vaccine candidates as these lie in extracellular and outer-membrane regions which are of great interest for TVCs/PVCs prediction.

4.3.2 Essential Genes Prediction

Screened out proteins by VacSol are also identified as essential genes using online DEG and results are mentioned in Table 4.4.

Table 4.4: Database of Essential Genes (DEG) Online Results. Same parameters were used as applied in VacSol. These results depicted that all the prioritized proteins can be the good PVCs, as these proteins are coded from pathogenic essential genes.

Parameters	Bit Score > 100 E-Value < 1.0 e ⁽⁻⁵⁾
Total protein-coding genes in your sequence:	10 genes
In your sequence, No. of genes having homologs with DEG:	10 genes
In DEG, the No. of genes having homologs with your sequence:	20 genes.

4.3.3 Virulence Factor Detection Results

Online VFDB results (coincide with the pipeline generated results) indicated that all prioritized proteins contain virulent factor, thus these results may helpful to generate the conclusion that 10 screened out proteins may be the vaccine targets.

4.3.4 Number of Transmembrane Helices Prediction

Screened out proteins by VacSol are also qualify the transmembrane helices less than two criteria using online TMHMM and results are mentioned in Table 4.5.

Table 4.5: TMHMM Results for Transmembrane Helices. Online TMHMM results verify that screened out proteins through VacSol might be the good PVCs, as all these proteins comprised < 2 number of helices.

Protein Name	Prioritized Protein ID	Gene Symbol	No. of Transmembrane Helices
fecA	3	HP1400	0
flagellin A (flaA)	285	HP0601	0
Putative beta-lactamase	534	HP1098	0

Iron(III) dicitrate transport protein (FecA)	825	HP0807	0
Flagellin B (flaB)	837	HP0115	0
toxin-like outer membrane protein	907	HP0289	1
Toxin-like outer membrane protein	995	HP0922	0
Beta-lactamase HcpA	982	HP0211	0
toxin-like outer membrane protein	1184	HP0610	0
Iron(III) dicitrate transport protein (FecA)	1359	HP0686	0

4.4 Protein Annotation Results

Screened out proteins by VacSol are then blasted at UniProt to know about the further details and results are mentioned in table 4.6.

Table 4.6: Functional Annotation of Prioritized proteins. Prioritized proteins were subjected for functional annotation by using different online tools and databases. All the following results best match to VacSol screened out putative candidates, *e.g.* the functional annotation (by using UniProt) and domain analysis (by using InterPro Scan) described that these proteins highly interact with pathogenic proteins such as: TonB, VacA, and Sell1-like proteins and involve in some important pathways.

Bacterial Proteins	Gene Symbol (NCBI)	Molecular Weight (kDa)	Molecular Function (UniProt)	Domains (InterPro Scan)	References
(fecA)	HP1400	94.827 kDa	receptor activity	TonB-dependent receptor & plug domain	[5, 68]
flagellin A	HP0601	53.284 kDa	Cell motility, Signal transduction, and structural molecule activity	Flagellin, Flagellin_D0/D1, & Flagellin_hook_IN_motif	[69-71]
Iron(III) dicitrate transport protein (FecA)	HP0807	88.946 kDa	receptor activity	TonB-dependent receptor & plug domain	[68, 72]

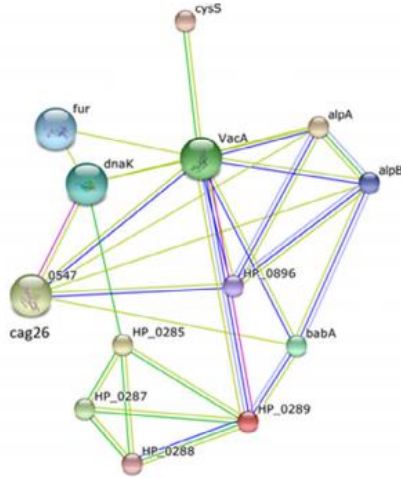
Flagellin B (flaB)	HP0115	53.882 kDa	structural molecule activity	Flagellin, Flagellin_D0/D1	[70, 73, 74]
toxin-like outer membrane protein (hypothetical)	HP0289	311.288 kDa	Not defined yet	Autotransporte_beta & Vacuolating_cytotoxin_put	[75]
Beta-lactamase HcpA	HP0211	27.366 kDa	peptidoglycan, involve in cell wall synthesis	Sel1-like, TPR-like_helical_dom	[76]
toxin-like outer membrane protein	HP0610	212.964 kDa	Not defined yet	Vacuolating cytotoxin putative & Autotransporter beta domain	[75]
Iron(III) dicitrate transport protein (FecA)	HP0686	87.698 kDa	receptor activity	TonB-dependent receptor, beta-barrel, plug domain	[77]
Toxin-like outer membrane protein	HP0922	274.563 kDa	Not defined yet	VacA2 (motif), Autotransporte_beta, PbH1	[78]
Putative beta-lactamase (HcpC)	HP1098	31.594 kDa	beta-lactamase activity	Sel1-like, TPR-like_helical_dom, TPR_2	[76]

Molecular weight analysis indicated that three proteins does not follow putative vaccine candidate criteria (as for best candidates, targeted proteins should be with mass <110 kDa). So protein-protein interaction was determined for these three proteins to check their involvement in pathogenesis.

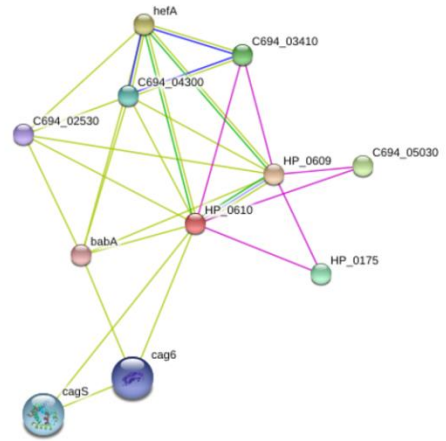
4.5 Protein-Protein Interaction Results

Three proteins which does not follow molecular weight < 110 kDa criteria are then further analyzed by using STRING and the snapshots of STRING results are shown in Figure 4.3.

Toxin like outer membrane protein (HP-0289)



Toxin-like outer membrane protein (HP-0610)



Toxin-like outer membrane protein (HP-0922) C694_02530

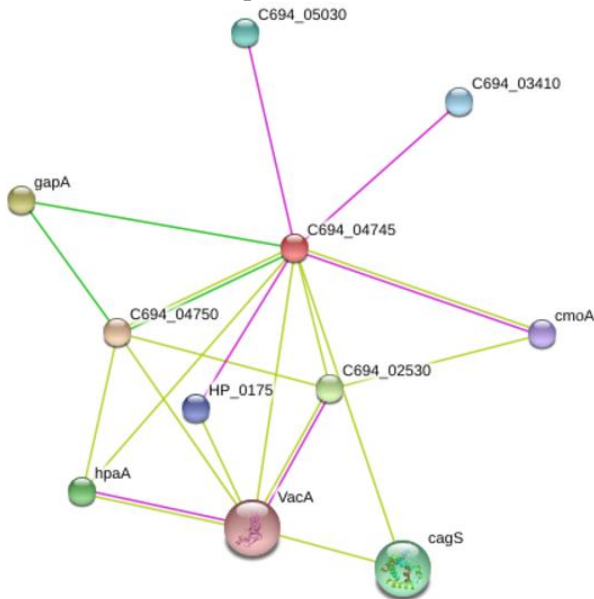


Figure 4.3: Interactive Depiction of Protein-Protein Interaction Network. STRING database was used to determine the protein-protein interactions. PPI network results clearly indicate that these toxin like outer membrane proteins highly interact with other virulent proteins such as *vacA*, *babA*, *cagS*, and *cag6* and might be involved in important pathogenic pathways.

CHAPTER 5: Discussion

5 Discussion

Infectious diseases are emerging globally, and have included the former diseases that still are posing immense threat to human health. Recent studies indicated that bacterial resistant strains lead to therapeutic failure [79], [5, 80] that can be the cause of high risk and re-infection [6]. Therefore, there is an urgent need to develop such type of vaccines or treatment that saves the human health against these resistant microbes. With the advent of bioinformatics approaches and high throughput sequencing technologies, reverse vaccinology technique was emerged [14]. Recently subtractive genomics is widely used to efficiently reduce cost and time in target screening for vaccine and drug discovery processes [81].

Various online tools and servers are available on reverse vaccinology, but these readily available pipelines requires days and even weeks for generating results [40, 41, 82]. Therefore, VacSol pipeline was used for inferring detailed inside information regarding gastric pathogen as it is a multifunctional, modular, configurable, and scalable pipeline. Due to its stand-alone version, user can gather inside pathogen information for predicting putative vaccine targets even without the need of internet connection. VacSol utility was tested on *Helicobacter pylori* (*H. pylori*) strain 26695 (RefSeq accession No. NC_000915) that contain 1,576 coding genes or proteins [83].

Whole *H. pylori* proteome was subjected for identifying prioritized proteins that would be therapeutic vaccine candidates. VacSol processed the whole *H. pylori* proteome for the identification of non-homology with human, subcellular localization prediction, essential gene prediction, detecting virulence factor, determination of proteins comprising < 2 number of transmembrane helices, and finally identification of epitope binding proteins. All these steps are necessary for prioritizing the proteins for TVCs/PVCs prediction [67] [5, 84].

VacSol screened out 10 proteins as prioritized that can be the putative vaccine candidates. Non-human homology was determined by performing the BLAST of the

sequences against RefSeq and SwissProt database to check out human homologous proteins by selecting Bit Score > 100, E-Value < $1.0 e^{(-5)}$, and Percentage Identity > 35% which is the normal criteria for the sequences lie in homology safe zone [65]. These homologous proteins were discarded and remaining 1,452 proteins were displayed by VacSol that were subjected for further processing. Subcellular localization prediction results displayed 65 proteins as secretome and exoproteome in which 23 lie in extracellular and 42 screened as outer-membrane proteins. These proteins are of great interest for the development of ideal vaccine candidates due to their feasible accessibility to antibody as compared to intracellular proteins [49, 85].

667 proteins were sorted out that are coded by essentially important genes for the survival of gastric pathogen (*H. pylori*). Out of these 667 proteins, 267 indicated as virulent proteins. Identification of essentially important and virulent proteins is an integral part in predicting the novel TVCs/PVCs targets for the therapy against infectious diseases [51]. *H. pylori* proteome was subjected for determining the proteins comprising < 2 number of transmembrane helices, and 1,254 proteins were explored that follow this criteria. Literature stated that more than one transmembrane helices in a protein make it difficult to clone and express [5].

Above described results were separately generated for extracting the inside pathogenic information by parallel VacSol processing. For prioritization process, VacSol worked sequentially (following subtractive reverse vaccinology procedure) and generated final results for 10 prioritized proteins. These proteins were further cross checked through available online tools and servers to ensure their putative vaccine candidacy. CELLO2GO results indicated (Figure: 4.2) that screened out prioritized proteins through VacSol can be the putative vaccine candidates as 7/10 proteins screened as extracellular and 3/10 as outer-membrane proteins [24] which are of great interest for PVCs prediction.

Prioritized proteins were assessed by performing BLAST against online DEG by selecting Bit Score > 100, E-Value < $1.0 e^{(-5)}$, and Percentage Identity > 35% as input parameters. DEG results (Table: 4.4) demonstrated that all these 10 proteins showed homology with 20 genes present in database of essential genes (DEG). Thus, according

to these results, it can be concluded that these 10 prioritized proteins might be the putative vaccine candidates.

Transmembrane Helices base on Hidden Markov Model (TMHMM) results mentioned in Table: 4.5 clearly display that all proteins comprising <2 number of transmembrane helices, interestingly leading to the conclusion that prioritized proteins might be integral part in vaccine development against gastric pathogens.

Protein annotation results (Table 4.6) describing the information of screened out proteins. Interestingly, these results giving some more prioritized information based on molecular weight information. Literature indicated that proteins with <110 kDa molecular weight are considered to be the suitable targets for vaccine candidates [5]. Seven proteins *fecA* (HP1400), *FecA* (HP0807), *FecA* (HP0686), *FlaA* (HP0601), *FlaB* (HP0115), *HcpA* (HP0211), *HcpC* (HP1098) were calculated with molecular weight less than <110 kDa by using Expasy ProtParam tools [86]. Whereas, three toxin like outer membrane proteins (HP0289, HP0610, and HP0922) comprised of molecular weight 311.288 kDa, 212.964 kDa, and 274.563 kDa respectively. On the basis of molecular weight results, it is concluded that 7 can be the best putative candidates, whereas HP0289, HP0610, and HP0922 might not be the therapeutic vaccine candidates. Surprisingly molecular weight results are somewhat varying than all the previous results, one of the reasons is that VacSol does not check molecular weight criteria. So, molecular weight criteria can be the future perspective of VacSol pipeline.

Toxin like proteins with molecular weight >110 kDa were further analyzed through their protein-protein interaction analysis through STRING to check their candidacy and involvement in pathogenesis. Protein-protein interaction (PPI) network results again indicated that VacSol sorted out proteins can be the good targets for putative vaccine and these results coincide with the previous results. PPI network for HP0610 (Figure 4.3) clearly displays that this toxin like outer membrane protein shows interaction with *babA* and *cagS* and *cag6* proteins. Literature indicated that *babA* is also a good vaccine candidate, whereas *cagS* and *cag6* proteins belong to *cag* family and these *cag* proteins form an island in *H. pylori*, termed as *cag* Pathogenicity Island. These *cag* proteins are pathogenic proteins, involving in some pathogenic pathways, so by targeting HP0610,

pathogenic pathway can be controlled, so this can be the putative vaccine candidate, but on the basis of molecular weight this protein is not taken under consideration. Similarly, other two extracellular proteins with >110 kDa mass may not be the therapeutic vaccine target, even though these interact mostly with virulent proteins. Literature stated that HP0289 interact with *babA* and *vacA* which are targets for putative vaccine candidates [5]. HP0922 PPI network indicated that this protein also interact with *cagS*, *hpaA*, and *vacA*. All these three proteins are involved in pathogenesis and different virulent pathways, *vacA* is directly involved in gastric ulcer [87]. On the basis of molecular weight criteria, and hypothetical protein (even their functions are still not known) these three proteins may not be considered as targets for vaccine candidates.

Iron(III) dicitrate transport protein (*FecA*) (HP1400, HP0807, and HP0686) interact with *TonB* protein which is involved in virulence. It is studied that controlled and mutated *TonB* lead to increase immunization [68], so by targeting HP1400, HP0807, and HP0686, *TonB* can be controlled, thus these three proteins can also be putative vaccine candidates.

Flagellin proteins (*flaA* and *flaB*) are responsible for the pro-inflammation of gastric mucosa that lead to gastric/peptic ulcer, thus *flaA* and *flaB* are good candidates for vaccine development [69]. Likewise, Beta-lactamase *HcpA* and *HcpC* are highly pathogenic proteins that are directly involved in different infections caused by *H. pylori* [76]. *HcpA* protein is also involved in bacteria and eukaryotic host interaction [88]. Thus, in short protein annotation results verify that VacSol screened only those proteins which can be good (even best) putative/therapeutic vaccine candidates.

Naz *et al.* studied on *H. pylori* which implicates that there are six putative vaccine candidates (two are of >110kDa mass), but VacSol describes for a few more PVCs as it extracted 10 proteins (7 are more reliable). *FecA* (HP0807) and HP0289 matched with that Anam's results [69]. VacSol screened *FacA* (HP1400, HP0807, and HP0686) proteins, whereas previous study is focused only on one type of *FecA* protein. *FecA* are iron(III) dictate transport protein which are mainly involved in transport and receptor activity as well as directly link to *tonB* protein, so by controlling three *FacA* proteins it is easy to attenuate the microbe. Previously studied six results were tested through VacSol, but 4 out of those six proteins did not pass the essentiality criteria. Surprisingly

one out of these four proteins also did not fulfill the localization criteria, and literatures states that *VacA*(even though is directly involved in hepatic ulcer) is > 110k Da protein [69].

Only two screened proteins have matched with studied by proteins. DEG versions difference might be the reason of not matching the remaining 4 proteins. While the other 8 screened out proteins by VacSol are highly involved in pathogenesis and are essential for *H. pylori* microbe, see the Result section for further details.

Thus, VacSol pipeline till date performs well with its flexible nature of input and output ways as it accepts input FASTA file in different ways (sequence, default, file, and online), and generate results in five different formats (Fasta, XML, JSon, HTML page, and PDF format). See Appendix A for further details.

5.1 Future Perspective

There are still more aspects to work on this pipeline in different perspectives

1. **Enhancement:** This pipeline can be enhanced and expanded by incorporating some more functionalities such as: Molecular Weight Calculation, Protein-Protein Interaction detection, and visual display of available protein structures in 3D model.
2. **Multi-Node:** Currently this pipeline is multicore parallel based on threading mechanism. This can be further optimized by adding multi-node feature. Multi-node feature can boost its performance and also make it capable of running on multiple computers in distributed fashion.
3. **Multi-Platform:** This pipeline is implemented in Java and currently tested only on Ubuntu 12.04 64 version. Further research can also make it platform independent and test on different operating systems with some minor efforts. Java is platform independent but all prerequisite tool are not platform independent.

Appendices

1 Appendix A: Installation Guide

VacSol itself does not require any installation. It is just an executable jar file. But it depends upon other standalone tools. VacSol functionality depends on the installation of various tools that behave as pre-requisite for this pipeline execution. VacSol has tested and analyzed to be fully functional on Ubuntu 12.04.5 64 bit. It is developed in Java which is platform independent. VacSol will work on any operating system, if its pre-requisite tools perform their entire functionality on other operating systems. Current guide is intended for Ubuntu 64 bit desktop environment. User may test it on other environments but should take care of compatibility issues and configurations of all prerequisite tools and environment set.

1.1 Pre Requisite Tools/Languages

- PSORTb
- NCBI Blast+
- Perl
- Bioperl
- Pftools
- Hmmtop
- ABCPred
- ProPred-I
- ProPred
- Java

Above mentioned tools have their own installation requirements. It is mandatory to install these tools correctly and successfully.

Before installing any tool, please open terminal using ALT+CNTRL+T and run following commands.

```
sudo apt-get update
```

```
sudo apt-get install gcc build-essential
```

1.1.1 PSORTb Installation

PSortb Installation instructions are available at following mentioned url <http://www.psort.org/downloads/INSTALL.html> which is the primary source of PSORTb installation guide.

We have also mentioned below the installation steps for the ease of user, so that he/she can easily install PSORTb.

1.1.1.1 PSORTb Prerequisites

PSORTb need several prerequisites that must be installed for a fully functional version.

1.1.2 Perl (5.6.X or higher)

Install Perl using command or get latest version of Perl from <http://www.cpan.org>

```
sudo apt-get install perl
```

1.1.3 Bioperl (1.2.X or higher)

Install Bioperl using command or Bioperl can be obtained from <http://www.bioperl.org>

```
sudo apt-get install bioperl
```

1.1.4 Stand Alone NCBI Blast

Following commands are available for install NCBI Blast or it can be downloaded from NCBI FTP site at <ftp://ftp.ncbi.nih.gov/blast/executables/>

```
sudo apt-get install blast2
sudo apt-get install ncbi-blast+
```

1.1.5 PFTOOLS

Download pre compiled pftools executable from <https://goo.gl/3hqQof>.

User is suggested to kindly create a folder named **VacSol** and place/put pftools and

other material in that folder. This step is not mandatory but suggested for the organization of installation material.

1.1.5.1 Libpsortb

Download libpsortb 1.0 from <https://goo.gl/MGJbcb> or from a primary source <http://www.psort.org/download/libpsortb-1.0.tar.gz>.

Please also extract and place the above.ar.gz in VacSol folder.

Open the folder VacSol/libpsortb-1.0 in terminal and run following commands in terminal one by one.

```
sudo ./configure
```

```
sudo make
```

```
sudo make install
```

```
sudo ldconfig
```

At any step, if system prompts for any input, please press Enter and use default options. After executing above mentioned commands, user will see that a folder has created **/usr/local/lib64** with libraries libhmmmer, libmodhmm, libsquid and libsvmloc etc

Note: Final libraries must be in a path that the dynamic linker from Perl can find.

Add following paths in .bashrc file

```
export $HOME:/usr/bin
export PATH=$PATH:$HOME:/usr/bin
export PSORT_PFTOOLS=<basePath>/VacSol/pftools
export BLASTDIR=/usr/bin
export LD_LIBRARY_PATH=/usr/local/lib64
export PSORT_ROOT=/usr/local/psort/bin
source ./bashrc
```

User should update the path, if he/she places pftool or blast at any other location. Set the above environment variables at the end of .bashrc file.

1.1.5.2 bio-tools-psort-all

Download bio-tools-psort-all 3.0.3 from <https://goo.gl/5iVtoB> or from a primary source <http://www.psort.org/download/bio-tools-psort-all.3.0.3.tar.gz>

Please also extract and place this in folder VacSol.

Open file VacSol/bio-tools-psort-all/algorithm-hmm/*hmm-binding.cpp* in any text editor and add a line `#include <string.h>` after the line `#include "hmm.h"` at the top of the file (If this line is missed, user should add this line from avoiding to face the exception during installation).

Open the folder VacSol/bio-tools-psort-all in terminal and run following command in terminal

```
perl Makefile.PL
```

During installation, system will require path of blastall, pftool, libhmm, libmodhmm, libsquid and libsvmloc libraries.

Please set the correct path of each library carefully.

Note: If user uses default installations as mentioned above, then blastall will be available at location /usr/bin/. Pftool will be available in folder VacSol/pftools/ and other libraries will be available in folder /usr/local/lib64/

Download hmmtop 2.1 from <https://goo.gl/VeJkdP> or from a primary source <http://www.enzim.hu/hmmtop/html/download.html>

Please also extract and place this in folder VacSol.

Navigate to folder VacSol/hmmtop_2.1 in terminal.

Run following commands in terminal or open README file and follow the instructions.

```
cc hmmtop.c -lm -o hmmtop
```


After executing above mentioned steps, please add further environment variables in `.bashrc` file before source `./bashrc` line.

```
export PSORT_ROOT=/usr/local/psort/bin
export PSORT_HMMTOP=<basePath>/VacSol/hmmtop_2.1
export HMMTOP_ARCH=<basePath>/VacSol/hmmtop_2.1/hmmtop.arch
export HMMTOP_PSV=<basePath>/VacSol/hmmtop_2.1/hmmtop.psv
```

Navigate back to folder `VacSol/bio-tools-psort-all` in terminal and execute following commands

```
sudo make
sudo make test (optional, but recommended)
sudo make install
```

Run command `cp -r psort /usr/local/`, this command will copy a `psort` folder from `/home/<username>/Pipeline /bio-tools-psort-all/` to `/usr/local/psort`

In terminal open a directory using following command

```
cd /usr/local/psort/conf/analysis/scblast
```

and then run a command

```
sudo ./makedb.sh
```

After installation please verify that in file `/usr/local/psort/bin/psort`, the mentioned path at line [`my$root = '/usr/local/psort'`] is correctly set. If you not correctly set then update this path.

Now a standalone `PSORTb` program is available in `$PSORT_ROOT/bin` and can be used to test.

1.1.6 OSDDlinux Installation

Install OSDlinux for the installation of ABCPred, Propred and Propred-I.

OSDlinux installation on an existing machine is available at

<http://osddlinux.osdd.net/installe.php>

There is no need to follow all steps mentioned in installation guide available at

<http://osddlinux.osdd.net/installe.php>

Please follow only following sections of that guide

- Pre installation setup for all type of machines
- Installation on any Unix based machine
- Post installation for all machines

Required steps are mentioned below for the ease of researcher, user can also easily install it after applying the following steps one by one.

1. Make a directory "/gpsr/" (directory for installing osddlinux, mkdir /gpsr)
2. Now you need to download only following files
 - base.tar.gz (basic or minimum infrastructure)
 - data.tar.gz (BLAST data for creating PSSM profile & similarity search)
 - models.tar.gz (SVM models used for prediction & classification)
3. Uncompressed and detar above files in directory "/gpsr" (e.g., tar -zxvf base.tar.gz)
4. Set the environment paths by using command

```
cat /gpsr/gpsr_env.sh >> ~/.bashrc"
```

This step will update environment variable HOME as mentioned below.

```
$HOME:/gpsr
```

Review the environment variables in .bashrc file

5. Copy perl of system in /gpsr/local/bin/perl ;directory by using command

```
cp /usr/bin/perl /gpsr/local/bin/
```

6. Change ownership of files in /gpsr ("chown -R <username> /gpsr/*")
7. Change group of all files in /gpsr ("chgrp -R <username> /gpsr/*")
8. Start crontab jobs to excute background jobs ("crontab /gpsr/cronjobs")
9. Restart your computer

1.1.6.1 Modifications

Modify some code files so that VacSol require files in its required format.

Download modified files from <https://goo.gl/TNcOUQ> and replace existing files with these files.

Replace /gpsr/standalone/abcpred/abcpred.pl with downloaded abcpred.pl file

Similarly replace /gpsr/standalone/propred1/propred1.pl with downloaded propred1.pl file

and replace /gpsr/standalone/propred/propred.pl with downloaded propred.pl file

Now abcpred, propred and propred1 are installed.

Please test abcpred, propred and propred1 individaully to ensure that these tools are installed and configured correctly.

1.1.7 Databases

Download configured database files of DEG, SWISS-Prot, MVIRDB, VFDB etc from <https://goo.gl/6aKKUb>

1.1.8 VacSol

Download executable VacSol from <https://goo.gl/9Uzq5o> and install jre and jdk version 7 or higher.

If there is any jre or jdk version 6, please uninstall it and install version 7 or higher.

1.1.9 Configurations

Download Config files from <https://goo.gl/grr11P> and place folder Configs in same directory in which VacSol.jar is placed.

1.1.9.1 Log4j Configurations

Create a Temp directory and place its path in Configs /log4j.properties file as mentioned below

```
log = <temp directory path>/log
```

1.1.9.2 VacSol Configurations

Open config.properties and configure paths of DBs and tools in this configuration file as mentioned below.

```
# HSDB Database Path
HSDB=<baseURL>/VacSol/DBs/HSDB/HSDB
# PSortB Bin Directory Path
PSORTB=/usr/local/psort/bin
# DEG Database path
DEGDB=<baseURL>/VacSol/DBs/DEG
# Virulence Factors Database path
VIRFDB=<baseURL>/VacSol/DBs/VIRF
# HMMTOP directory path
HMMTOP=<baseURL>/VacSol/hmmtop_2.1
# Uniprot Database Path
```

```

ANNOT=<baseURL>/VacSol/DBs/ANNOT/uniprot_sprot
# ABCPred for BCell Epitops Path
ABCPred=/gpsr/standalone/abcpred
# Propred1 for TCell MHC1 Epitops Path
MHC1=/gpsr/standalone/propred1
# Propred1 for TCell MHC2 Epitops Path
MHC2=/gpsr/standalone/propred
# XML Template path for using STANDALON mode
TEMPLATE=<baseURL>/VacSol/vacSol.xml
# DEFAULT Submission Method files directory
DEFAULT=<baseURL>/VacSol/Genomes
# Path of TEMP folder
TEMPFOLDER=<baseURL>/VacSol/Temp
# Path for gpsr internal tools TEMP folder
GPSRTEMP=/gpsr/temp
# Directory in which you want to store results
OUTPUT=<crate directory and give its path here>
# Tell system to clear temp directories at the end of
processing
CLEAR=false
# Tell system that Internet is available to use, on false
value system will not use internet consuming features
INTERNET=true
# Options applicable if internet feature is true
SocketTimeout=30000
# Socket communication time out
ConnectTimeout=30000
# Maximum proteins to process for epitope mapping in one
go. Note: 1 Sequences will take around maximum 3 minutes if
internet consuming feature is true, it's better to keep this
value small not greater than 10 and use multiple passes to
save time

```

MaxSeqEpitopProcessing = 10

Caution: Please do not change path variables, VacSol will recognize only the mentioned path variables.

Table 1.1: Description of Different Pathway Variables.

PATH VARIABLE	PURPOSE
HSDB	This variable will require path of homo sapiens proteins database which will be used by Blaster for homology checking
DEGDB	This variable will require path of essential genes database which will be used by Blaster for essentiality checking
VIRFDB	This variable will require path of virulent proteins database which will be used by Blaster for virulence checking
PSORTB	This variable will require path of PSORTb which will be used by Localizer to predict localization.
HMMTOP	This variable will require path of hmmtop which will be used by Helicer to calculate transmembrane helices
ANNOT	This variable will require path of database which will be used for annotation. Currently UniProt/SwissProt database is configured.
ABCPRED	This variable will require path of ABCPred tool which will be used by Epitoper to calculate B Cell epitopes
MHC1	This variable will require path of Propred1 tool which will be used by Epitoper to calculate MHC-I T Cell epitopes
MHC2	This variable will require path of Propred tool which will be used by Epitoper to calculate MHC-II T Cell epitopes
TEMPLATE	This path variable will require the path of XML Template that is used in case of STANDALONE mode. (For further detail of this XML Template please review Appendix B User Guide)

DEFAULT	This variable require the path of the folder where genome or proteome files are already placed which can be used in DEFAULT input mode. (For further detail please review Appendix B User Guide)
TEMPFOLDER	VacSol create some intermediate files during its processing. This variable require the path of such temporary folder. This folder must be created and its path should be configured at this path variable
GPSRTEMP	Tools used by Epitoper also require a temporary folder. Path of such temporary folder must be configured at this path variable. Such temporary folder will be with name temp under gpsr directory (/gpsr/temp).
OUTPUT	VacSol generate result in five different formats and placed in an output directory. This directory must be created and its path should be configured at this path variable
CLEAR	This variable is actually a flag which require true or false value. This indicate VacSol to clear or not temporary directories at the end of processing. Recommended value is true.
INTERNET	This variable is also a flag which require true or false value. This indicate VacSol to use internet (if available) for some internet dependent.
SOCKETTIMEOUT	This variable require socket time-out. Recommended value is 30000
CONNECTTIMEOUT	This variable require connection time-out. Recommended value is 30000
MAXSEQEPITOPPRO CESSING	Prioritized proteins will be used for epitope mapping. This variable will limit the number of proteins processing for Epitoper.

1.1.9.3 Additional Configurations

These additional configurations are require to enable DEFAULT input mechanism of VacSol.

Place the default proteome or genome files in a folder configured under path variable DEFAULT.

For example a folder <basePath>/Genome contains files like

- Acinetobacter.fasta
- Campylobacter.fasta
- Bacillus.fasta

And this folder is configured under path variable

DEFAULT=<basePath>/Genome

Open a configuration file proteomes.config and configure the name of these files like

1=Acinetobacter

2=Campylobacter

3=Bacillus

These names will be displayed in dropdown of default input mode and VacSol will automatically accepts the path and name of the file. VacSol will pick that file and process it.

2 Appendix B: User Guide

Open Terminal and type command

```
java -jar VacSol.jar [Tool Type] [Debug Level<Optional>] [Number of Cores <Optional>]
```

Tool Type is first parameter that is composed of two modes (i) GUI and (ii) STANDALONE. GUI option refers to graphical user interface view, whereas STANDALONE option requires input values in XML template and xml file path should be mentioned in configuration file (config.properties) under TEMPLATE variable. Please consult Installation Guide for detail of these path variables.

Debug Level is second parameter and it exhibits values ALL, DEBUG, ERROR, FATAL, INFO, OFF, TRACE or WARN. It is an optional parameter. If user does not select this parameter, then VacSol will use a default value for debugging level as ALL. This parameter informs VacSol to set logging level. Details of logging levels are mentioned in the following table.

Table 1: Description of VacSol Logging Levels.

Level	Description
ALL	All levels including custom levels.
DEBUG	Designates fine-grained informational events that are most useful to debug an application.
ERROR	Designates error events that might still allow the application to continue running.
FATAL	Designates very severe error events that will presumably lead the application to abort.
INFO	Designates informational messages that highlight the progress of the application at coarse-grained level.
OFF	The highest possible rank and is intended to turn off logging.
TRACE	Designates fine-grained informational events than the DEBUG.
WARN	Designates potentially harmful situations.

Note: DEBUG and TRACE logging levels are not supported in current VacSol version.

Number of cores is third parameter that is required to inform how VacSol can utilize as many cores available in host machine. It is also an optional parameter, and by default VacSol will use all available cores of the host machine.

Example: `java -jar VacSol.jar GUI ALL 3`

2.1 GUI Mode

On opening VacSol in GUI mode it will show a following window.

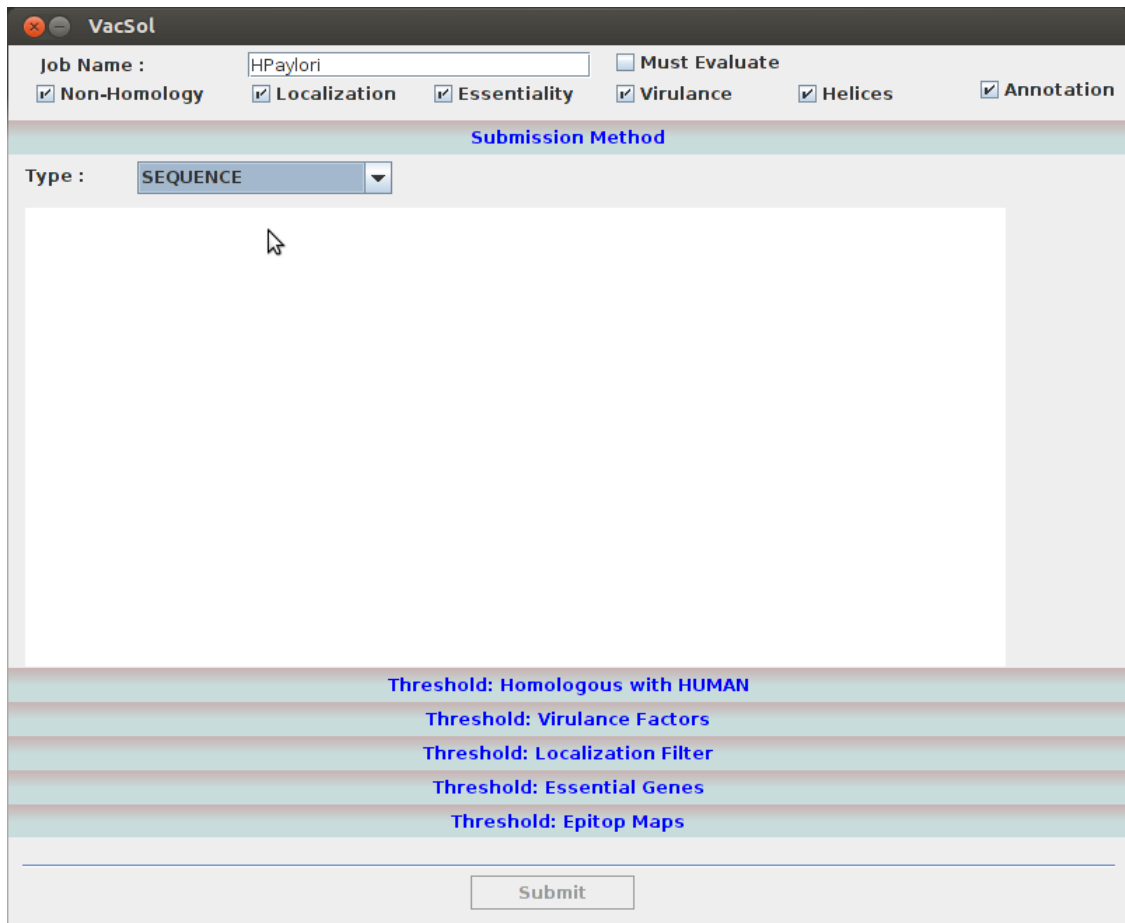


Figure 1: VacSol GUI Mode. It depicts the user friendly graphical user interface.

In top section of window VacSol requires some input in the form of checkboxes.

Non-Homology: This check box indicates that non-host homology checking is required or not.

Localization: This check box indicates that localization prediction is required or not.

Essentiality: This check box indicates that essentiality checking is required or not.

Virulence: This check box indicates that virulence protein identification is required or not.

Helices: This check box indicates transmembrane helices calculation is required or not.

Annotation: This check box indicates that the annotation of protein sequence with UniProt data base is required or not.

These check boxes control the behavior of VacSol, and make it feasible to work as a single tool or as a package of tools. To identify prioritize protein, all these checks must be checked.

Must Evaluate: This check is very important when user wants to use VacSol to identify prioritize protein(s). VacSol works in a flow as mentioned in Fig 3.1 and Fig. 3.5. If you see in Fig 3.5 there are some decision boxes and Must Evaluate check override these decision boxes. If Must Evaluate check is checked then VacSol will must evaluate all steps weather the previous step qualify the criteria or not.

Job Name: A text field which requires name of the job and it will save results in a directory with name provided as Job Name(s).

2.1.1 Submission Method

VacSol pipeline offers four different submission options. Accordion panel bar has title Submission method. This panel contains a drop down list of types through which user can select the required type.

2.1.1.1 SEQUENCE

In Sequence option, a small window will open in which user can paste the protein sequence in plain format and submit it.

VacSol

Job Name : Must Evaluate

Non-Homology Localization Essentiality Virulence Helices Annotation

Submission Method

Type :

```
MLHKKVLLALTASLICQESLFAKEKDYTLGKVSTAGKKDRSDYSQVNLGYSGITAPKSWQDEEVKKTGSRTVISNKALTQQANQSIEEAL
QNVPLGLQIRNATGVGAMPTIQIRGFGAGGSGHSDATLMLVNGIPVYMAPYAHIELDIFPVTFQAIDRIDVKGGGVSQYGPNTYGGIVNIITK
PIPNQWENQAERITYWAKARNAGFAAPPDKTGDPSFIKSLGNLNTYVRSRGGMINKHVGIQAQANWVRGQGFRDNSPNSISNYWL
DGVYDINENNGIKAYYQYYDFAIAQP GSLSEQDYKINRFANLRPLNQKGGRSQRF GAVYENRFGDLKVGGTFSFTYYGQLMTRDFQVSS
SYNSANMVTCSFAACRAAGLPAGYNLAVPYATNYNGWAEVENPVRISINNAFEPKVNLIWNTGKVKQTFIMGLRFMTTFLQRQYLNTN
ECATKTSGEGAGFLCEGANVMSGWKPHIKHGVYRNWNWNRNNYAVYLSDRIEAWDGRFFIVPGLRYAFVQYNNENASNWMQIPEKD
LRKIKHMNNWMPSTNIGFIPVQGDHNVLTYNQYRSFVPPQLDVLVSYGGAEYFTQHFDTVEAGARYTYKDKFSFNADYFRWARDFATG
QYSVYTSGPMKGNVRPINGYSQGVELELYRPIRGLQFHAAFNYIDTRVTSHGPLTDLNGDVLKGTSYNKHFFVSPFQFILDARYNWRKT
TIGISSYFYSRAYSGISNSAAGGYGMQYSSGGNNYESVLNSGYCEAWCMTQHEGLLPWYWWNIQVSQIFWENGRHRVTSGLQINNI
FNMKYFTGIGSSPAGLQAPGRSVTAYLNYTF
```

Threshold: Homologous with HUMAN

Threshold: Virulence Factors

Threshold: Localization Filter

Threshold: Essential Genes

Threshold: Epitop Maps

Figure 2: VacSol Input Acceptance Way in Sequence Form. If a user selects a sequence way, then a text box appears to paste the sequence in FASTA format.

2.1.1.2 FILE

This submission method provides option to user of selecting the desired file and submit it to VacSol pipeline. On selection of file, VacSol displays the count of protein sequences in front of Browse button.

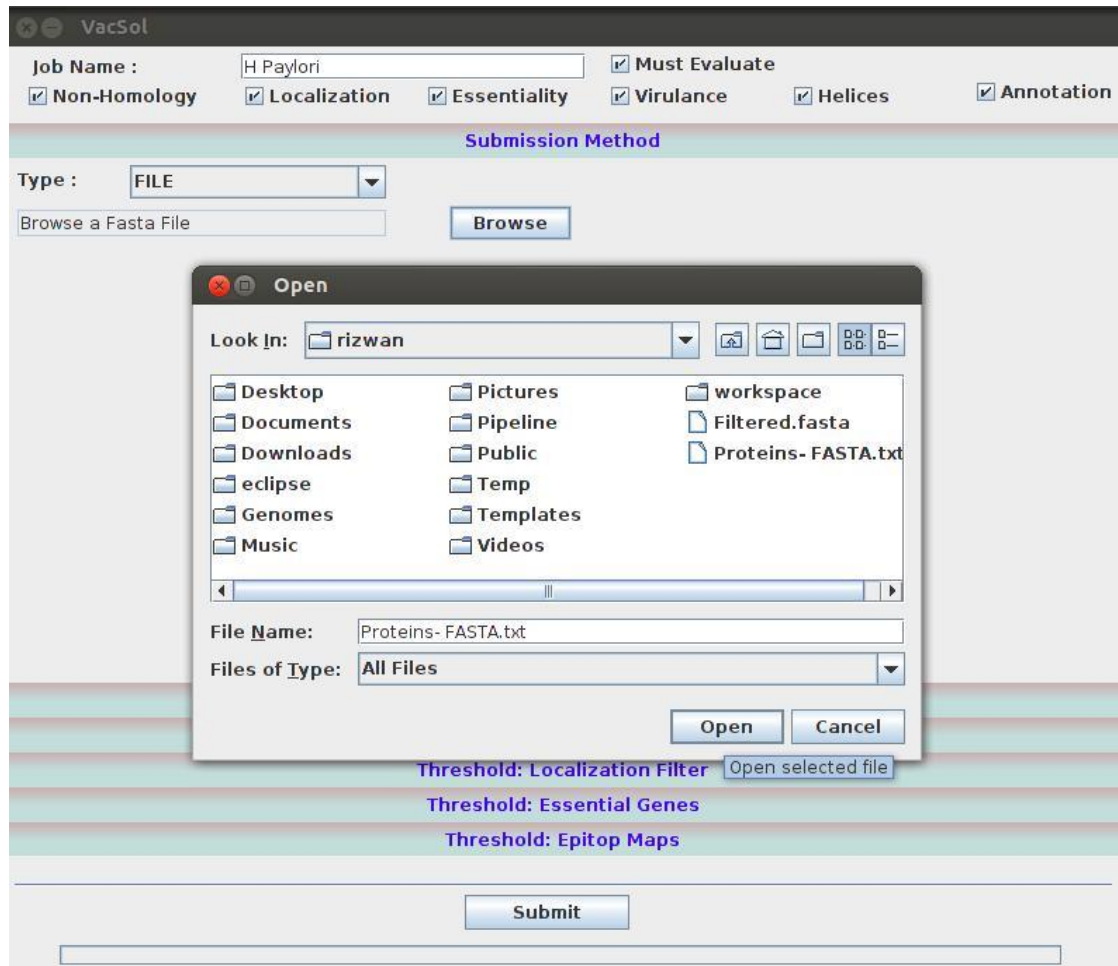


Figure 3: VacSol Input Acceptance Way in File Form. A pop up window will open to select proteome sequence file, if user selects a File form.

2.1.1.3 DEFAULT

This option display another dropdown to select one of the pre-configured default fasta file.

The screenshot shows the VacSol application window. At the top, the 'Job Name' is 'H Paylori'. There are several checked checkboxes: 'Non-Homology', 'Localization', 'Essentiality', 'Virulance', 'Helices', 'Annotation', and 'Must Evaluate'. The 'Submission Method' section is highlighted in blue. Under 'Type', a dropdown menu is open, showing a list of bacterial species: Helicobacter pylori (selected), Caulobacter crescentus, Bacillus subtilis, Acinetobacter, Staphylococcus aureus, pseudomonas aeruginosa, and Mycoplasma pulmonis. To the right of this list, it says 'You choose 115 sequences'. Below the list, there are five threshold options: 'Threshold: Homologous with HUMAN', 'Threshold: Virulance Factors', 'Threshold: Localization Filter', 'Threshold: Essential Genes', and 'Threshold: Epitop Maps'. At the bottom, there is a 'Submit' button and a progress bar.

Figure 4: VacSol Input Acceptance Way from Already Available Sequences. On selecting Default option, a drop down will appear to show the bacterial proteomes that have already present in VacSol.

2.1.1.4 ONLINE

This option requires a UniProt protein ID. INTERNET variable must be configured as true and internet should be available at host machine, for using this online option. VacSol will download the protein of provided id and execute it based on provided parameters.

The screenshot shows the VacSol web interface. At the top, the window title is "VacSol". Below it, there is a "Job Name" field containing "HPaylori" and a "Must Evaluate" checkbox which is unchecked. A row of checkboxes includes "Non-Homology", "Localization", "Essentiality", "Virulence", "Helices", and "Annotation", all of which are checked. A section titled "Submission Method" contains a "Type" dropdown menu set to "ONLINE". Below this is an input field for a UniProt ID, with a mouse cursor pointing to it and the text "Enter Uniprot ID e.g; Q6GZX4". At the bottom of the form, there are five threshold settings: "Threshold: Homologous with HUMAN", "Threshold: Virulence Factors", "Threshold: Localization Filter", "Threshold: Essential Genes", and "Threshold: Epitop Maps". A "Submit" button is located at the very bottom of the interface.

Figure 5: VacSol Input Acceptance through Online Retrieving Sequence. Online option requires a UniProt protein ID.

2.1.2 Threshold: Homologous with Human

This panel requires the threshold values of evaluation criteria. VacSol evaluates homologous proteins with human based on provided criteria and discards them from results and identify non-homologous proteins.

The screenshot shows the VacSol web interface. At the top, the window title is 'VacSol'. Below it, the 'Job Name' field contains 'H Paylori'. To the right of the job name are several checkboxes, all of which are checked: 'Must Evaluate', 'Non-Homology', 'Localization', 'Essentiality', 'Virulence', 'Helices', and 'Annotation'. Below these is a section titled 'Submission Method' with a sub-section 'Threshold: Homologous with HUMAN'. This section contains three dropdown menus: 'Bit Scores' set to '100', 'E Value' set to '1.0E-5', and 'Percentage Identity' set to '35.0'. Below these are four more sections, each with a title: 'Threshold: Virulence Factors', 'Threshold: Localization Filter', 'Threshold: Essential Genes', and 'Threshold: Epitop Maps'. At the bottom of the form is a 'Submit' button and a long empty input field.

Figure 6: Threshold Representation for Homology Detection. This figure clearly displays the threshold values that have set for determining homology among given bacterial sequences and the human sequences.

2.1.3 Threshold: Virulence Factors

This panel requires the threshold values of virulence factor identification criteria. VacSol evaluates virulence proteins based on given threshold criteria.

The screenshot shows the VacSol web interface. At the top, the window title is "VacSol". Below it, there is a "Job Name" field containing "H Paylori". To the right of the job name are several checkboxes: "Must Evaluate", "Non-Homology", "Localization", "Essentiality", "Virulence", "Helices", and "Annotation", all of which are checked. Below these are several horizontal bars representing different threshold categories: "Submission Method", "Threshold: Homologous with HUMAN", "Threshold: Virulence Factors", "Threshold: Localization Filter", "Threshold: Essential Genes", and "Threshold: Epitop Maps". The "Threshold: Virulence Factors" section is currently active and contains four dropdown menus: "Bit Scores" (set to 100), "E Value" (set to 1.0E-5), "Percentage Identity" (set to 35.0), and "Data Base". The "Data Base" dropdown menu is open, showing options "ALL", "MvirDB", and "VFDB", with "MvirDB" selected. At the bottom of the interface is a "Submit" button.

Figure 7: Threshold Representation for Determining Virulence Factors. It displays the threshold values for determining virulence factor by selecting MvirDB, VFDB, or both databases.

2.1.4 Threshold: Localization Filter

This panel requires threshold values for localization identification. VacSol evaluates and identifies proteins which have localizations mentioned in threshold criteria.

The screenshot shows the VacSol web interface. At the top, the 'Job Name' is 'HPaylori'. Below it, there are several checkboxes: 'Non-Homology' (checked), 'Localization' (checked), 'Essentiality' (checked), 'Virulance' (checked), 'Helices' (checked), and 'Annotation' (checked). There is also a 'Must Evaluate' checkbox which is unchecked. Below these are several sections with blue headers: 'Submission Method', 'Threshold: Homologous with HUMAN', 'Threshold: Virulance Factors', 'Threshold: Localization Filter', 'Threshold: Essential Genes', and 'Threshold: Epitop Maps'. In the 'Threshold: Localization Filter' section, there are three dropdown menus: 'Organism Type' set to 'Bacteria', 'Gram Stain' set to 'Negative', and 'Localizations' which is currently open, showing a list of options: 'Cytoplasmic', 'CytoplasmicMembrane', 'Cellwall', 'Extracellular', 'Periplasmic', 'Unknown', and 'OuterMembrane'. At the bottom of the interface is a 'Submit' button.

Figure 8: Representation of Threshold Set for Cytoplasmic & Extracellular Localization Filtering. This diagram describes that VacSol is very helpful in determining the protein subcellular locations. It also offers users to select any one of the mentioned locations to determine the protein residential position in cell that would be helpful in determining its function.

2.1.5 Threshold: Essential Genes

This panel requires the threshold values of essential proteins identification criteria. VacSol evaluates essential proteins based on provided threshold criteria.

The screenshot shows the VacSol application window. At the top, the job name is 'H Paylori'. Below it, several checkboxes are checked: 'Non-Homology', 'Localization', 'Essentiality', 'Virulence', 'Helices', and 'Annotation'. The 'Submission Method' is set to 'Threshold: Homologous with HUMAN'. Under the 'Threshold: Essential Genes' section, the 'Bit Scores' are set to 100, 'E Value' to 1.0E-5, and 'Percentage Identity' to 35.0. The 'Organism' dropdown menu is open, displaying a list of bacterial species. At the bottom of the interface, there is a 'Submit' button.

Figure 9: Representation of Threshold Set for Identifying Essential Genes. It also describes threshold values that have been set for essential genes determination. Users can select any bacterial proteome to determine the presence of essential genes in their provided proteome.

2.1.6 Threshold: Epitope Maps

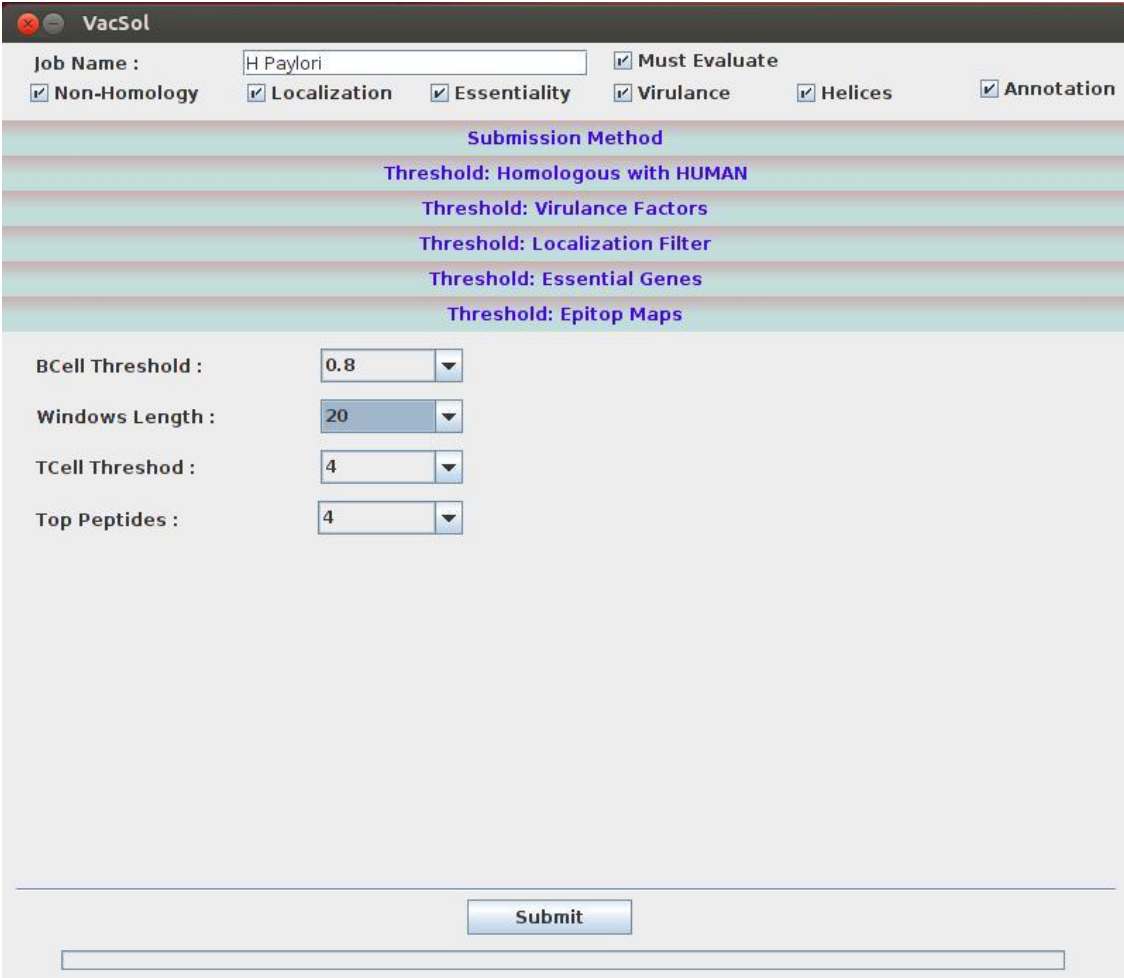
This panel requires the threshold values for epitope maps.

B-Cell Threshold: BCell highest scoring peptides, threshold is a pre-defined numerical value used to make decision.

Windows Length: Number of peptide mers to select. For example 20 mers

T-Cell Threshold: %age of highest scoring peptides.

Top Peptides: Number of top peptides to select.



The screenshot shows the VacSol application window. At the top, the 'Job Name' is 'H Paylori'. Below it, several checkboxes are checked: 'Non-Homology', 'Localization', 'Essentiality', 'Virulence', 'Helices', and 'Annotation'. A section titled 'Submission Method' lists several thresholds: 'Homologous with HUMAN', 'Virulence Factors', 'Localization Filter', 'Essential Genes', and 'Epitop Maps'. Below this, four dropdown menus are visible: 'BCell Threshold' (0.8), 'Windows Length' (20), 'TCell Threshold' (4), and 'Top Peptides' (4). A 'Submit' button is located at the bottom center of the window.

Figure 10: Representation of Threshold Set for Epitope Maps. It displays the threshold values for epitope mapping.

2.2 STANDALONE Mode

Standalone mode is an alternative of GUI mode. In this mode user can provide input values in the form of an XML template. All fields are same and equivalent to GUI mode, but the difference is that in GUI mode user selects values from graphical user interface. While in STANDALONE mode, user will fill values in XML template and set the path of XML template under path variable TEMPLATE .

```

<VacSol>
  <JobName>MyTestJob</JobName>
  <MustEvaluate>true</MustEvaluate>
  <isPerformNonHomology>true</isPerformNonHomology>
  <isPerformEssentiality>true</isPerformEssentiality>
  <isPerformVirulance>true</isPerformVirulance>
  <isPerformEpitopMapping>true</isPerformEpitopMapping>
  <isPerformTransmembraneHelices>true</isPerformTransmembraneHelices>
  <SubmissionMethod>
    <Type>SEQUENCE</Type>
    <FilePath>/home/rizwan/File.fasta</FilePath>
    <Sequence>ABCDEFGHGIJK</Sequence>
  </SubmissionMethod>
  <Thresholds>
    <Homology>
      <Threshold>
        <BitScore>100</BitScore>
        <eValue>1.0E-5</eValue>
        <PercentageIdentity>35.0</PercentageIdentity>
      </Threshold>
    </Homology>
    <VirulanceFactors>
      <DataBaseType>ALL</DataBaseType>
      <Threshold>
        <BitScore>100</BitScore>
        <eValue>1.0E-5</eValue>
        <PercentageIdentity>35.0</PercentageIdentity>
      </Threshold>
    </VirulanceFactors>
    <Localizations>
      <Localization>Extracellular</Localization>
      <Localization>CytoplasmicMembrane</Localization>
      <OrganismType>Bacteria</OrganismType>
      <GramStain>n</GramStain>
    </Localizations>
    <Essentiality>
      <Organism>DEG</Organism>
      <Organism>DEG1001</Organism>
      <Organism>DEG1013</Organism>
      <Threshold>
        <BitScore>100</BitScore>
        <eValue>1.0E-5</eValue>
        <PercentageIdentity>35.0</PercentageIdentity>
      </Threshold>
    </Essentiality>
    <EpitopMap>
      <BCELL>
        <Thresold>0.5</Thresold>
        <WindowLength>20</WindowLength>
      </BCELL>
      <AntigenThreshold>0.5</AntigenThreshold>
      <TCELL>
        <Thresold>5</Thresold>
        <TopPeptides>5</TopPeptides>
      </TCELL>
    </EpitopMap>
  </Thresholds>
</VacSol>

```

Figure 10: Input Template for Standalone Mode. This figure describes the XML input template for using VacSole as a standalone mode.

Bibliography

- [1]. Moriel, D.G., et al., *Genome-based vaccine development: a short cut for the future*, in *Pharmaceutical Biotechnology*. 2009, Springer. p. 81-89.
- [2]. Loscalzo, J., I. Kohane, and A.L. Barabasi, *Human disease classification in the postgenomic era: a complex systems approach to human pathobiology*. *Molecular systems biology*, 2007. **3**(1).
- [3]. Casadevall, A. and L.-a. Pirofski, *Ditch the term pathogen*. *Nature*, 2014. **516**(7530): p. 165-166.
- [4]. Wilson, B.A., et al., *Bacterial pathogenesis: a molecular approach*. 2011: American Society for Microbiology (ASM).
- [5]. Naz, A., et al., *Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: A reverse vaccinology based approach*. *Infection, Genetics and Evolution*, 2015. **32**: p. 280-291.
- [6]. Reid, G., J. Howard, and B.S. Gan, *Can bacterial interference prevent infection?* *Trends in microbiology*, 2001. **9**(9): p. 424-428.
- [7]. Roush, S.W., T.V. Murphy, and V.-P.D.T.W. Group, *Historical comparisons of morbidity and mortality for vaccine-preventable diseases in the United States*. *Jama*, 2007. **298**(18): p. 2155-2163.
- [8]. Henderson, D., *Smallpox: The death of a disease—The inside story of eradicating a worldwide killer (pp. 193–195)*. Amherst, NY: Prometheus Books, 2009.
- [9]. Fenner, F., et al., *Smallpox and its eradication continued*. 1988: World Health Organization.
- [10]. Dales, S. and B.G. Pogo, *Biology of poxviruses*. Vol. 18. 2013: Springer Science & Business Media.
- [11]. Kanampalliwar, A.M., et al., *Reverse Vaccinology: basics and applications*. *Journal of Vaccines & Vaccination*, 2013. **2013**.
- [12]. Delany, I., R. Rappuoli, and K.L. Seib, *Vaccines, reverse vaccinology, and bacterial pathogenesis*. *Cold Spring Harbor perspectives in medicine*, 2013. **3**(5): p. a012476.

- [13]. Capecchi, B., et al., *The genome revolution in vaccine research*. Current issues in molecular biology, 2004. **6**: p. 17-28.
- [14]. Rappuoli, R., *Reverse vaccinology, a genome-based approach to vaccine development*. Vaccine, 2001. **19**(17): p. 2688-2691.
- [15]. Rinaudo, C.D., et al., *Vaccinology in the genome era*. The Journal of clinical investigation, 2009. **119**(9): p. 2515-2525.
- [16]. Pizza, M., et al., *Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing*. Science, 2000. **287**(5459): p. 1816-1820.
- [17]. Moriel, D.G., et al., *Genome -based vaccine development: A short cut for the future*. Human vaccines, 2008. **4**(3): p. 184-188.
- [18]. Rappuoli, R., *Bridging the knowledge gaps in vaccine design*. Nature biotechnology, 2007. **25**(12): p. 1361-1366.
- [19]. Lefébure, T. and M.J. Stanhope, *Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition*. Genome Biol, 2007. **8**(5): p. R71.
- [20]. Zhao, Y., et al., *PGAP: pan-genomes analysis pipeline*. Bioinformatics, 2012. **28**(3): p. 416-418.
- [21]. He, Y., Z. Xiang, and H.L. Mobley, *Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development*. BioMed Research International, 2010. **2010**.
- [22]. Ali, A., et al., *Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets*. BioMed research international, 2015. **2015**.
- [23]. Gardy, J.L., et al., *PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis*. Bioinformatics, 2005. **21**(5): p. 617-623.
- [24]. Yu, C.-S., et al., *CELLO2GO: a web server for protein subCELLular LOcalization prediction with functional gene ontology annotation*. PloS one, 2014. **9**(6): p. e99368.

- [25]. Goodswen, S.J., P.J. Kennedy, and J.T. Ellis, *Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology*. *Bioinformatics*, 2014: p. btu300.
- [26]. Barh, D., et al., *In silico subtractive genomics for target identification in human bacterial pathogens*. *Drug Development Research*, 2011. **72**(2): p. 162-177.
- [27]. Barh, D., et al., *A novel strategy of epitope design in Neisseria gonorrhoeae*. *Bioinformatics*, 2010. **5**(2): p. 77-85.
- [28]. Song, C.M., S.J. Lim, and J.C. Tong, *Recent advances in computer-aided drug design*. *Briefings in bioinformatics*, 2009. **10**(5): p. 579-591.
- [29]. Wadood, A., et al., *In-silico drug design: An approach which revolutionarised the drug discovery process*. *OA Drug Design & Delivery*, 2013. **1**(1): p. 4.
- [30]. Njogu, P.M., et al., *Computer-Aided Drug Discovery Approaches against the Tropical Infectious Diseases Malaria, Tuberculosis, Trypanosomiasis, and Leishmaniasis*. *ACS Infectious Diseases*, 2015.
- [31]. Vivona, S., F. Bernante, and F. Filippini, *NERVE: new enhanced reverse vaccinology environment*. *BMC biotechnology*, 2006. **6**(1): p. 35.
- [32]. Larsen, J.E.P., O. Lund, and M. Nielsen, *Improved method for predicting linear B-cell epitopes*. *Immunome research*, 2006. **2**(1): p. 1.
- [33]. Feldhahn, M., et al., *EpiToolKit—a web server for computational immunomics*. *Nucleic acids research*, 2008. **36**(suppl 2): p. W519-W522.
- [34]. Dimitrov, I., et al., *EpiTOP—a proteochemometric tool for MHC class II binding prediction*. *Bioinformatics*, 2010. **26**(16): p. 2066-2068.
- [35]. Ponomarenko, J.V. and P.E. Bourne, *Antibody-protein interactions: benchmark datasets and prediction tools evaluation*. *BMC Structural Biology*, 2007. **7**(1): p. 1.
- [36]. Sachdeva, G., et al., *SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks*. *Bioinformatics*, 2005. **21**(4): p. 483-491.
- [37]. Tusnady, G.E. and I. Simon, *The HMMTOP transmembrane topology prediction server*. *Bioinformatics*, 2001. **17**(9): p. 849-850.

- [38]. Bigelow, H. and B. Rost, *PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins*. Nucleic acids research, 2006. **34**(suppl 2): p. W186-W188.
- [39]. Consortium, U., *The universal protein resource (UniProt)*. Nucleic acids research, 2008. **36**(suppl 1): p. D190-D195.
- [40]. Jaiswal, V., et al., *Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions*. BMC bioinformatics, 2013. **14**(1): p. 1.
- [41]. Doytchinova, I.A. and D.R. Flower, *VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines*. BMC bioinformatics, 2007. **8**(1): p. 4.
- [42]. Li, L., C.J. Stoeckert, and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes*. Genome research, 2003. **13**(9): p. 2178-2189.
- [43]. Vita, R., et al., *The immune epitope database 2.0*. Nucleic acids research, 2010. **38**(suppl 1): p. D854-D862.
- [44]. Horton, P., et al., *WoLF PSORT: protein localization predictor*. Nucleic acids research, 2007. **35**(suppl 2): p. W585-W587.
- [45]. Emanuelsson, O., et al., *Locating proteins in the cell using TargetP, SignalP and related tools*. Nature protocols, 2007. **2**(4): p. 953-971.
- [46]. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*. Journal of molecular biology, 2001. **305**(3): p. 567-580.
- [47]. Nancy, Y.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes*. Bioinformatics, 2010. **26**(13): p. 1608-1615.
- [48]. Chan, J.N., C. Nislow, and A. Emili, *Recent advances and method development for drug target identification*. Trends in pharmacological sciences, 2010. **31**(2): p. 82-88.
- [49]. Kaufmann, S.H. and P.-H. Lambert, *The Grand Challenge for the Future: Vaccines for Poverty-Related Diseases from Bench to Field*. 2005: Springer Science & Business Media.

- [50]. Grandi, G., *Bacterial surface proteins and vaccines*. ITALIAN JOURNAL OF BIOCHEMISTRY, 2007. **56**(3): p. R.
- [51]. Cooper, I. and M. Duffield, *The in silico prediction of bacterial essential genes*. Science against microbial pathogens: communicating current research and technological advances, 2011.
- [52]. Luo, H., et al., *DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements*. Nucleic acids research, 2013: p. gkt1131.
- [53]. Liu, W., B. Schmidt, and W. Muller-Wittig, *CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware*. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2011. **8**(6): p. 1678-1684.
- [54]. Casadevall, A. and L.-a. Pirofski, *Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity*. Infection and immunity, 1999. **67**(8): p. 3703-3713.
- [55]. Zhou, C., et al., *MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications*. Nucleic acids research, 2007. **35**(suppl 1): p. D391-D394.
- [56]. Chen, L., et al., *VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors*. Nucleic acids research, 2011: p. gkr989.
- [57]. Käll, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method*. Journal of molecular biology, 2004. **338**(5): p. 1027-1036.
- [58]. Camacho, C., et al., *BLAST+: architecture and applications*. BMC bioinformatics, 2009. **10**(1): p. 1.
- [59]. ; Available from: <http://web.expasy.org/pftools/>.
- [60]. Waldmann, J., et al., *FastaValidator: an open-source Java library to parse and validate FASTA formatted sequences*. BMC research notes, 2014. **7**(1): p. 365.
- [61]. Saha, S. and G.P. Raghava, *Prediction methods for B-cell epitopes*. Immunoinformatics: Predicting Immunogenicity In Silico, 2007: p. 387-394.
- [62]. Singh, H. and G. Raghava, *ProPred1: prediction of promiscuous MHC Class-I binding sites*. Bioinformatics, 2003. **19**(8): p. 1009-1014.

- [63]. Singh, H. and G. Raghava, *ProPred: prediction of HLA-DR binding sites*. Bioinformatics, 2001. **17**(12): p. 1236-1237.
- [64]. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic acids research, 2003. **31**(1): p. 365-370.
- [65]. Pertselmidis, A., J.W. Fondon, and W. John, *Having a BLAST with bioinformatics (and avoiding BLASTphemy)*. Genome Biol, 2001. **2**(10): p. 1-10.
- [66]. Surla, D., D. Ivanovic, and Z. Konjovic. *Development of the software system CRIS UNS*. in *Intelligent Systems and Informatics (SISY), 2013 IEEE 11th International Symposium on*. 2013: IEEE.
- [67]. Baltrus, D.A., et al., *The complete genome sequence of Helicobacter pylori strain G27*. Journal of bacteriology, 2009. **191**(1): p. 447-448.
- [68]. Hsieh, P.-F., et al., *Serum-induced iron-acquisition systems and TonB contribute to virulence in Klebsiella pneumoniae causing primary pyogenic liver abscess*. Journal of Infectious Diseases, 2008. **197**(12): p. 1717-1727.
- [69]. CONSTANTINESCU, C. and E. CONSTANTINESCU, *COULD KNOWLEDGE OF H. PYLORI PATHOGENICITY FACTORS LEAD TO THE EMERGENCE OF NEW METHODS FOR IDENTIFYING BACTERIA?* Bulletin of the Transilvania University of Brasov, Series VI: Medical Sciences, 2014. **7**(1).
- [70]. Schirm, M., et al., *Structural, genetic and functional characterization of the flagellin glycosylation process in Helicobacter pylori*. Molecular microbiology, 2003. **48**(6): p. 1579-1592.
- [71]. Hayashi, F., et al., *The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5*. Nature, 2001. **410**(6832): p. 1099-1103.
- [72]. Koebnik, R., *TonB-dependent trans-envelope signalling: the exception or the rule?* Trends in microbiology, 2005. **13**(8): p. 343-347.
- [73]. Hilbi, H., et al., *Environmental predators as models for bacterial pathogenesis*. Environmental microbiology, 2007. **9**(3): p. 563-575.
- [74]. Brahmachary, P.P., *Novel aspects of flagellar biogenesis and virulence in Helicobacter pylori*. 2004: uga.

- [75]. Dutta, A., et al., *In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori*. In silico biology, 2006. **6**(1, 2): p. 43-47.
- [76]. Mittl, P.R., et al., *Detection of high titers of antibody against Helicobacter cysteine-rich proteins A, B, C, and E in Helicobacter pylori-infected individuals*. Clinical and diagnostic laboratory immunology, 2003. **10**(4): p. 542-545.
- [77]. Louvel, H., I. Saint Girons, and M. Picardeau, *Isolation and characterization of FecA-and FeoB-mediated iron acquisition systems of the spirochete Leptospira biflexa by random insertional mutagenesis*. Journal of bacteriology, 2005. **187**(9): p. 3249-3254.
- [78]. Naville, M. and D. Gautheret, *Transcription attenuation in bacteria: theme and variations*. Briefings in functional genomics & proteomics, 2009. **8**(6): p. 482-492.
- [79]. Levy, S.B. and B. Marshall, *Antibacterial resistance worldwide: causes, challenges and responses*. Nature medicine, 2004. **10**: p. S122-S129.
- [80]. Tenover, F.C., *Mechanisms of antimicrobial resistance in bacteria*. The American journal of medicine, 2006. **119**(6): p. S3-S10.
- [81]. Barh, D., A. Kumar, and A.N. Misra, *Genomic Target Database (GTD): a database of potential targets in human pathogenic bacteria*. Bioinformatics, 2010. **4**(1): p. 50-51.
- [82]. Xiang, Z. and Y. He, *Vaxign: a web-based vaccine target design program for reverse vaccinology*. Procedia in Vaccinology, 2009. **1**(1): p. 23-29.
- [83]. Stover, C.K., et al., *Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen*. Nature, 2000. **406**(6799): p. 959-964.
- [84]. Giuliani, M.M., et al., *A universal vaccine for serogroup B meningococcus*. Proceedings of the National Academy of Sciences, 2006. **103**(29): p. 10834-10839.
- [85]. Mulić, R., *The Grand Challenge for the Future. Vaccines for Poverty-Related Diseases from Bench to Field*. 2007.
- [86]. Gasteiger, E., et al., *Protein identification and analysis tools on the ExPASy server*. 2005: Springer.

- [87]. Ramaswamy, V., et al., *Listeria-review of epidemiology and pathogenesis*. Journal of Microbiology Immunology and Infection, 2007. **40**(1): p. 4.
- [88]. Mittl, P.R. and W. Schneider-Brachert, *Sell-like repeat proteins in signal transduction*. Cellular signalling, 2007. **19**(1): p. 20-31.