MOLECULAR CHARACTERIZATION OF OUTER MEMBRANE PROTEINS OF *Shigella flexneri* AS VACCINE CANDIDATES

ADA YOUSEF KAZI

UNIVERSITI SAINS MALAYSIA

2018

MOLECULAR CHARACTERIZATION OF OUTER MEMBRANE PROTEINS OF *Shigella flexneri* AS VACCINE CANDIDATES

By

ADA YOUSEF KAZI

Dissertation is submitted in partial fulfillment of the requirement for the Degree of Master of Health Science (Biomedicine)

Mixed Mode

January 2018

ACKNOWLEDGEMENTS

"In the name of Allah, the most Gracious and the most Merciful"

Alhumdulillah, all praises to Allah, for giving me the strength, courage and inspiration in completing this study. I would like to express my deepest gratitude to all those who have contributed and supported me in each and every way.

I would like to thanks all my supervisors, Assoc. Prof Dr. Few Ling Ling, Assoc. Prof. Dr. Lim Boon Huat, Dr.Leow Chiuan Yee, Dr. Chuah Candy, Assoc. Prof. Dr. Kirnpal Kaur Banga Singh and Dr. Leow Chiuan Herng for their support, excellent guidance and supervision throughout the research project and also during the writing of this thesis.

A special thanks to all my friends and colleagues especially, the research group of Dr. Leow Chiuan Yee, INFORRM and the microbiology department of Pusat Pengajian Sains Perubatan (PPSP) who have directly or indirectly contributed to my work and provided assistance and guidance.

Also I would like to share my deepest appreciation to all at INFORRM for providing me the support, guidance and opportunity to conduct my experiment in their laboratory.

Lastly I would like to thank my parents who have encouraged and inspired me throughout my journey.

Thank You all for your great support, patience, love and encouragement

TABLE OF CONTENTS

CHAPTER 1- INTRODUCTION & LIT. REVIEW

INTRODUCTION

LITERATURE REVIEW

CHAPTER 2- MATERIALS & METHOD

MATERIALS

METHODS

CHAPTER 3 – RESULTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBEREVATIONS

ANN	Artificial Neural Network
$Cacl_2$	Calcium Chloride
cDNA	complementary DNA
DES	Database of essential genes
dH_2O	Deionized water
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
FAE	Follicular associated epithelium
FAO	Food and Agriculture Organization
HLA	Human leukocyte antigen
HMM	Hidden Markov Model
Hr	Hours
IL	Interleukin
IMAC	Immobilized metal-affinity chromatography
IPTG	Isopropyl thiogalactoside
LB	Luria-Bertani Broth
LPS	Lipopolysaccharide
M Cells	Membranous epithelial cells
MDR	Multiple Drug Resistance
MenB	*Meningococcus group B*
$Mgcl_2$	Magnesium Chloride
MHC	Major Histocompatibility Complex

Min	Minutes
NaCl	Sodium chloride
NGS	Next Generation Sequencing
OD	Optical Density
PAMPS	Pathogen Associated Molecular Patterns
PCR	Polymerase chain reaction
PGN	Peptidoglycan
PMNs	Polymorphonuclear neutrophils
PSSMs	Position Specific Scoring Matrices
rDNA	recombinant DNA technologies
SDS-PAGE	Sodium dodecyl sulphate – polyacrylamide gel electrophoresis
Sec	Seconds
*Sf*OM	*Shigella flexneri* Outermembrnae channel protein
*Sf*PL	*Shigella flexneri* Putative lipoprotein
Spp.	Species
SS Agar	*Salmonella-Shigella* agar
SVM	Support Vector Machine
T3SS	Type III secretion system
TAE	Tris Acetate Buffer
TEMED	Tetramethylethylenediamine
WHO	World Health Organization
X-gal	5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

ABSTRAK

Jangkitan kuman disentri adalah penyebab jangkitan silang diarea yang paling umum di negara-negara membangun. Setiap tahun, terdapat kira-kira 165 juta kes *shigellosis* yang dilaporkan di seluruh dunia. Penghidratan semula dan penggunaan antibiotik adalah rawatan yang biasa untuk *shigellosis*. Namun demikian, pengedaran dan penggunaan antibiotik yang tidak terkawal telah menyebabkan wujudnya strain rintang dadah. *Shigella* telah menunjukkan pelbagai kerintangan terhadap ubat-ubatan yang boleh digunakan untuk merawat jangkitannya. Dengan peningkatan kerintangan ubat yang ketara, cara terbaik untuk memerangi jangkitan ini ialah dengan mengenalpasti dan membangunkan protein immunogenic baru yang boleh digunakan untuk menghasilkan vaksin yang sesuai. Dalam kajian ini enam protein immunogenic luar membran telah dikenalpasti menggunakan pendekatan *immunoformatics* sebagai calon vaksin yang berpotensi pada masa akan datang. Daripada enam protein ini, dua protein panjang penuh dipilih dan di satukan menjadi rekombinan yang ditandai dengan protein His. Protein *chimeric* kemudiannya dihasilkan dengan menggabungkan sebahagian protein daripada kedua-dua protein panjang penuh tersebut. Ekspresi protein dilakukan dengan menggunakan sistem bakteria BL 21(DE3). Secara keseluruhannya penilaian kawasan pengekodan protein dengan kereaktifan imunologi boleh menghasilkan pengenalan antigen tambahan *Shigella* yang akan berguna sebagai vaksin dan reagen baru untuk diagnosis *shigellosis*.

ABSTRACT

In the developing world, bacillary dysentery is one of the most common communicable diarrheal infection. Approximately every year there are approximately 165 million cases of shigellosis that is reported worldwide. Treatment of shigellosis include oral rehydration and the use of antibiotics. However due to uncontrolled use and distribution of antibiotics, has led to the emergence of resistant strains. *Shigella* is once organism that has shown multiple drug resistance (MDR) to most of the drugs that have been used against it. With rapid increase in resistant strain the best way to combat this infection is by identifying and developing new immunogenic proteins that can be used to develop suitable vaccines. In this study, six immunogenic outer membrane proteins that could be used as potential vaccine candidates in the future were identified using immunoinformatics approach. From these six immunogenic proteins, two full length proteins were selected and expressed as recombinant His-tagged proteins. A chimeric protein that was then engineered by combining partial fragment of these two full length proteins. The expression of these proteins were carried out using BL21 (DE3) bacterial system. Overall, evaluation of these regions for encoding proteins with immunological reactivity can lead to the identification of additional antigens of *Shigella* which are useful as new vaccines and reagents for specific diagnosis of shigellosis.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1. Overview

Shigella is a gram-negative intracellular pathogenic enterobacteria genus, which is responsible for causing bacillary dysentery or shigellosis worldwide. It consists of 4 species namely *Shigella flexneri, S. sonnei, S. dysenteriae* and *S. boydii*. Each of these species is composed of different serotypes which can be identified based on the structure of their lipopolysaccharide O-antigen. *S. flexneri*, *S. boydii* and *S. dysenteriae* are the common causative agents of shigellosis in the developing world whereas *S. sonnei* is found to cause infections in developed countries. *Shigella* sp. invades and colonizes the colonic mucosa leading to its disruption. The bacilli invade the villi of the large intestine, multiply and spread laterally to adjacent epithelial cells and also penetrate into the lamina propria (*Warren, Parish and Schneider, 2006; Banga Singh et al., 2011)*.

Since the pathogen is intracellular it become even more difficult to treat. The symptoms associated with shigellosis range from mild self-limited diarrhea to severe dysentery with frequent passage of blood and mucus, high fever, abdominal cramps, malaise, chills, nausea, vomiting and in rare cases known to cause bacteremia. The diagnosis depends on the presence of erythrocytes, polymorphonuclear neutrophils (PMNs) and mucus in the patients stool sample *(Warren, Parish and Schneider, 2006; Banga Singh et al., 2011; Anderson, Sansonetti and Marteyn, 2016).*

This organism is mostly seen to infect children of age 5 years or younger and is one of the third most common bacterial agent responsible for childhood diarrhea. Apart from young children this pathogen also targets elderly and immunocompromised patients. Its mode of transmission is by fecal-oral route and requires a minimum infectious dose of <10 to 100 bacilli to cause the infection, due to its ability to survive gastric acidity better than other enterobacteria *(Banga Singh et al., 2011)*. However if immediate action is taken to cure the infection with the use of effective antimicrobial agents it may be possible to curtail the clinical symptoms and control the spread of the infection. The uncontrolled use of antimicrobial agents against *Shigella sp*. has inadvertently led to the development of resistant strains, which hinder appropriate selection of effective antibiotics. The variation in antimicrobial resistance from region to region further exacerbates selection of an appropriate antibiotic. According to a report by *Kotloff et al. (1999)*, the percentage distribution of *Shigella* species in develop and developing countries was found to be as follows:

Species	Developing countries	Developed countries
S. flexneri	60%	16%
S. sonnei	15%	77%
S. boydii	6%	2%
S. dysenteriae	6%	1%

Outbreak of shigellosis is mainly due to consumption of contaminated water, food, overcrowded communities, unhygienic food handlers and houseflies *(Kapperud et al., 1995; Shears, 1996).*

The treatment includes both rehydration and antibiotic therapy. As recommended by WHO a 5-day course of antibiotic is effective in treating diarrhea. However *Shigella* has evolved over the years and has become multi-drug resistant. It has shown resistance to sulphonamide, tetracycline, chloramphenicol, ampicillin and many more with increase in uncontrolled use and distribution of antimicrobial agent's future control of antimicrobial resistance will be very difficult to achieve *(Niyogi, 2007)*. Effective approach against shigellosis in developing countries relies on environmental sanitization, personal hygiene and to develop safe and effective vaccines *(Niyogi, 2007)*.

Development of vaccine through conventional techniques take decades to develop and with an increasing rate of emerging antimicrobial resistant organisms conventional methods of vaccine development is not a good idea, but now due to the availability of whole genome sequences, the genomic data could be used *in silico* to help identify and screen for potential vaccine candidates. This technique to develop vaccine using *in silico* method is termed as 'Reverse Vaccinology *(Mora et al., 2003; Rappuoli, 2009)*. To select a potential vaccine candidate it is essential to identify the virulent protein that is capable to evoke an immune response within the host organism. Some features for an effective vaccine candidate protein include: (i) sub-cellular localization; (ii) presence of a signal peptide; (iii) transmembrane domain; and (iv) antigenic epitopes. The main strategy behind identification of potential vaccine candidates is recognizing the antigenic and virulence factor as well as predicting those sequences which are likely to bind MHC class I and II proteins within the host. One such example that made use of reverse vaccinology was for the development of vaccine against Group B *meningococcus* (MenB) *(Rappuoli,*

2001, 2009; Mora et al., 2006; Doytchinova and Flower, 2007; Movahedi and Hampson, 2008; Naz et al., 2015).

The current study is based on the use of computational pipelines for the identification of putative vaccine targets against *S. flexneri*. Essential proteins were first identified, followed by determining which were antigenic. Subsequently the proteins were scanned for those epitopes having the ability to bind with both B-cells and T-cells. Identification of potential antigens can help in the development of protective immunogens against various other pathogens *(Naz et al., 2015)*. Using the identified vaccine candidate, a chimeric molecule was designed and its immunogenic response was tested using molecular techniques (Figure 1).

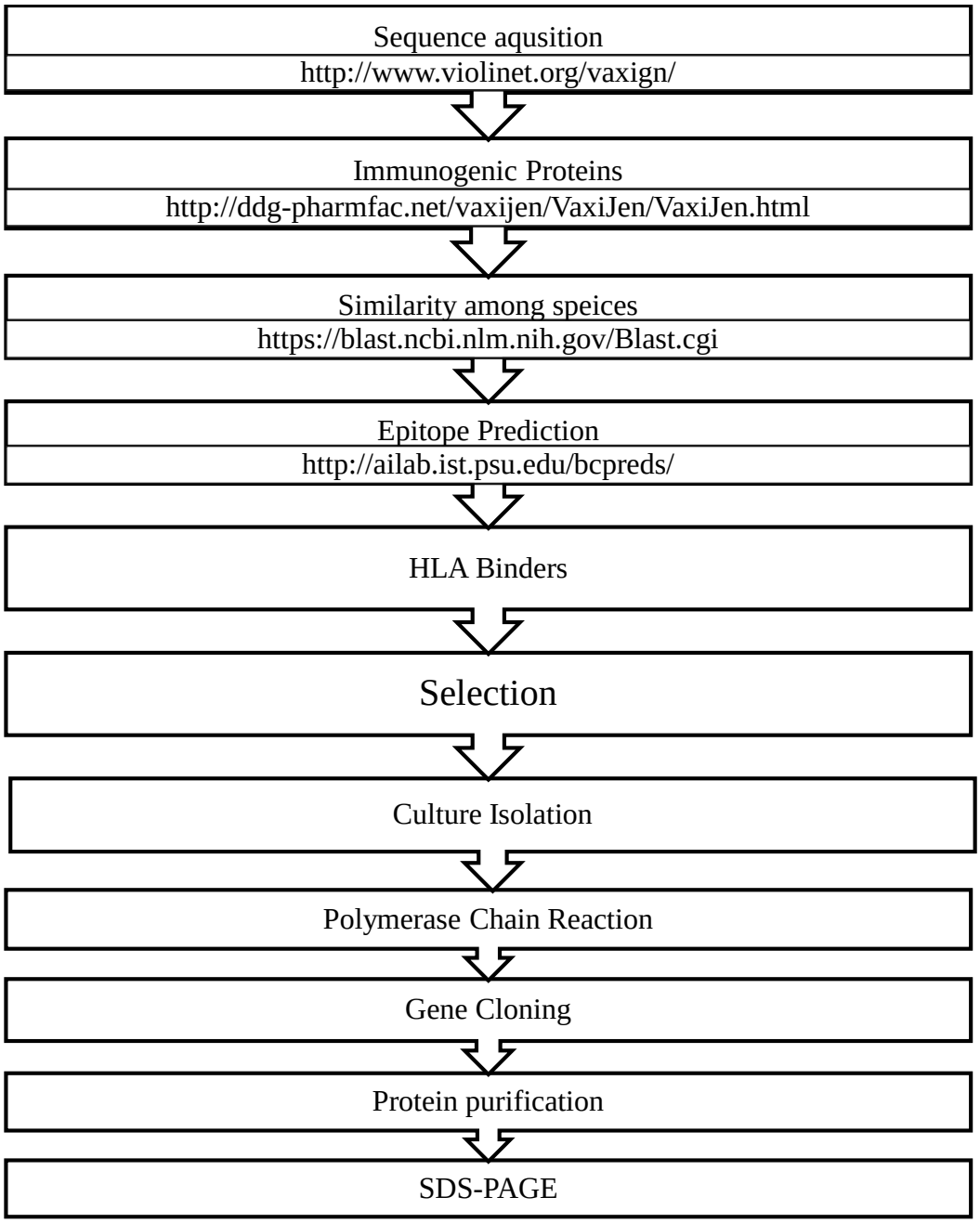

Fig1. Workflow used in this experiment.

1.1 History of *Shigella*

The first organism that was identified from the genus *Shigella* was *Shigella dysenteriae type 1.* This bacterium was named after Kiyoshi Shiga, a Japanese scientist who was the first person to identify and examine dysenteric stools of patients and managed to isolate the bacteria. After the identification of *S. dysenteriae,* the next few decades saw the identification of other species such as *S. flexneri, S. sonnei* and *S. boydii* that were named according to their lead workers *(Taneja and Mewara, 2016).*

Modern medical and public health efforts in controlling the global spread of shigellosis has not been very successful especially in the developing countries. During the 1970s dehydration related with diarrhea could be controlled by giving the patient a strong dose of oral rehydration thereby reducing the mortality of the problem. However oral rehydration had very little benefits to individuals who were infected with invasive bacterial enteropthogens such as *Shigella* thereby making dysentery a clinical problem in developing countries *(Kotloff et al., 1999).* In Bangladesh between the year 1975 and 1985 the data collected by diarrheal disease center showed that the deaths attributed to acute or chronic dysentery between age groups 1-4 years outnumbered the deaths attributed to acute or chromic watery diarrhea by a factor ranging from 2.1 to 7.8 *(Bennish and Wojtyniak, 1991).*

As time progressed antimicrobial treatments became available that could be used to treat shigellosis if diagnosed at an early stage, but due to uncontrolled use and spread of antimicrobial agents the organism has developed extraordinary prowess in acquiring plasmid-encoded resistance to the antimicrobial drugs that were considered as the first line of therapy. Few of these first line drugs that were considered to be highly effective

against this organism were sulfonamides, tetracycline, ampicillin and trimethoprim-sulfamethoxazole however these drugs are currently impotent against *shigellosis (Kotloff et al., 1999).*

1.2 Classification

Shigella belongs to the Enterobacteriaceae family and the in species is classified into different serotypes based on their biochemical difference and variations in O-antigen. There are 4 serogroup A to D, group A (*S. dysenteriae*) has 17 serotypes, group B (*S. flexneri*) has 14 serotypes, group C (*S. sonnei*) has a single serotype and group D (*S. boydii*) has 20 serotypes. Each of these serogroup differ in their distributions *(Taneja and Mewara, 2016).* A multi-centric study from six Asian countries showed that 68% of isolated species in China, Vietnam, Bangladesh, Pakistan and Indonesia were *S. flexneri*, whereas in Thailand the most common isolated species (84%) was found to be *S. sonnei*, while only 4% of *S. dysenteriae* isolates are common in southern Asia and sub-Saharan Africa. This heterogeneous distribution of *Shigella* species and its serogroup indicates that a multivalent or a cross-protective vaccines is required to control the burden of shigellosis in these developing countries *(von Seidlein et al., 2006).*

1.3 An etiology of *Shigella*

Shigella is a gram negative, lactose negative, facultative intracellular pathogen that is closely related to *Escherichia coli (E. coli).* This organism is the etiological agent of bacillary dysentery or shigellosis. The chromosomal genes of this organism controls the two most studied pathogen associated molecular patterns (PAMPS): the lipopolysaccharide (LPS) on the bacterial surface and the peptidoglycan (PGN) between

the inner and outer membrane. The virulence is based on the presence of large virulence plasmid, that encode a Type III secretion system (T3SS), and a set of secreted protein that are the virulence effectors.

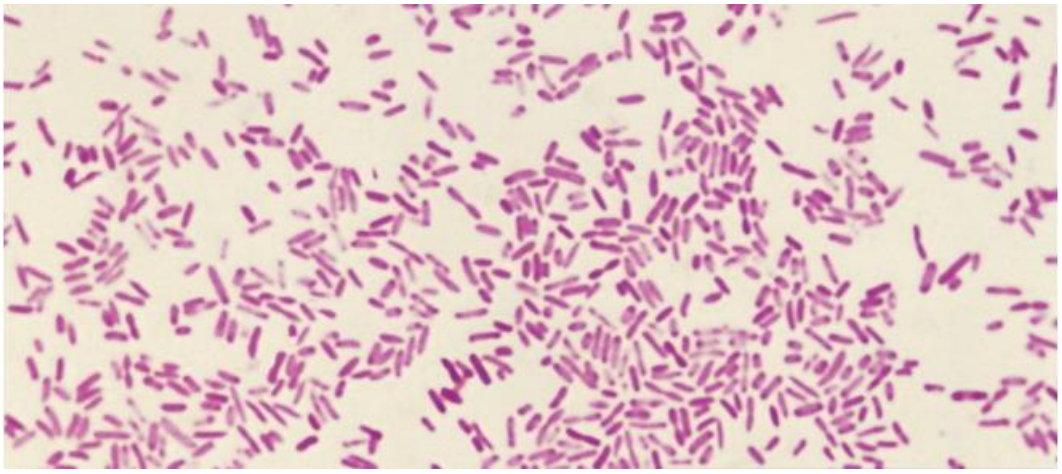

Fig 2: Rod shape bacilli of Shigella Spp. (Getty Images, 2017)

1.4 Epidemiology of *Shigella*

Shigella is responsible for causing bacillary dysentery in children below the age of 5 and elderly individuals with an immunocompromised immune system. It's one of the major species with high morbidity and mortality rate, nearly 165 million causes of shigellosis and about 1 million associated death are reported annually worldwide most of this reports come from developing countries such as sub-Saharan Africa and south Asia *(Kotloff et al., 1999; Kotloff et al., 2013). Shigella* species are endemic in temperate and tropical climates *(Prevention, 2017).*

Its mode of transmission is mostly through the fecal-oral route and requires a minimum dose of less than 10 to 100 bacilli to cause an infection *(Banga Singh et al., 2011).* The impact of *Shigella* on developing countries is higher compared to developed countries, this is mainly because of the poor infrastructure and hygienic conditions. The most

common route of transmission in developing countries is through consumption of contaminated water, food and unhygienic practices where as in develop countries most cases are transmitted by fecal-oral spread from people with symptomatic infection. Outbreaks in developed countries are most commonly related with institutions such as day care centers, schools etc. *(Rabia Agha, MDMarcia B Goldberg, 2017).*

Out of all the four species of *Shigella*, shigellosis in industrial countries is mostly caused by *S. sonnei*, whereas in developing countries the most common organism responsible for shigellosis is *S. flexneri.* There have been several reports of outbreaks across many Asian countries such as Bangladesh (1972-1978, 2003), Sri Lanka (1976), Maldives (1982), Nepal (1984-1985), Chandigarh India (2003). During the 1984 outbreak in West Bengal and Tripura approximately 3, 50,000 people were affected with 3500 death giving an indication that *Shigella* has the ability to cause an outbreak involving a large population. Multiple Drug resistant *S. dysenteriae* serotype 1 is the current cause of epidemic dysentery faced by developing countries *(Taneja and Mewara, 2016).*

1.5 Multiple Drug Resistance (MDR) to *Shigella*

Multiple drug resistance (MDR) to *Shigella* has been reported all over the world including countries like USA, China, Iran, Indonesia and many more *(Taneja and Mewara, 2016)*. The first drug which was used against *Shigella* was sulphonamide that was introduced in the early 1940's and was highly sensitive against all *Shigella* strains. However by the late 1940's it became ineffective, thereby using other drugs like tetracycline, chloramphenicol, ampicillin, co-trimoxazole, nalidixic acid and fluoroquinolones. *Shigella* became resistant to tetracycline and chloramphenicol in 1980's followed by nalidixic acid in 1988 due to

the outbreak in Tripura, which made way for fluoroquinolones as the drug of choice, which was very effective against MDR *Shigella*.

In 1990 ciprofloxacin proved to be effective against shigellosis, yet due to uncontrolled used of such fluoroquinolones resulted in the emergences of resistance strains against these antibiotics as well. Currently the World Health Organization (WHO) recommends ceftriaxone, pivmecillinam and azithromycin as an alternative to fluoroquinolone resistant *Shigella.* However over the period of nine years (2000-2009) resistance to at least one of the third generation cephalosporin's (ceftriaxone/ cefotaxime) has been noted. First isolate showing resistance to ceftriaxone was seen in 2001 followed by an increase in resistant strains by 2005 and currently on the rise, thereby making therapeutic treatment a challenge *(Taneja and Mewara, 2016).*

1.6 Clinical Manifestation

The clinical manifestation of shigellosis begins with most common symptoms such as fever, anorexia and malaise. Initially the diarrhea of the infected patient is watery, but as the disease progresses it may contain blood and mucus. Usually the incubation period required for the organism to develop the symptoms ranges from one to seven days, with an average of three days *(Rabia Agha, MDMarcia B Goldberg, 2017).* The severity of the disease depends on the serogroup of the infecting organism. *S. sonnei* commonly causes mild disease, which is limited to watery diarrhea, while *S. dysenteriae* or *S. flexneri* are most commonly responsible for bloody diarrhea *(Khan, Griffiths and Bennish, 2013).* In a normal host the disease is generally self-limited, when left untreated will last of less than seven days.

There is a possibility that during a shigellosis infection there may be rare cases of intestinal and systemic complications such as in table 1.1:

Table1.1: Complications associated with Shigellosis

Intestinal Complication	**Systemic Complications**
Proctitis/ rectal prolapse	Bacteremia
Toxic megacolon	Metabolic disturbances
Intestinal obstruction	Leukemoid reaction
Colonic Perforation	Neurologic disease
	Reactive arthritis
	Hemolytic-uremic syndrome

Few other complications of shigellosis include vaginitis or vulvovaginitis with or without diarrhea which occurs in young girls. In very rare cases *Shigella* may cause Keratitis or Conjunctivitis in young children who have recently been exposed to the illness. *Shigella* may also cause acute myocarditis. However these complications differ from species to species *(Rabia Agha, MDMarcia B Goldberg, 2017).*

1.7 Host and Risk groups

The primary reservoirs for *Shigella* are humans and primates. It has been isolated from various sources viz. aquatic bodies (rivers, surface waters), free living amoebae, insects, birds and wild animals *(Taneja and Mewara, 2016).*

Risk group associated with *Shigella* are mostly children who are less than five years of age. This pathogen can have a serious effect especially on children's who are

malnourished causing further impairment of nutrition, growth retardation and recurrent infection. In developed countries the individuals who are most often infected are migrant populations, travelers to developing countries, children in day-care facilities, prisoners and military personnel, and homosexual men. However no individual is immune to shigellosis especially during an epidemic outbreak where all age groups are at the risk of acquiring the infection *(Taneja and Mewara, 2016).*

1.8 Route of transmission

In the developing countries, high incident of shigellosis is mainly due to lack of clean water, poor hygiene conditions, malnutrition and close personal contact. However the most common route of transmission is by fecal oral route. There have been outbreaks that have been associated with person to person especially seen in crowded or unhygienic environments like prisons and asylums (Taneja and Mewara, 2016). Transmission can also be triggered by environmental factors such as rainfall and temperature. It has been seen that the organism can express its virulence gene when the temperature shifts from 30^{o}C to 37^{o}C when left in a medium of moderate osmotic stress with a pH of 7.4 (Dorman and Porter, 1998). Outbreaks associated with *Shigella* can occur round the year but higher cases of outbreaks are mostly seen during summer, this may be due hot and dry weather possibly due to lack of water thereby limiting hygiene conditions like washing hands. Another route of transmission is food, especially foods and beverages that are prepared by street-vendors. Thereby making food-handlers as a source of transmission. Areas with inadequate facilities for disposal of human faeces can provide a breeding ground for flies particularly *Musca domestica*; the common housefly, thereby making them a vector for transmission *(Taneja and Mewara, 2016).*

1.9 Pathogenesis

S. flexneri requires less than 100 cells to cause an infection, this is mainly because the organism has the ability to survive in the low acidic conditions of the host's stomach via an up-regulation in acid resistant genes. The clinical symptoms of shigellosis is seen when the organism successfully destroys the colonic epithelium of the host. To be able to destroy the epithelium the organism must first reach the colon, invade, penetrate, replicate within the mucosa and spread between the mucosal epithelial cells resulting in a subsequent inflammatory response of the host causing the destruction of the colonic epithelium *(Jennison and Verma, 2004).*

Through the follicular associated epithelium (FAE) which is located above the mucosa-associated lymph nodes the organism penetrates the epithelial lining. This region consists of endocytic M cells (Membranous epithelial cells) which have transcytotic (transcellular transport) properties. Many organism including *Shigella* exploit of this property of the M-cells and use it as a route for invasion of the impermeable epithelial lining *(Jennison and Verma, 2004).*

In order for the organism to amplify its penetration into the colonic epithelium, it exploits the host's immune response. The macrophages that are infected by the organism undergo apoptosis, releasing large amounts of IL-1 (Interleukin-1), who play an important role in inducing inflammatory response. The invasion of the epithelial cells by the organism activates the transcription and secretion of IL-8, which is responsible in recruiting Polymorphonuclear cells (PMN) to the luminal bacteria in the infected area. This influx of PMN disrupts the integrity of the epithelium, allowing the luminal bacteria to cross the

submucosa and cause further inflammation and destruction of tissues *(Jennison and Verma, 2004).*

1.10 Diagnosis, Treatment and Prevention

Based on typical clinical suspicion like fever, pain, watery or bloody diarrhea in patients who are likely to have been exposed to the bacteria makes a doctor suspicious of shigellosis. Preliminary test includes checking for the presence of white blood cells and red blood cells on direct microscopic examination of the stool sample which is consistent with the diagnosis of *Shigella*. To confirm the diagnosis, stool sample is taken from the patients and sent to the laboratory for further test for organism identification; such as stool culture, molecular testing and susceptibility testing (*Shigella infection - Diagnosis and treatment - Mayo Clinic*, no date; *Rabia Agha, MDMarcia B Goldberg, 2017).*

In a healthy individual *Shigella* may be treated by giving the patients lots of fluids containing salts in order to replace the lost fluids. Antibiotics may be used to shorten the duration of infection with the exception if the strain is multiple drug resistant than antibiotics should be avoided (*Shigella infection - Diagnosis and treatment - Mayo Clinic*, no date, *Shigellosis - Infections - Merck Manuals Consumer Version*, no date).

The best way to avoid the disease is to use precaution methods like: (i) Follow proper personal and sanitary hygiene; (ii) infected individuals should not be allowed to prepare food for others; and (iii) infected children should not be allowed to interact with other healthy children till they are cured (*Shigellosis - Infections - Merck Manuals Consumer Version*, no date). One of the best way to prevent the disease is to design safe and effective vaccines that can be used to combat the infection.

1.11 Overview on vaccine design

The journey of development of vaccines began in 1796 when Edward Jenner created the first vaccine to protect human from smallpox. This work became the foundation for the design of various vaccines. Vaccines against bacterial infections such as diphtheria, tetanus, anthrax, cholera, plague, typhoid, tuberculosis were developed in 19th century *(Anon., 2011)*. By the middle of 20th century, vaccine research and development became an active area of interest which led to the development of vaccines against viral infections like polio, measles, mumps, and rubella *(Serruto and Rappuoli, 2006; Movahedi and Hampson, 2008; Bambini and Rappuoli, 2009)*.

While effective vaccines have been developed with conventional methods (Figure 3) to fight against microbial infections, there are still some limitations yet to overcome particularly: (i) not all pathogens can be grown in culture (ii) some cell associated microorganisms require specific cell cultures for growth (expensive cost) (iii) extensive safety procedure for personnel and environment may be required to prevent exposure to pathogenic microorganisms (iv) insufficient killing or attenuation may result in the introduction of virulent organisms into the final vaccine and inadvertently cause disease (v) time consuming *(Movahedi and Hampson, 2008; Bambini and Rappuoli, 2009)*.

Therefore, innovative techniques have been devised to design vaccines that are safe, more efficacious, and less expensive than traditional vaccines, also there has been an expansion in the disease targets, scientists now are not only focusing on developing vaccines for infectious diseases but are also focusing on designing vaccine for non-infectious conditions like allergies and cancers.

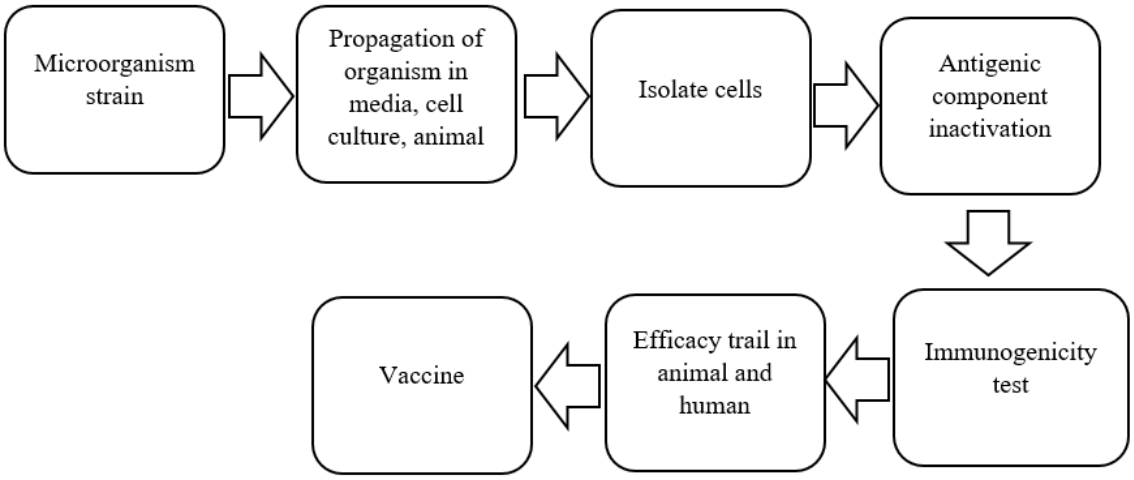

Fig 3: Schematic representation of the conventional approach to bacterial vaccine development.

1.12 The *in silico* Approach

In silico approach provides computational methods to predict and design new vaccine components. Two promising *in silico* fields that have widely contributed to the development of vaccine are bioinformatics and immunoinformatics. Bioinformatics is a field of science that represents the convergence of several different disciplines in order to organize and store large amounts of biological information which are generated from experiments conducted in the field of genetics, molecular biology and biotechnology *(Soria-Guerra et al., 2015).*

The field of bioinformatics comprises of various studies including proteomics, genomics and vaccinomics. The information obtained from these disciplines speed up the identification and characterization of new antigens. Over the years, sequencing experiments have generated a large amount of data relevant to immunology research. With recent advancement in research methodologies, huge amounts of functional, clinical and

epidemiologic data are being reported and deposited in various specialist repositories and clinical records *(Soria-Guerra et al., 2015)*. Combining this accumulated information together can provide insights into the mechanisms of immune function and disease pathogenesis. Hence, the need to handle this rapidly growing immunological resource has given rise to the field known as immunoinformatics *(Tomar and De, 2014; Soria-Guerra et al., 2015)*.

The main objective of immunoinformatics is to convert large-scale immunological data, using a wide variety of computational, mathematical and statistical methods that range from text mining, information management, sequence analysis and molecular interactions to understand and organize these large scale data to obtain immunologically meaningful interpretations. Also attempts are being made for the extraction of interesting and complex patterns from non-structured text documents in the immunological domain, including categorization of allergen cross-reactivity information, identification of cancer-associated gene variants, and the classification of immune epitopes *(De and Tomar, 2014)*. The tools used in this field are based on statistical and machine learning system and are used for studies in modeling molecular interaction. They also play a role in defining new hypothesis related to understand the immune system mechanism *(Backert and Kohlbacher, 2015)*.

Advanced knowledge in the field of infectious disease such as pathogenesis, virulence factor and the host immune response has assisted us in overtaking the traditional method of vaccine development *(Movahedi and Hampson, 2008)*. The advancement in the field has led to the development of second generation vaccines. Later, with the integration of the development in bioinformatics tools along with the advancement in recombinant DNA

technologies (rDNA), such approach provides another series of advancement in the field of vaccine development resulting in the rises of third generation vaccines known as Reverse Vaccinology. The Reverse Vaccinology approach coined by Rino Rappuoli was first developed to develop potential pathogenic microbial vaccine based on genomic data *(Movahedi and Hampson, 2008; Bambini and Rappuoli, 2009)*. It starts from the whole genomic sequence and through computer analysis to predict those proteins that are most likely to be effective vaccine components. This approach is widely applicable as there are many genome sequences available publicly. It was firstly applied for the development of vaccines against serogroup B *Neisseria meningitids (Rappuoli, 2000)*.

In 2008, a cutting-edge reverse vaccinology program known as Vaxign was successfully developed. Typically, the program comprises of a comprehensive pipeline that using bioinformatic technology to find potential genes from the genomes for developing vaccines. The program has also been developed to predict possible antibody targets based on various criteria by utilizing microbial genomic and protein groupings as data information *(Xiang and He, 2009)*. The major predicted features include subcellular location of proteins, transmembrane domain, adhesion probability, sequence similarity to host proteome and MHC class I and II epitope binding *(Xiang and He, 2009)*. Out of these features, subcellular localization is considered to be as one of the main criterion for target prediction *(He, Xiang and Mobley, 2010)*. This program is a part of web based system called Vaccine Investigation and Online Information Network (VIOLIN, http://www.violinet.org). Vaxign has widely been used for the prediction of vaccine targets against bacteria like *Brucella spp.*, *Neisseria meningitides* and *Mycobacterium tuberculosis* since it was developed *(Xiang and He, 2009; He, Xiang and Mobley, 2010)*.

1.13 From genomics to epitope prediction

High throughput human leukocyte antigen (HLA) binding assay and Next Generation Sequencing (NGS) has led to major progress in this field exclusively on epitope prediction tools development *(Backert and Kohlbacher, 2015)*. Basically, the immunogenicity of an antigen is associated with its ability to interface with the humoral (B-cell) and cellular (T-cell) immune systems. An immunogen construct containing both B- and T-cell epitopes is crucial to effectively induce strong immune responses when in contact with host immune system *(He et al., 2010; Moise et al., 2011)*. The progression from genomics to epitope prediction has led us toward the evolution of designing epitope based vaccines.

Epitopes hold huge potential for vaccine design, disease prevention, diagnosis and treatment and are therefore of particular interest both clinical and basic biomedical researchers *(Soria-Guerra et al., 2015)*. With recombinant DNA technologies (rDNA) it is possible to isolate specific epitopes which replace the whole pathogen in a vaccine, also can design vaccines that consist of chimeric proteins (multi-epitope vaccine) in order to enhance a strong immune response. These features offer the possibility of designing multi-target highly efficient vaccine, but it is not an easy process as proper identification of the immunogenic peptide is necessary in order to elicit an immune response *(Soria-Guerra et al., 2015)*. A good epitope-based vaccine should possess two features: - (i) it should be known if the selected epitope for the vaccine is conserved across different stages of the pathogen and its variant; (ii) the desired immune response should also be taken into consideration. All the information mentioned above for the design of epitope-based vaccine can now be performed using immunoinformatic approach (Figure 4).

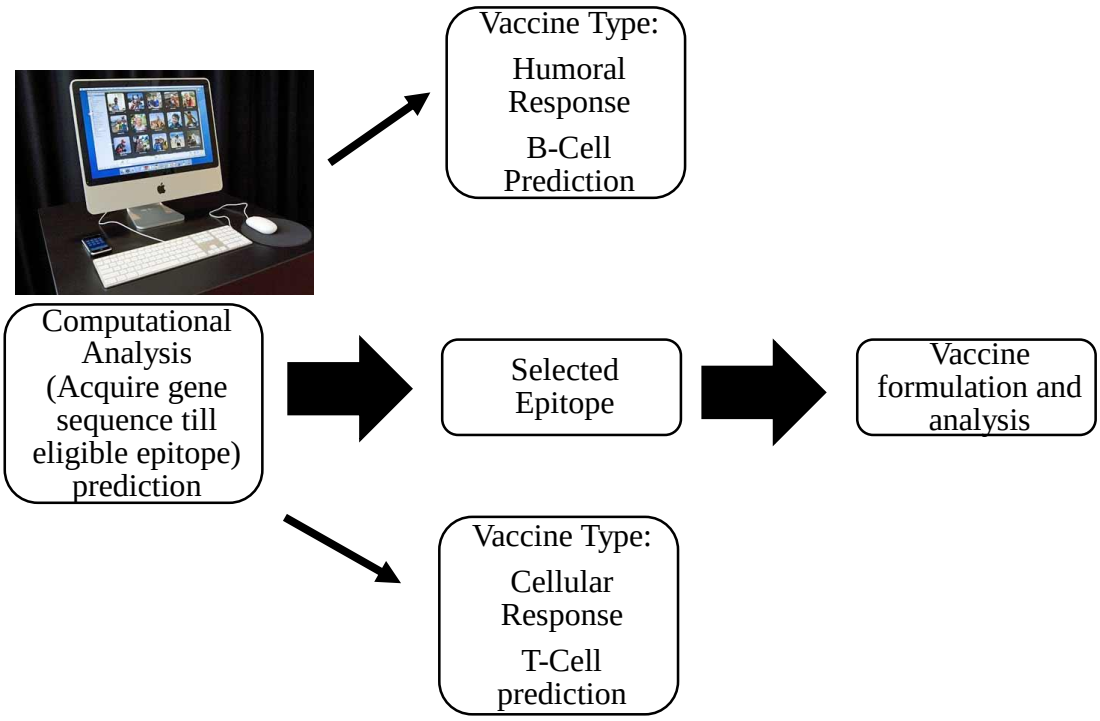

Fig 4: Schematic workflow to identify epitope for vaccine development.

1.14 *B*-cell epitope prediction and tools

Identification of B-cell epitopes plays an important role in development of epitope-based vaccines, therapeutic antibodies, and diagnostic tools. Surface immunoglobulins of B-cells and antibodies recognize the B-cell epitope of an antigen in its native conformation. There are two types of B-cell epitope which are known as continuous (linear/sequential) or discontinuous (conformational). In general, B-cell epitopes bind to B-cell receptors (BCR) and the binding site of these receptors (BCR) are primarily hydrophobic, consisting of 6 hyper variable loops of varied lengths and amino acid composition *(Tong and Ren, 2009; Tomar and De, 2014; Soria-Guerra et al., 2015).*

Prediction of B-cell epitopes is primarily based on its amino acid properties such as hydrophilicity, the charged exposed surface area and secondary structure. It is estimated

that about 85% of documented B-cell epitopes could be considered to be continuous in sequence, but it should be noted that not all residues within an epitope are functionally essential for binding, and this binding efficacy depends on the amino acid sequence such that, if a single amino acid is substituted or eliminated then the binding efficacy is reduced *(Kringelum et al., 2013; Tomar and De, 2014; Backert and Kohlbacher, 2015; Soria-Guerra et al., 2015).*

In contrast, in discontinuous B-cell epitopes the discontinuity is because distant residues are brought into spatial proximity by protein folding. These complex structures of folded proteins lead to spatial proximity of amino acids that can be remote in the antigen sequence. Prediction of discontinuous B-cell epitope is difficult because two reasons (i) not many prediction software are available for discontinuous B-cell epitopes and (ii) because classic machine learning based system require a continuous sequenced data. *(Saha, Bhasin and Raghava, 2005; Tong and Ren, 2009; Tomar and De, 2010; Backert and Kohlbacher, 2015).*

Selected specific web-based servers available for the prediction of continuous or discontinuous B-cell epitopes are listed in Table 1.2. These integrative B-cell prediction tools shed light on the identification of immunogenic B-cell epitopes which are potentially to trigger antibody response in the host immune system *(Soria-Guerra et al., 2015).*

Table 1.2: List of some servers used for the prediction of B cell epitopes

Server name	Type of prediction	Link	Reference
BcPred	Continuous B cell epitopes	http://ailab.ist.psu.edu/bcpreds/predict.html	*(Chen et al., 2007; El-Manzalawy, Dobbs and Honavar, 2008a, 2008b)*
BepiPred	Continuous B cell epitopes	http://www.cbs.dtu.dk/services/BepiPred/	*(Jespersen et al., 2017)*
ABCPred	Continuous B cell epitopes	http://www.imtech.res.in/raghava/abcpred/	*(Saha and G P S Raghava, 2006)*
BcePred	Continuous B cell epitopes	http://www.imtech.res.in/raghava/bcepred/	*(Saha and Raghava, 2004)*
EPCES	Discontinuous B cell epitope	http://sysbio.unl.edu/EPCES/	*(Liang et al., 2007)*
DiscoTope	Discontinuous B cell epitope	http://www.cbs.dtu.dk/services/DiscoTope/	*(Kringelum et al., 2012)*
BEPro (PEPITO)	Discontinuous B cell epitope	http://pepito.proteomics.ics.uci.edu/	*(Sweredoski and Baldi, 2008)*
Ellipro	Discontinuous B cell epitope	http://tools.iedb.org/ellipro/	*(Ponomarenko et al., 2008)*

1.15 *T-cell* epitope prediction and tools

T-cell epitopes are parts of intracellular processing antigens that are presented to T lymphocytes in association with molecules of the major histocompatibility complex (MHC). Antigenic peptides bound with MHC are presented to T-cell for immunity stimulation. Therefore identification of these MHC binding peptides is a central part of any algorithm that is used to predict T- cell epitopes *(Tomar and De, 2014)*. T-cell epitopes are bound in a linear form to MHCs, the interface between ligands and T-cells can be modeled with accuracy. These epitopes are linked together into the binding grove of MHC class I and class II molecules, through interaction between their R group side chain and

pockets located on the floor of the MHC *(Falk et al., 1991; Rötzschke et al., 1991; Zinkernagel and Doherty, 1997; Soria-Guerra et al., 2015)*. Hence based on this knowledge large number of computational tools have been established and used for identifying putative T-cell epitopes and MHC binding peptides (Table 1.3). The identification of these putative T-cell epitopes and MHC binding peptides is based on various algorithms such as artificial neural network (ANN), protein threading and docking techniques, hidden Markov model (HMM) and decision tree *(Tong and Ren, 2009).*

Among all the MHC I and MHC II binding predictor software's, MHC class I predicators have displayed to be more efficient in their prediction with an estimated accuracy of 90-95 %. Among all the available servers for MHC alleles, RANKPEP is one server which makes use of Position Specific Scoring Matrices (PSSMs) to predict peptide binders to MHC I and MHC II molecules from protein sequences. It is a friendly platform which offers a wide range of allelic coverage to MHC I and MHC II alleles for human and mouse. The algorithm used by this database is written in Python, and it sorts and stores all protein segments with the length of the PSSM width. The scoring starts at the beginning of each sequence and the PSSM is slid over the sequence one residue at a time until reaching the end of the sequence. Furthermore a threshold value is set in order to narrow down the potential binders from the list of ranked peptides; a binding threshold is defined as a score value that includes 90% of the peptides within the PSSM. This binding threshold value is built into each matrix, delineating the range of putative binders among the top scoring peptides *(Reche, Glutting and Reinherz, 2002; Reche et al., 2004; Reche and Reinherz, 2007).*

Kernel–based Inter–allele peptide binding prediction system (KISS), predicts if a 9-mer peptide will bind to a MHC I molecule of 64 alleles using a support vector machine (SVM) multitask kernel to leverage the available training information across the alleles, which improves it accuracy especially for the alleles with few known epitopes. The predictor is trained on databases which contain known epitopes from SYFPEITHI, MHCBN and IEDB databases *(El-Manzalawy, Dobbs and Honavar, 2008b)*. There are many servers available for the identification of MHC I binding predictors, the most complete server in terms of allelic coverage and identification of alleles were as mentioned above. These servers can also identify alleles in other organism besides humans.

Though MHC class I prediction seems to have achieved a good success, it's not the same for MHC class II. MHC Class II prediction has achieved a very limited success in predicting the potential binding epitopes. This is mainly because the predicting accuracy is low. There are several factors which is the reason for low accuracy, such as insufficient or low quality training data, difficulty in identifying 9-mer binding cores within longer peptides used for training and lack of consideration of the influence of flanking residues and the relative permissiveness of the binding groove of MHC II molecules which limit the binding stringency. For instance, Propred is a server used for the prediction of MHC Class II binding regions present in an antigen sequence using quantitative matrices. The server implement matrix based prediction algorithm, employing amino-acid/position coefficient table deduced from literature. This server assists in locating promiscuous binding regions that are useful in selecting vaccine candidates that can bind to 51 HLA-DR alleles *(Sturniolo et al., 1999; Singh and Raghava, 2002).*