STRUCTURE PREDICTION AND SCFV DESIGN FOR *Bm*R1 ANTIGEN OF *Brugia Malayi*

BY

KHOR BEE YIN

UNIVERSITI SAINS MALAYSIA

2014

STRUCTURE PREDICTION AND SCFV DESIGN AGAINST *BM*R1 OF *BRUGIA MALAYI*

BY

KHOR BEE YIN

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

December 2014

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my supervisor Dr. Choong Yee Siew for her invaluable guidance and encouragement throughout my studies. Thank you for being more than just a supervisor, giving continuous advices, motivations and inspiring ideas when facing bottleneck in the study. I am also thankful to my co-supervisor Dr. Lim Theam Soon and Prof. Rahmah binti Noordin for their constant support and encouragement in making this study successful. My sincere thanks also go to Dr. Tye Gee Jun for spending time in giving comments throughout the study.

I would like to thank the laboratory mate Yie Vern and Syazana for their helping hands in guiding me when I am still new to this field and their suggestion in problem solving. Not to forget Kevin, Tommy, Roy and Siew Wen for their time during our discussion sessions. Their contributions in giving advices and comments are greatly appreciated and acknowledged. I am also thankful to all INFORMMERs, including all the staffs, lecturers and students, for all the supports and helps when I am facing difficulties.

I would also like to thank Exploratory Research Grant Scheme (ERGS) (203/CIPPM/6730058) and Higher Institutions Centre of Excellence (HICoE) Grant (311/CIPPM/44001005) from the Malaysia Ministry of Education for funding the research. Thanks also due to MyBrain by Ministry of Education for the scholarship.

Last but not least, a heartfelt of appreciation goes to my family for their love and unfailing support that drove me to complete this study successfully.

A million thanks to everyone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

3D	Three-dimensional
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
atm	Atmospheric pressure
BLAST	Basic Local Alignment Search Tool
CASP	Critical assessment of protein structure prediction
CD Search	Conserved Domain Search Service
CDR	Complementarity-determining region
CDR H1	Complementarity-determining region heavy chain 1
CDR H2	Complementarity-determining region heavy chain 2
CDR H3	Complementarity-determining region heavy chain 3
CDR L1	Complementarity-determining region light chain 1
CDR L2	Complementarity-determining region light chain 2
CDR L3	Complementarity-determining region light chain 3
Cys	Cysteine
DOPE	Discrete Optimized Protein Energy
DUF148	Domain of unknown function 148
ELISA	Enzyme-linked immunosorbent assays
ExPASy	Expert Protein Analysis System
FAR	Fatty acid and retinol
FM	Template-free modelling
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
GPELF	Global Programme for the Elimination of Lymphatic Filariasis
His	Histidine
HIV-1	Human immunodeficiency virus type 1
HMM	Hidden-Markov models
I-TASSER	Iterative Threading ASSembly Refinement
ICT	Immunochromatographic test
Ile	Isoluecine
L3	Third stage larvae
Leu	Leucine
LF	Lymphatic filariasis
LOMETS	Local Meta-Threading-Server
Lys	Lysine
MD	Molecular dynamics
Met	Methionine
Mf	Microfilariae
molpdf	MODELLER objective function
NCBI	National Center for Biotechnology
NMR	Nuclear magnetic resonance
nr	non-redundant
ns	Nanosecond
PCR	Polymerase chain reaction
PDB	Protein Data Bank

Pfam	Proteins Families database
Phe	Phenylalanine
PPA	Profile-profile alignment
Pro	Proline
ps	Picosecond
R.E.U	Rosetta energy unit
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
scFv	Single chain variable fragment
SCOP	Structural Classification of Proteins
Ser	Serine
TBM	Template-based modelling
Thr	Threonine
TIP3P	Transferable intermolecular potential 3P
Trp	Tryptophan
Tyr	Tyrosine
V_H	Variable regions of heavy chain
V_L	Variable regions of light chain
WHO	World Health Organization
α	Alpha
β	Beta

LIST OF APPENDICES

LIST OF PUBLICATIONS

Page

RAMALAN STRUKTUR DAN REKA BENTUK SCFV UNTUK *BM*R1 ANTIGEN DARIPADA *BRUGIA MALAYI*

ABSTRAK

Filariasis limfatik yang disebabkan oleh salah satu cacing filaria, *Brugia malayi*, telah dikenalpasti oleh Badan Kesihatan Sedunia untuk dihapuskan sejak tahun 1997. Antigen rekombinan daripada *B. malayi* (*Bm*R1) telah menunjukkan spesifisiti dan sensitiviti yang tinggi dalam ujian pengesanan antibodi terhadap filariasis brugian (Brugia Rapid). Maka, antigen ini merupakan calon yang unggul untuk pembangunan ujian pengesanan antigen untuk melengkapkan Brugia Rapid. Dalam kajian ini, pendekatan siliko telah dilaksanakan untuk mendapat struktur tiga dimensi bagi antigen *Bm*R1, mengkaji domain fungsi, meramalkan epitop yang berpontesi dan reka bentuk rantai tunggal fragmen Fv (scFv) spesifik terhadap *Bm*R1. Di sini, ramalan struktur telah diperolehi melalui pendekatan pemodelan perbandingan, 'threading' dan *ab initio*. Penilaian menunjukkan bahawa struktur yang diramalkan dengan pendekatan *ab initio* (Rosetta) paling optimum di kalangan semua struktur. Struktur yang telah diperhaluskan dan dioptimalkan lagi telah menunjukkan kestabilan dan kualiti yang baik sepanjang 5 ns dinamik simulasi molekul. Tiga epitop berpotensi (residu 37-47, 104-112 dan 125-148) telah diramalkan oleh server epitop linear dan konformasi. ScFv spesifik terhadap epitop-epitop ini telah direka melalui dua langkah utama: generasi 'hot spot libraries' dan konfigurasi bentuk saling melengkapi. Sebanyak 200 reka bentuk dengan tenaga ikatan (<-20 R.E.U) and bentuk saling melengkapi (Sc>0.5) telah diperolehi. Residu-residu penting dalam peningkatan interaksi dengan antigen *Bm*R1 telah dikenalpasti. Reka bentuk scFv ini diramalkan sangat spesifik terhadap antigen *Bm*R1 dan mungkin dapat digunakan dalam ujian *in vitro* sebelum pembangunan ujian pengesanan berdasarkan antigen.

STRUCTURE PREDICTION AND SCFV DESIGN AGAINST *BM*R1 ANTIGEN OF *BRUGIA MALAYI*

ABSTRACT

Lymphatic filariasis which is caused by one of the filarial nematode, *Brugia malayi*, has been identified by World Health Organization for elimination since year 1997. The recombinant antigen from *B. malayi* (*Bm*R1) was found to be highly specific and sensitive in the antibody detection test against brugian filariasis (Brugia Rapid). Therefore, this antigen is an ideal candidate for the development of antigen detection test to complement the current Brugia Rapid. In this study, computational approach was implemented to obtain three-dimensional structure of *Bm*R1 antigen, to study the functional domain, to predict the potential epitopes and to design single chain variable fragment (scFv) specific to *Bm*R1. Predictive structures were obtained via comparative modelling, threading and *ab initio* approaches. Evaluation studies showed that the structure built by *ab initio* (Rosetta) was the most optimal compared to structures built by other methods. The further refined and optimized structure was shown to be stable and posses good quality throughout 5 ns of molecular dynamics simulations. Three potential epitopes (residues 37-49, 104-112 and 125-148) were predicted by linear and conformational epitopes prediction server. Specific scFv towards these epitopes were designed using two main methods: the generation of hot spot libraries and high shape complementarity conformations. A total of 200 scFv designs with binding energy (<-20 R.E.U) and shape complementarity (Sc>0.5) were obtained. Key residues that played crucial roles in enhancing the interactions with *Bm*R1 antigen were also identified. These designed scFv were predicted to be highly specific towards the *Bm*R1 antigen and may be useful for *in vitro* testing prior to the development of antigen detection test.

CHAPTER ONE

GENERAL OVERVIEW

1.1 Introduction

Lymphatic filariasis has been identified as one of the leading causes of permanent and long-term disability and has been targeted for elimination by the year 2020 (World Health Organization, 2011). Therefore, the availability of an easy on-site lymphatic filariasis diagnostic test which is rapid, affordable and accessible for disease management and therapy is one of the important factors for the ongoing lymphatic filariasis elimination programme (Noordin *et al.*, 2004; Noordin, 2007; Rahman *et al.*, 2007). Earlier studies showed that *Brugia malayi* recombinant antigen (*Bm*R1) expressed from *Bm17DIII* gene has been employed in ELISA and rapid immunochromatographic dipstick test format (Brugia Rapid) for specific and sensitive detection of anti-filarial IgG4 antibodies against brugian filariasis (Rahmah *et al.*, 2001a; Noordin *et al.*, 2007; Rahman *et al.*, 2007). Studies also showed that the *Bm*R1 recombinant antigen is a highly specific and sensitive marker, thus making it a promising candidate for the detection of IgG4 antibody for brugian filariasis (Noordin *et al.*, 2004; Rahman *et al.*, 2007; Noordin *et al.*, 2007). The development of a specific and sensitive *Bm*R1 antigen detection test would also be a complement to currently available antibody detection test (Brugia Rapid) for detection of active lymphatic filariasis infection (Rahmah *et al.*, 2001b; Noordin *et al.*, 2004; Rahman *et al.*, 2007). Thus, this study is to model and identify epitopes of *Bm*R1 antigen and subsequently to design binders, namely single chain variable fragment (scFv), specific against *Bm*R1 antigen.

1.2 Lymphatic filariasis

Lymphatic filariasis (LF) is a neglected disease caused by filarial nematodes namely *Wuchereria bancrofti*, *Brugia malayi* and *Brugia timori*. *W. bancrofti* (bancroftian filariasis) accounts for 90% of LF infection while the remaining 10% is by *B. malayi* and *B. timori* (brugian filariasis; commonly found in Asia countries). LF, also known as elephantiasis, has infected over 120 million people worldwide and it is estimated that nearly 1.4 billion people are at risk of infection (World Health Organization, 2011). Due to this, WHO has launched the Global Programme for the Elimination of Lymphatic Filariasis (GPELF) in the year 2000 with two principal goals: to interrupt the transmission and to alleviate the disability (reduce morbidity) caused by LF (Ottesen, 2000). In conjunction with the program, National Lymphatic Filariasis Elimination Programme was being introduced in 2001 by Ministry of Health Malaysia and WHO (Ministry of Health Malaysia, 2004; Hamid, 2012). Although statistic showed that the number of LF cases had decreased over the years (from 528 cases in 1999 to 267 cases in 2012) (Ministry of Health Malaysia, 2012), LF still remains as a public health problem in the states of Johor, Kedah, Kelantan, Pahang, Perak, Terengganu, Sabah and Sarawak (Noordin, 2007; Talip, 2014).

LF is transmitted to human by mosquitoes such as *Culex*, *Anopheles*, *Aedes* and *Mansonia* (Erickson et al., 2009). When an infected mosquito (intermediate host) takes a blood, it deposits infective third stage larvae (L3) into human (definitive host). The larvae are then mature and develop into lymph-dwelling adult worms that reside in the lymphatic vessels. Adult female worms produce hundreds to thousands of sheathed microfilariae (Mf) daily and are then ingested by mosquitoes during a blood meal, thereby completing the life cycle of *B. malayi* (Figure 1.1) (Centers for Disease Control and Prevention, 2010).

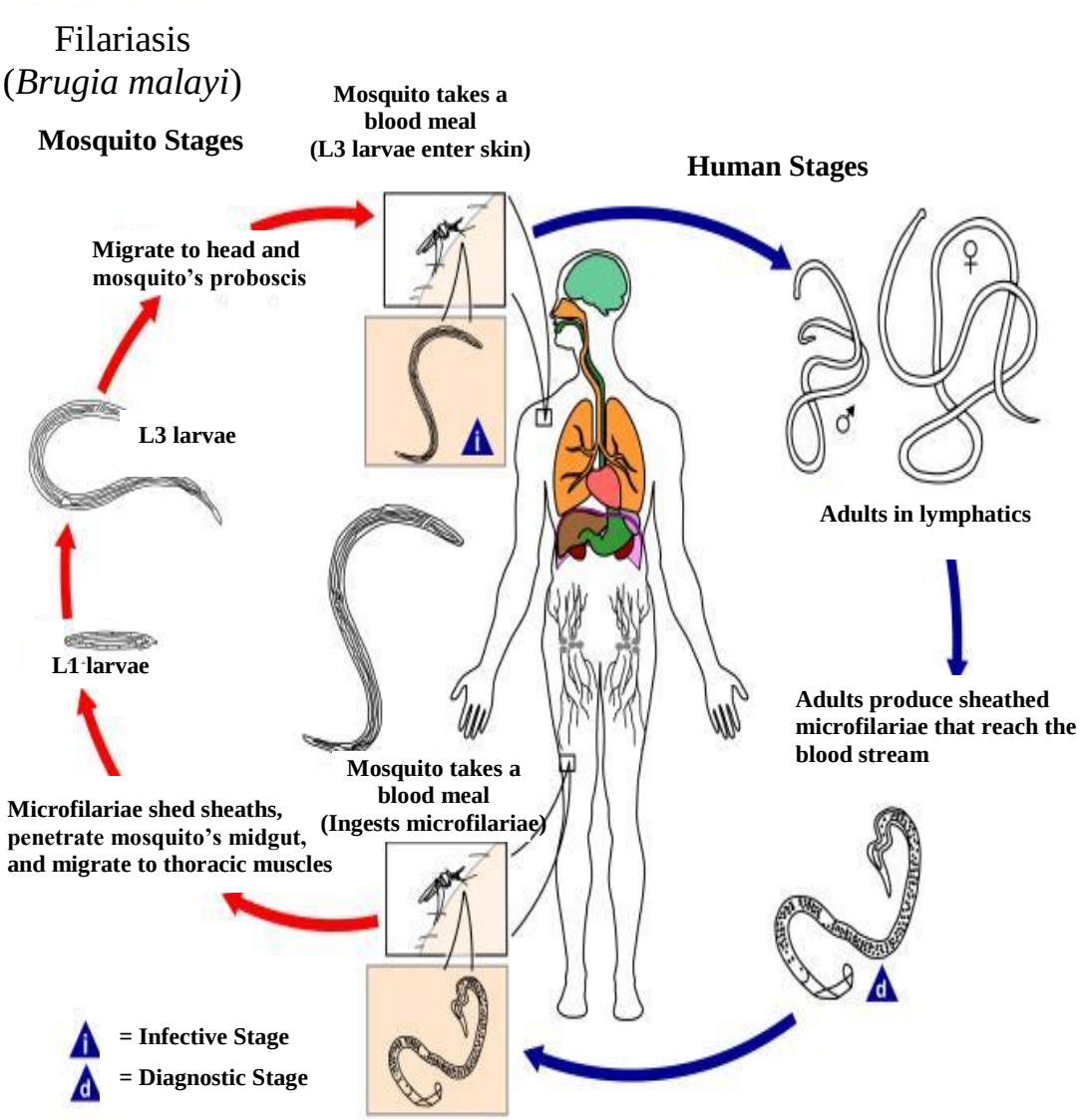

Figure 1.1 Life cycle of *Brugia malayi* in the mosquito (intermediate host) and human (definitive host) adopted from Centers for Disease Control and Prevention, 2010.

1.3 Symptoms and treatment of lymphatic filariasis

LF patients show different clinical symptoms ranging from no evident of clinical disease to overt clinical manifestation. Clinical manifestations of lymphatic filariasis can be further divided into asymptomatic stage, acute manifestation (inflammatory episodes) and chronic manifestation (lymphoedema, elephantiasis and hydrocele). The most common clinical manifestation is the asymptomatic stage. Individuals in this asymptomatic stage are carriers of the infection and the reservoir for the ongoing transmission of the disease (Babu and Nutman, 2012; Centers for Disease Control and Prevention, 2013). Acute manifestation is characterized by recurrent attacks of fever associated with the local inflammation of the skin, lymph nodes (lymphadenitis) and lymphatic vessels (lymphangitis) (Babu *et al.*, 2005; Babu and Nutman, 2012). Lymphoedema (tissue swelling) and elephantiasis (skin/tissue thickening) are more commonly observed in brugian filariasis during chronic manifestation. They are developed at the leg, arm, scrotum, vulva and breast of the infected individuals (Partono, 2007). Lymphoedema or swelling of the limbs is a common progression from acute manifestation and is due to the malfunction of lymphatic vessels by the adult worms. When lymphoedema (Figure 1.2 a) becomes severe, it is often referred to as elephantiasis (Figure 1.2 b).

LF treatment under GPELF mainly focuses on mass drug administration with single-dose diethylcarbamazine or ivermectin combined with albendazole. These drugs are to interrupt transmission of LF and morbidity alleviation. The drugs can eliminate the Mf from bloodstream and annual drug regimens need to be administered for at least 5 years continuously (World Health Organization, 2011). For effective control of LF, it lies in early diagnosis, appropriate treatment and follow-up of drug administration.

(a)

(b)

Figure 1.2 Chronic manifestation of lymphatic filariasis. (a) Lymphoedema and (b) Elephantiasis (Babu and Nutman, 2009; Centers for Disease Control and Prevention, 2013).

1.4 Lymphatic filariasis diagnosis

LF mainly affects those who reside in remote areas and with limited health and laboratory facilities. Besides, most of the LF patients have silent manifestations that have no symptoms (asymptomatic) and did not develop clinical symptom. The infections remain unknown in some cases unless tested or showed clinical signs. Thus, on-site diagnosis which is simple, rapid and accessible are essentials for the LF elimination in order for early detection and treatment (Noordin *et al.*, 2007).

Detection methods for brugian filariasis include night blood examination, polymerase chain reaction (PCR) and diagnostic kits such as Brugia Rapid and panLFrapid. Traditional diagnosis of brugian filariasis is by using finger-prick blood sample to detect the presence of Mf in blood under microscope. This method is lack of sensitivity (25%-40% sensitive) because of the low density of Mf, inability to detect cryptic and occult infections (Noordin, 2007). Apart from low sensitivity, the main disadvantage of this diagnosis is the nocturnal periodicity of Mf as they are primarily circulated at night and require night blood collection (Rahmah *et al.*, 2001a). The PCR method has high sensitivity for low levels of Mf but still require the night blood sampling as the parasite must be circulating during the sampling time (Noordin, 2007; TWAS Newsletter, 2011).

The advantages of antigen detection test to detect active filarial infections are the convenience of examination because blood sample can be collected anytime and it is more sensitive compared to the traditional Mf detection (Weil and Ramzy, 2007). For bancroftian filariasis, the BinaxNOW filariasis immunochromatographic test (ICT card) has been widely used for LF elimination programs to detect the *W. bancrofti* antigen (DPDx- Laboratory Identification of Parasitic Diseases of Public Health Concern, 2013). A new rapid antigen test (Alere Filariasis Test Strip) had been

introduced by Weil and colleagues in year 2013. In their studies, both ICT card and Alere Filariasis Test Strip possessed similar high rates of sensitivity and specificity (both >99% of sensitivity and specificity) (Weil *et al.*, 2013). Therefore, for the detection of brugian filariasis, there is a need for antigen detection test which can provide good results (similar with ICT card test and Alere Filariasis Test Strip) and can be easily performed anytime without the need of highly trained technical staff for on-site diagnosis.

1.5 *Bm*R1 antigen from *Brugia malayi*

The *Bm*R1 recombinant antigen expressed by gene *Bm17DIII* from *B. malayi* has been employed in both enzyme-linked immunosorbent assays (ELISA) and Brugia Rapid. Preliminary test based on ELISA demonstrated that *Bm*R1 recombinant antigen possessed high diagnostic value and is patented in Malaysia (2007) and Indonesia (2009) (TWAS Newsletter, 2011). Studies showed that the immunogenicity of *Bm*R1 antigen is clearly different from other filariasis diseases including bancroftian filariasis. Evaluation of the diagnostic tests revealed that the antigen had a sensitivity of 93%-100% for the detection of Mf and it is highly specific (99%-100%) to detect brugian filariasis (Noordin *et al.*, 2004; Rahmah *et al.*, 2001b). Therefore, *Bm*R1 antigen would be a convincing candidate for antigen detection test.

1.6 Application of computational approach

As mentioned in Section 1.5, *Bm*R1 antigen has been shown to be important as a promising biomarker to detect brugian filariasis. However, its structure and function has not been characterized. The identification of the structure and function of this antigen can provide valuable insights to enhance the understanding of this protein.

Besides, knowledge and understanding of antigen-antibody interactions is important in leading to the development of antigen detection test. In order to achieve this, computational approach plays crucial roles.

Advances in experimental and computational approach are important to obtain the protein three-dimensional (3D) structure and information of antigen-antibody interactions. With the increasing number of deposited protein structure from X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy in PDB, the data is highly beneficial to the computational approach that utilized information from the experimentally determined protein structures. Molecular modelling may not be as accurate as experimental method but they often reveal the molecular insight from the predicted structure which is useful to complement the experimental approach and provide fundamental understanding (Petrey and Honig, 2005).

On the other hand, molecular docking can provide a detailed understanding of the protein-protein interaction e.g. specific interaction between antigen and antibody. Protein-protein docking is important especially when the binding interaction of the antibody and antigen is unknown (Simonelli *et al.*, 2010; Kuroda *et al.*, 2012). In order to predict the antigen-antibody complexes, the docking simulation will require the structures of the interacting proteins which are important to yield accurate antigen-antibody complexes (Rajkannan and Malar, 2007; Kuroda *et al.*, 2012). This is due to the antigen-antibody docking algorithms utilize the information that the epitopes of an antigen bind specifically to the paratope of an antibody and these binding sites can be predicted from their structures.

Due to the rapid progress in computational method, the successful rates are increasing over the years with numerous successful examples have been reported. In year 2011, *Chlamydia trachomatis* protein CT296 was determined using both

computational method (I-TASSER) and X-ray crystallography method. The result showed that the structure of CT296 predicted by I-TASSER has overall structural similarity (root mean square , RMSD of 2.72 Å for 101/137 residues) to the high-resolution X-ray crystallography structure (1.8 Å). This clearly showed that computational method is able to predict accurate protein structures despite having no homologs to deposited protein structure (Kemege *et al.*, 2011).

Successes in the structure prediction for gas vesicle protein GvpA have also been reported (Strunk *et al.*, 2011; Ezzeldin *et al.*, 2012). The structure prediction of GvpA protein from haloarchaeon *Haloferax mediterranei* was first carried out by Strunk and colleague via computational *ab initio* method (Rosetta). The predicted structure suggests that GvpA possess two α-helices and two β-strands. The secondary structure elements (α-β-β-α) is similar with the NMR structures obtained for GvpA from cyanobacterium *Anabaena flos-aquae* (Sivertsen *et al.*, 2010). Mutation in α-helix and β-turn affected the ability to form gas vesicle. This *in vivo* data on GvpA mutants support the major structural features from the proposed structures (Strunk *et al.*, 2011). In subsequent year, Ezzeldin and colleagues predicted GvpA protein from *Halobacterium sp. NRC-1* with computational comparative modelling (MODELLER and SCRATCH), threading (by I-TASSER) and *ab initio* modelling (Rosetta) (Strunk *et al.*, 2011; Ezzeldin *et al.*, 2012). All the predicted structures were equilibrated through molecular dynamics (MD) simulation. Average MM-PBSA energy and standard deviation were calculated and ranked. From the comparison of the top ranked predictions structures and an earlier model proposed by Strunk *et al.*, it showed that despite belonging to different organisms, the two sequences possess 93% identity (Strunk *et al.*, 2011; Ezzeldin *et al.*, 2012). Furthermore, the structures with an α-β-β-α secondary structure is in agreement with the previous experiment data

(Sivertsen *et al.*, 2010) and their studies of secondary structure prediction (Ezzeldin *et al.*, 2012). The predicted model supports the hypothesis that homologous sequences synthesized by different organisms should possess similar structures (Ezzeldin *et al.*, 2012).

Computational approach was performed to identify the important residues in the binding site of single chain variable fragment (scFv) against epitope of Gag p55 polyprotein of human immunodeficiency virus type 1 (HIV-1) by Lee and colleagues (Lee *et al.*, 2010). In their study, scFv was modelled from its X-ray structure homologues and the antigen-antibody complexes were generated by computational docking (Lee *et al.*, 2010). Their studies revealed that the key residues were located at the complementarity-determining regions (CDRs) of scFv and they played a crucial role in the binding interaction. In addition, the binding activities from their result showed that the calculated binding free energy had a good correlation with the experimental data ($r^2 = 0.88$). In the subsequent study, Tue-Ngeun and colleagues identified that the key residues were located at the hot spot of the surface between scFv and the antigen through computational alanine scanning. From the analysis, new antibodies were designed by the mutation of these key residues (Tue-Ngeun *et al.*, 2014). Therefore, it showed that computational approach is efficient to generate protein structure and identify the key residue in antigen-antibody interaction as well as to design a new specific scFv with better binding affinity with the given antigen.

1.7 Content of the thesis

This chapter (Chapter One) discusses the statement of the problem, various aspects of lymphatic filariasis and the importance of *Bm*R1 antigen. Applications of computational approaches along with the success cases were also being described here.

Chapter Two provides the details on the prediction of the 3D structure of *Bm*R1 antigen via different approaches. The refinement and dynamics of *Bm*R1 protein were also being described for the identification of potential epitopes.

Chapter Three details the design and generation of scFv specific to *Bm*R1 antigen. The selection criteria for the scFv models were provided and insights into *Bm*R1-scFv interactions were also being discussed.

Chapter Four summarizes the studies in Chapter Two and Three. The possible future directions were also included in this chapter.

1.8 Objectives

The general objectives are:

i) To predict the structure and epitopes of *Bm*R1 antigen, and

ii) To design scFv specific toward *Bm*R1 antigen

CHAPTER TWO

THE STRUCTURE AND EPITOPES PREDICTION OF *BM*R1 PROTEIN FROM *BRUGIA MALAYI*

2.1 Protein structure prediction

Specific function and mechanism of a protein can be derived from the 3D structure of a protein. The most accurate way to determine a high resolution protein structure is through experimental methods such as X-ray crystallography or NMR spectroscopy (Wu and Zhang, 2009; Nguyen and Madhusudhan, 2011). Even though experimentally determined structures will normally possess high resolution structure information about a protein, it is a time consuming process without guaranteed success (Schwede *et al.*, 2003). There are also limitations in experimental method such as cost in handling the experiment and some proteins cannot be crystallized easily (Mizianty and Kurgan, 2011). When experimentally determined structures are unavailable, the predictive structures may serve as starting points to study the possible function of a protein. They can often reveal important information of a protein. These information, including the protein's folding, evolution, function as well as the protein-protein interactions that could generate testable hypotheses which is useful to complement the experimental information (Petrey and Honig, 2005; Wooley and Ye, 2007). Therefore, it leads to the rapid growth and importance of computational protein structure prediction.

Protein structure prediction is a method of translating the protein sequence into 3D structure by employing computer algorithms. Protein structure prediction is still one of the most challenging problems in computational structural biology nowadays despite the significant progress in recent years showed by critical

assessment of protein structure prediction (CASP) experiments (Zhang, 2008). As of December 2014, the Protein Data Bank has over 100,000 deposited protein structures (www.rcsb.org) (Berman *et al.*, 2000). Although the number of experimentally solved protein structures is increasing at an accelerated rate, at the same time, numbers of known protein sequences from genome sequencing projects are increasing at a breathtaking pace. To bridge the protein sequence-structure gap, computational protein 3D structure predictions from its amino acid sequence provide potential solution (Webb and Sali, 2014).

Computational methods for predicting protein 3D structures can be generally divided into three categories: comparative modelling, threading and *ab initio*. It can also be categorized into template-based (TBM) and template-free (FM) modelling (Källberg *et al.*, 2012; Maurice, 2014). Comparative modelling and threading method are categorized into TBM as they depend on the availability of a template from solved protein structure (Fiser, 2010). FM (also known as *ab initio* or *de novo* method) is potentially able to predict protein structures without any template (Moult *et al.*, 2011; Maurice, 2014). The general overview of protein structure prediction is illustrated in Figure 2.1.

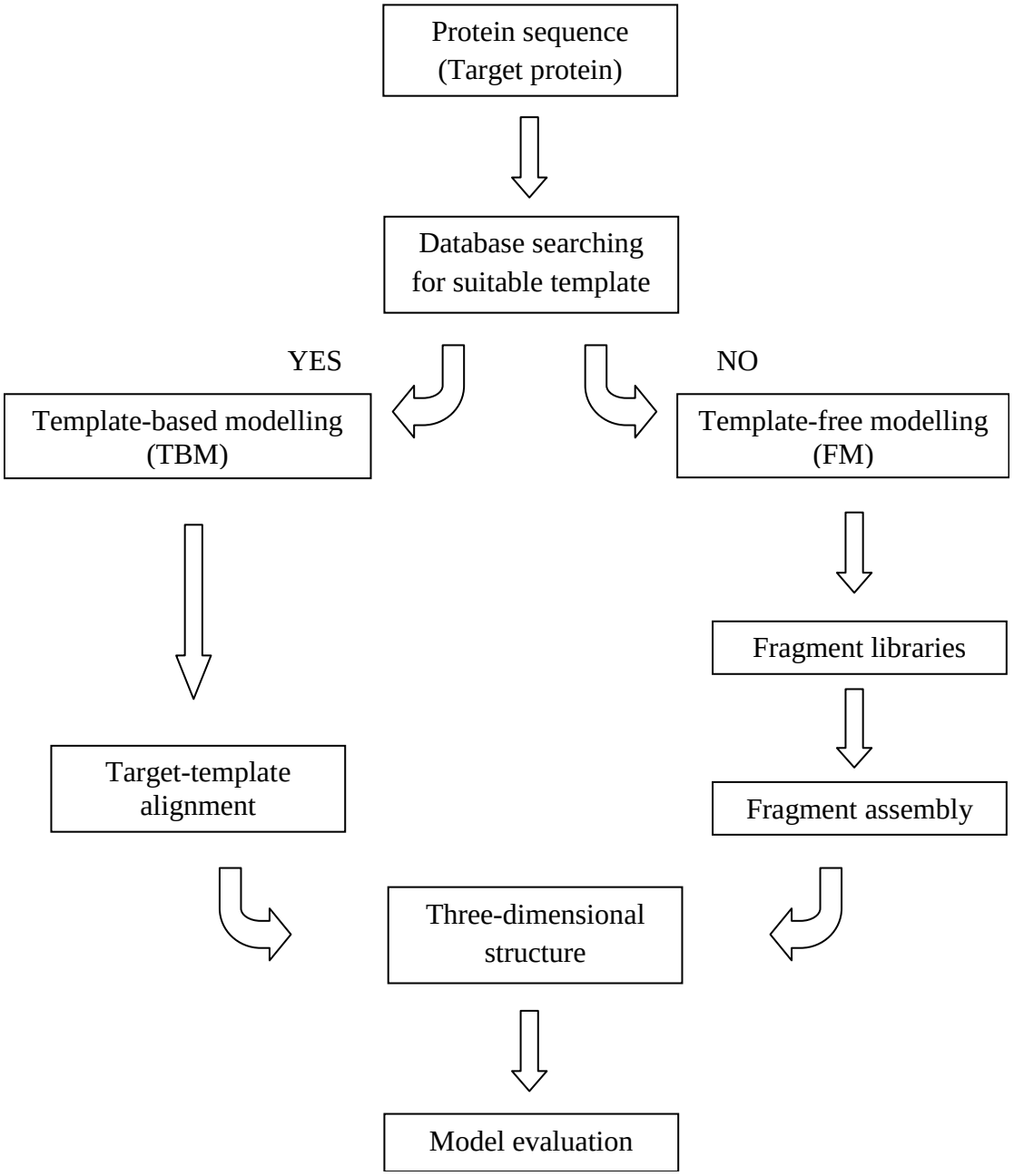

Figure 2.1 General overview of protein structure prediction from its primary sequence to three-dimensional (3D) structure through template-based modelling (TBM) or template-free modelling (FM).

2.1.1 Template-based modelling (TBM)

To date, comparative modelling is the most successful and accurate computational method to produce a reliable protein 3D structure from its amino acid sequence. The accuracy of protein structure increases with the accumulation of experimentally determined structures in the database (Yan *et al.*, 2013). The idea behind comparative modelling is evolutionarily related proteins with at least 30% sequence identity will usually share similar structures as the structures and functions of proteins are often conserved (Errami *et al.*, 2003; Schwede *et al.*, 2003; Mihăşan, 2010; Choong *et al.*, 2011). Comparative modelling predicts the 3D structure of a target protein based on its alignment to one or multiple templates from protein of known structures deposited in a database (Martí-Renom *et al.*, 2000; Eswar *et al.*, 2008). Comparative modelling consists of four main steps: fold assignment that identifies similarity between target sequence and templates, template-target alignment, model building and model evaluation (Martí-Renom *et al.*, 2000; Eswar *et al.*, 2003; Eswar *et al.*, 2008). The identification of templates can be facilitated by numerous structure databases available online such as Protein Data Bank (PDB) (Berman *et al.*, 2000), CATH (Knudsen and Wiuf, 2010) and Structural Classification of Proteins (SCOP) (Conte *et al.*, 2000). The target sequence can be compared with the solved protein sequences using pair-wise sequence-sequence comparison in order to obtain related protein structures by sequence similarity alignment and search tools e.g BLAST (Altschul *et al.*, 1990), FASTA (Pearson and Lipman, 1988), GPU-BLAST (Vouzis and Sahinidis, 2011), mpiBLAST (Darling *et al.*, 2003) and G-BLAST (Zhao and Chu, 2014). BLAST from NCBI is the most widely used bioinformatics tool to search for similar sequences (Vouzis and Sahinidis, 2011; Zhao and Chu, 2014). Other factors that need to be taken into consideration during the selection of suitable

template(s) including the protein family and quality of the experimentally determined structures (Marti-Renom *et al.*, 2002).

Threading, also known as fold recognition, is another TBM technique used to identify structural similarity of evolutionally remotely related protein templates (Roy *et al.*, 2010; Roy and Zhang, 2012). The concept for threading is similar to comparative modelling but comparative modelling only considers sequence similarity between target protein and template while protein threading considers the structural information in the template (Xu *et al.*, 2008). The critical step of threading is to identify correct template proteins with similar folds to the target protein and make correct alignment (Wu and Zhang, 2008). Threading methods compare a target sequence against one or more protein 3D structures to detect and obtain best compatibility of sequence-structure template pair (Xu *et al.*, 2008; Wu and Zhang, 2009). The best fit of target sequence with the fold template based on the generated alignments is identified and each template is calculated according to different scoring function. Commonly used alignment scores to identify precise target-template alignments include sequence profile-profile alignments (PPA), sequence-structural profile alignments, secondary structure match, hidden-Markov models (HMM) and residue-residue contacts (Wu and Zhang, 2009). The alignment algorithms are able to search for remotely homologous sequences in the databases. Therefore, even if the sequence similarity is relatively low (<30%), threading method can be used to obtain the similar fold or similar structural motif for the target sequence. PPA, which can be used to detect weak similarities between protein families, is most often-used and popular threading approach which performed better than other algorithm as it can detect weak similarities between protein families (Yona and Levitt, 2002; Wooley and Ye, 2007; Yan *et al.*, 2013).

2.1.2 Template-free modelling (FM)

When there is no homologous structure in PDB or the relationship is so distant until it cannot be detected, an *ab initio* approach is the alternative way to generate the structure from scratch (Wu and Zhang, 2009). This type of prediction is termed template-free modelling (FM) (also termed as *ab initio* or *de novo* modelling) as it originally referred to methods that based on the first principle laws of physics and chemistry. The idea is based on Anfinsen's thermodynamic hypothesis. According to the hypothesis, the protein structure was determined solely by its amino acid sequence (Anfinsen, 1973; Wooley and Ye, 2007; Hoque *et al.*, 2009). The prerequisite of these modelling methods is that the native structure has the global minimum free energy among all available conformations (Roy and Zhang, 2012). Therefore, efficient and reliable algorithm is important to limit the conformational space in order to minimize the energy function so that the protein tends to be in its native state (Bonneau and Baker, 2001; Ishida *et al.*, 2003).

There are a variety of methods developed for *ab initio* protein structure generation. The leading approach is the fragment-based assembly method, an idea of Bowie and Eisenberg (Bowie and Eisenberg, 1994). Based on this idea, Rosetta (Bonneau *et al.*, 2001) is developed and is a leading method for FM approach as Rosetta is able to produce accurate models that are nearer to native structures (Simons *et al.*, 1997; Bonneau and Baker, 2001; Simoncini and Zhang, 2013). The idea of this approach is that the smaller fragments are restricted to the local structures by the closely related sequence in protein structure database (Simons *et al.*, 1997; Ishida *et al.*, 2003). The lengths of the fragments vary with programs and the fragment libraries are fragments from high resolution solved protein structures. In Rosetta, fragment libraries of three- and nine-residue are exploited (Bonneau *et al.*, 2001). Generation of

fragments is important in Rosetta after the completion of secondary structure prediction and it can be performed through Robetta server (Chivian *et al.*, 2003; Chivian *et al.*, 2005). This program iterates over three- and nine-residue of the sequence and looks for similar sequences from the fragment libraries to guide the search of conformational space in predicting protein structures (Kim *et al.*, 2004). In QUARK, the models are assembled from small continuous fragments ranged from 1-20 residues excised from unrelated proteins (Zhang, 2013). Both Rosetta and QUARK showed the importance of assembling structural models using small fragments by their significant performance in CASP9 (Kinch *et al.*, 2011). In CASP10, QUARK successfully predicted model with largest size range in FM modelling (>150 residues) (Xu and Zhang, 2013a).

2.1.3 Current trend in protein structure prediction

Recently, the borders between the types of protein structure prediction methods have overlapped. Recent CASP experiments demonstrated that composite approaches can achieve additional advantages in structure prediction. Since no single approach can perform better than others for all protein prediction, the emergence of recent trend is combination/hybrid of different protein structure prediction approaches (Roy and Zhang, 2012; Zhang, 2013).

I-TASSER (Iterative Threading ASSEmbly Refinement) is one of the notable successful composite approaches in the CASP experiments (Roy *et al.*, 2010). I-TASSER method is based on the secondary structure enhanced profile-profile threading alignment extended from TASSER algorithm for iterative structure assembly and refinement of protein molecules (Zhang, 2008; Zhang and Skolnick, 2013). I-TASSER retrieves structural template from PDB library through a meta-threading server, termed LOMETS. By year 2010, the online I-TASSER server has

generated more than 30,000 full-length structure and function predictions for more than 6,000 registered users (Roy *et al.*, 2010). I-TASSER can consistently predict correct folds and sometimes high-resolution for small single-domain protein (<120 residues) with a lower computational time (5 CPU hours for I-TASSER but 150 CPU days per target for Rosetta) (Wu *et al.*, 2007). In CASP7, CASP8, CASP9 and CASP10, I-TASSER was ranked as the best server for protein structure prediction (Nahar *et al.*, 2014).

In year 2013, *Bhageerath*-H Strgen, another homology/*ab initio* hybrid algorithm was developed (Dhingra and Jayaram, 2013). The method was tested in CASP9 experiments and showed 93% of the targets were in the pool of decoys. The results showed that *Bhageerath*-H Strgen is capable to search the protein fold for near-native conformation. Strategies in *Bhageerath*-H Strgen include secondary structure prediction, database search for sequence based on the input amino acid sequence, fold recognition, template-target alignment and template-based modelling by MODELLER (Webb and Sali, 2014). The missing residues with no fragments are modelled using *Bhageerath ab initio* modelling. In their study, the results showed that *Bhageerath*-H Strgen performs better than Rosetta and I-TASSER.

2.2 Specific aim and objective

The specific objectives are:

i) to predict and evaluate the three-dimensional structure of *Bm*R1 antigen from *B. malayi* through different *in silico* approaches, and

ii) to predict the potential epitope(s) of the *Bm*R1 antigen.

2.3 Methodology

2.3.1 Overview of the method

The overview of this study is illustrated in Figure 2.2. Sequence alignment and analysis against the available online database and tools were carried out in order to obtain the suitable templates, to identify the domain of the protein and to annotate the protein function. Subsequently, the protein structure was predicted and the quality of the structure was evaluated. Molecular dynamics (MD) simulation was performed in order to further minimize and refine the best predicted structure. Finally, the linear and conformational epitopes as well as binding site of *Bm*R1 antigen were identified from the structure.

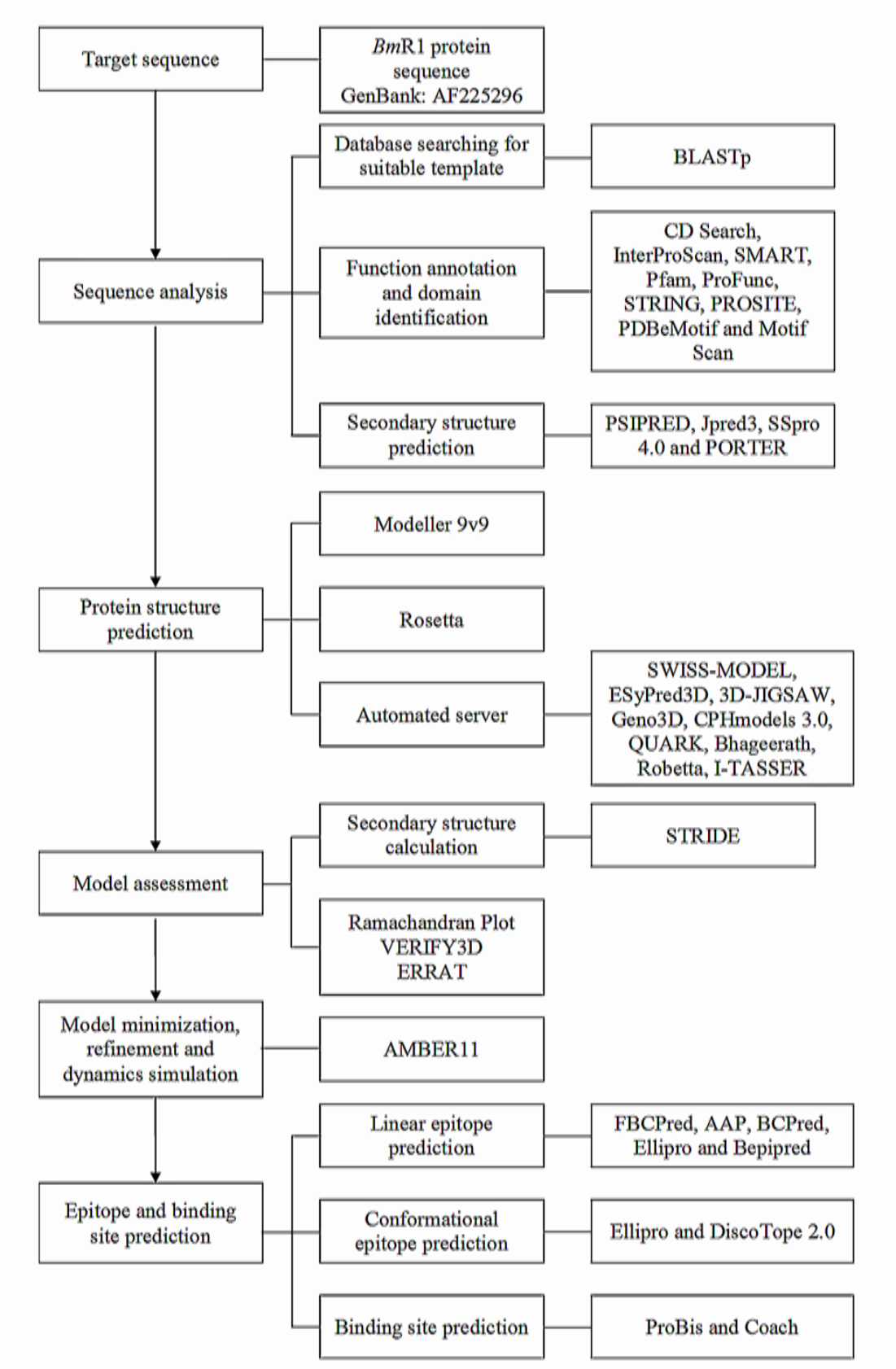

Figure 2.2 The overall methodology for structure and epitope prediction of *Bm*R1 protein from *B. malayi*.

2.3.2 Sequence analysis

The amino acid sequence of *Bm*R1 protein was retrieved from GenBank with the accession number: AF225296. ProtParam from Expert Protein Analysis System (ExPASy) Proteomics Server (Gasteiger *et al.*, 2003) was implemented to calculate the protein molecular weight. The 206 residues of *Bm*R1 protein was subjected to BLAST search against non-redundant (nr) protein sequences to identify similar protein family. BLASTp search with default parameters against PDB was performed to identify suitable templates (sequence identity >30%). To assist the study on tertiary structure prediction, secondary structure prediction on *Bm*R1 protein was performed using Jpred3 (Cole *et al.*, 2008), PORTER (Pollastri and McLysaght, 2005), PSIPRED (Buchan *et al.*, 2010) and SSpro 4.0 (Cheng *et al.*, 2005). Function annotation and identification of the conserved domain were carried out using Conserved Domain Search Service (CD Search) (Marchler-Bauer and Bryant, 2004; Marchler-Bauer *et al.*, 2011), InterProScan (Quevillon *et al.*, 2005), Proteins Families database (Pfam) (Finn *et al.*, 2010) and SMART (Letunic *et al.*, 2009). Protein function analysis were performed by Motif Scan from ExPASy (Gasteiger *et al.*, 2003), PDBeMotif (Golovin and Henrick, 2008), ProFunc (Laskowski *et al.*, 2005), PROSITE (Sigrist *et al.*, 2013) and STRING (Mering *et al.*, 2005).

2.3.3 Structural prediction and evaluation

2.3.3.1 Comparative modelling approach

Multiple sequence alignment was performed prior to the model construction. A total of 250 initial models were generated from multiple templates (PDB id: 2G3Y, 2IE8 (Lee *et al.*, 2006), 3QOE (Kim *et al.*, 2011), 1V32, 2E87, 2R3V (Fu *et al.*, 2008) and 3I4Q) to improve the accuracy of structures, followed by secondary structure

restraints and loop refinement using MODELLER 9v9 (Appendix A) (Martí-Renom *et al.*, 2000). After subsequent steps of secondary structure restraints and loop refinement, the best model was selected based on Discrete Optimized Protein Energy (DOPE) score, MODELLER objective function (molpdf) and the quality assessment of the model. The molpdf is used to measure the satisfaction of the model input spatial restraints (Eswar *et al.*, 2008). DOPE is an atomic distance-dependent statistical potential calculated from a set of known native protein crystallographic structures and can be used to identify native-like conformation of protein (Shen and Sali, 2006). Lower values of the molpdf and DOPE indicate more accurate model (Eswar *et al.*, 2008).

2.3.3.2 *Ab initio* approach

Protein structure prediction by *ab initio* approach was carried out through Rosetta (Bonneau *et al.*, 2001). AbinitioRelax (AbRelax), which is the combination of *ab initio* folding and refinement by Rosetta full-atom force field (Relax), was used to create five hundred initial structures (Appendix B). The Relax application makes small backbone and side chain torsion movements in order to find a minimum local conformation space and Rosetta energy function (Bonneau *et al.*, 2001; Combs *et al.*, 2013). The input files consisted of the FASTA file of protein sequence, fragments library files and the secondary structure prediction files. PSIPRED secondary structures file and fragment libraries of three- and nine- residue were generated by Robetta fragment server (Chivian *et al.*, 2003). AbRelax combined the generated three-mer and nine-mer fragment through fragment assembly method (Kim *et al.*, 2004) and secondary structure prediction from PSIPRED was used as guideline.