IN SILICO OPTIMISATION OF DOMAIN ANTIBODIES AGAINST HSP16.3 FROM *MYCOBACTERIUM TUBERCULOSIS*

SOONG JIA XIN

UNIVERSITI SAINS MALAYSIA
2018

IN SILICO OPTIMISATION OF DOMAIN ANTIBODIES AGAINST HSP16.3 FROM *MYCOBACTERIUM TUBERCULOSIS*

by

SOONG JIA XIN

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

March 2018

ACKNOWLEDGEMENT

First and foremost, I wish to convey my profuse thanks to my beloved supervisor, Assoc. Prof. Dr. Choong Yee Siew. Her relentless support and constructive comments are what pulled me through the difficult times during my Masters study. I could not have been luckier to learn from her. I would like to thank Assoc. Prof. Dr. Lim Theam Soon, my co-supervisor, for his academic support. My appreciation also extends to Dr. Tye Gee Jun.

My gratitude also goes to the Ministry of Higher Education for MyMaster scholarship under MyBrain15 programme. The government's continuing support for higher education has helped tremendously in relieving students' financial burden.

To my labmates, thanks for the help and support. Life in the lab is far from monotony with you guys around. I will always cherish those moments.

Last but not least, my most heartfelt thanks to my parents. Thank you for your understanding and encouragement. Thank you for your confidence in me. I cannot imagine being the person I am today without such amazing parents.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D	Three-dimensional
ACD	α-crystallin domain
ATP	Adenosine triphosphate
bisANS	4,4′-Dianilino-1,1′-binaphthyl-5,5′-disulfonic acid
BLAST	Basic Local Alignment Search Tool
CAS	Computational alanine scanning
CAPRI	Critical Assessment of Prediction of Interactions
CB	Beta-carbon
CDR	Complementarity determining region
CG	Conjugate gradient
CHARMM	Chemistry at Harvard Macromolecular Mechanics
dAb	Domain antibody
DOPE	Discrete Optimised Protein Energy
EM	Electron microscopy
FFT	Fast Fourier Transform
GA	Genetic algorithm
GAFF	General AMBER force field
G-factor	Geometry factors
HSP16.3	Heat shock protein 16.3
ICOSA	Icosahedra
LCPO	Linear Combination of Pairwise Overlaps
LGA	Lamarckian genetic algorithm
LTBI	Latent tuberculosis infection
Mtb	*Mycobacterium tuberculosis*

MD	Molecular dynamics
MM-GBSA	Molecular mechanics generalised Born surface area
MM-PBSA	Molecular mechanics Poisson-Boltzmann surface area
NMR	Nuclear magnetic resonance
NPT	Constant temperature, constant pressure ensemble
NVT	Constant temperature, constant volume ensemble
PDB	Protein Data Bank
PME	Particle Mesh Ewald
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SASA	Solvent-accessible surface area
sdAb	Single domain antibody
sHSP	Small heat shock protein
TB	Tuberculosis
TBM	Template-based modelling
UniProtKB	UniProt Knowledgebase
vdW	Van der Waals
VGP	Variable gap penalty
VH	Variable heavy
VTFM	Variable target function method
WHO	World Health Organisation
WT	Wild type

LIST OF SYMBOLS

π	Pi
ε_{in}	Solute dielectric constant
Å	Angstrom
Cα	Alpha carbon
fs	Femtosecond
K	Kelvin
kDa	Kilodalton
ns	Nanosecond
ps	Picosecond

PENGOPTIMUMAN *IN SILICO* ANTIBODI DOMAIN TERHADAP HSP16.3 DARIPADA *MYCOBACTERIUM TUBERCULOSIS*

ABSTRAK

Protein renjatan haba 16.3 (HSP16.3) daripada *Mycobacterium tuberculosis* (Mtb) adalah kritikal bagi kewujudannya semasa jangkitan pendam pada manusia, justeru menjadi sasaran menarik untuk strategi diagnostik dan terapeutik. Model struktur ramalan HSP16.3 didokkan terhadap prob hidrofobik HSP, iaitu *4,4'-dianilino-1,1'-binaphthyl-5,5'-disulfonic acid* (bisANS) dan pada model perbandingan HSP16.3-khusus antibodi domain tunggal (sdAbs), klon E3 dan F1. Interaksi pengikatan tersebut dijelaskan lebih lanjut dengan pengiraan tenaga bebas. Interaksi bukan-kutub telah dikenalpasti sebagai kuasa utama untuk peyatuan antigen-antibodi. Dengan menggunakan penguraian tenaga bebas setiap residu dan pengiraan pengimbasan alanina, residu-residu kelompok hangat pada E3 (Y391) dan F1 (M394, Y396, R397 dan M398) telah dikenalpasti. Residu-residu tersebut ditaklukkan kepada mutagenesis *in silico* berdasarkan cadangan oleh pelayan web mCSM-AB. Kesan mutasi pada kompleks HSP16.3-dAb dianalisis dengan simulasi dinamik molekul, pengiraan tenaga bebas dan penguraian tenaga bebas setiap residu. Kompleks HSP16.3-E3$_{Y391W}$ diramal mempamerkan peningkatan sebanyak 69% dalam tenaga bebas pengikatan berbanding dengan E3 jenis liar. Sebaliknya, tahap peningkatan tertinggi bagi F1 jenis liar adalah HSP16.3-F1$_{R397N}$ (44%), diikuti oleh HSP16.3-F1$_{M398Y}$ (33%), HSP16.3-F1$_{M394E}$ (29%) dan akhir sekali HSP16.3-F1$_{M398W}$ (6%). Kesimpulannya, dAbs (E3 dan F1) berjaya dioptimumkan terhadap HSP16.3 pada tahap *in silico*. Penemuan ini boleh digunakan sebagai garis panduan bagi reka bentuk dAbs yang lebih tinggi affiniti terhadap HSP16.3 pada tahap *in vitro* pada masa depan.

IN SILICO OPTIMISATION OF DOMAIN ANTIBODIES AGAINST HSP16.3 FROM *MYCOBACTERIUM TUBERCULOSIS*

ABSTRACT

Heat shock protein 16.3 (HSP16.3) from *Mycobacterium tuberculosis* (Mtb) is critical for its survival during latent infection in human, thus making it an attractive target for developing diagnostic and therapeutic strategies. The predicted structure of HSP16.3 was docked against a known HSP hydrophobic probe, namely 4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bisANS) and to the comparative models of HSP16.3 specific single domain antibodies (sdAbs), clone E3 and F1. The binding interactions were further elucidated by free energy calculations. The non-polar interactions were identified as the main force for antigen-antibody association. By using per-residue free energy decomposition and computational alanine scanning, the hot spot residues in E3 (Y391) and F1 (M394, Y396, R397 and M398) had been identified. These residues were subjected to *in silico* mutagenesis based on suggestions by mCSM-AB webserver. The mutational effects on HSP16.3-dAb complex were analysed using molecular dynamics simulation, free energy calculations and per-residue free energy decomposition. The HSP16.3-E3$_{Y391W}$ complex was predicted to exhibit up to 69% improvement in its binding free energy over the E3 wild type. On the other hand, the highest improvement in F1 wild type was HSP16.3-F1$_{R397N}$ (44%), followed by HSP16.3-F1$_{M398Y}$ (33%), HSP16.3-F1$_{M394E}$ (29%) and lastly HSP16.3-F1$_{M398W}$ (6%). Thus, it can be concluded that the dAbs (E3 and F1) have been successfully optimised against HSP16.3 at *in silico* level. These findings could serve as guidelines for design of higher affinity dAbs against HSP16.3 at *in vitro* level in the future.

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Mycobacterium tuberculosis (Mtb) survives long-term dormancy due to cell wall thickening and upregulation of heat shock protein 16.3 (HSP16.3) which stabilises its cell structures (Cunningham & Spreadbury, 1998). The HSP is predominantly found in latent tuberculosis infection (LTBI) individuals before it develops or reactivates into active stage. By treating LTBI, it minimises the risk of progression to active tuberculosis (TB). Therefore, detection of LTBI in time plays a crucial role in global effort to combat TB epidemic. In this study, the domain antibodies (dAbs) provided by collaborator had exhibited good binding affinity towards HSP16.3, major antigen detected in LTBI (Bahara et al., 2016). Thus, these dAbs could be promising candidates as diagnostic biomarkers. To achieve good diagnostic accuracy, binding affinity of dAbs against HSP16.3 can be further optimised. However, antibody optimisation via laboratory approach is expensive and time-consuming. Therefore, *in silico* approach offers an alternative method to enhance the binding affinity of dAbs against HSP16.3.

1.2 Background of Study

TB prevails as one of the most ominous global health threat. During the 18th and 19th centuries, TB emerged as an escalating epidemic of communicable disease (Daniel, 2006). The lack of knowledge about disease containment has left the world appalled at confirmed diagnostic evidence of the deadly TB infection in those days.

Despite advances in current medical technology, this age-old disease has ailed a staggering number of 10.4 million people and claimed at least 1.4 million lives in year 2015, according to World Health Organisation report (WHO, 2016). Thus, one should not make light of the threats posed by TB.

WHO estimated that 5-15% of people infected with Mtb stand a risk of developing TB disease later in life (WHO, 2016). In line with End TB Strategy, LTBI treatment coverage is one of the indicators used by WHO to monitor implementation of the strategy. Examples of high-risk groups for LTBI include children aged under five years with close contact with culture-positive pulmonary TB patients, HIV-positive patients and citizens in high TB burden countries.

As a means to prevent and contain the infectious disease, it is a priority to characterise mycobacterial antigens for better understanding of the molecular mechanism of pathogen to stay in dormant stage and its pathogenesis. A 16 kDa alpha crystalline-like small heat shock protein (sHSP) has been identified to be latency associated antigen predominantly expressed by Mtb (Yuan et al., 1996). Several studies have directed HSP16.3 as a potential diagnostic marker (Davidow et al., 2005; Kashyap et al., 2011; Rabahi et al., 2007; Shekhawat et al., 2016; Silva et al., 2014; Zhang et al., 2015), and shared principal findings whereby the 16 kDa antigen has elevated levels in latent TB subjects.

Antibodies targeting major antigen can contribute in the advancement of research, diagnosis and possible therapeutics of TB. Instead of using conventional antibodies, single-domain antibodies (sdAbs) were selected in order to study their interactions with HSP16.3 in this research. Other than being smaller in size, sdAbs possess high physical-chemical stability, good water solubility and better penetration in reaching target antigens (Eyer & Hruska, 2012). The dAbs have been reported to

have reasonable expression level, good solubility and capable of binding to HSP16.3 at denaturing environment (Bahara et al., 2016). Therefore, further optimisations of these dAbs against HSP16.3 in this study could be useful for development in diagnostics and therapeutics.

On the other hand, recent progress on *in silico* approach contributed to the exponential growth of computational modelling which in turn enables antibodies to be re-designed or optimised to portray better affinity or other favourable changes (Kuroda et al., 2012). This is due to *in silico* simulation can enable more directional mutations at complementarity determining regions (CDRs) to be studied at a relatively lower running cost. For example, an early study employed computational modelling to select favourable residues for random mutagenesis (Barderas et al., 2008). As a result, they obtained novel antibodies with 454-fold more enhanced binding affinities over the wild type (WT).

1.3 Scope of Study

The study started with the modelling of HSP16.3. The reliability of HSP16.3 model was evaluated by analysing its interactions with 4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bisANS), a known hydrophobic ligand for sHSPs. Next, dAbs were modelled and docked to predicted epitopes on HSP16.3. Interactions established in docked complexes were investigated by molecular dynamics (MD) simulation and binding free energy calculations. Computational alanine scanning and per-residue binding free energy calculations were then performed on the dAbs. *In silico* mutagenesis was performed on the identified hot spot residues in dAbs. Successful dAb mutants with enhanced binding affinity against HSP16.3 were identified by improvement in their binding free energy compared to WT.

1.4 General Objective

The general objective in this study is to optimise the HSP16.3-specific dAbs for possible binding affinity improvement with HSP16.3 at computational level.

1.5 Thesis Outline

Chapter one enlightens the current status of TB by providing background of study and addressing problems to be solved. It also included scope, general objective and outlined the thesis structure.

Chapter two covers the first stage of study which involved predicting the structure of HSP16.3 and exploring its interactions with bisANS. Specific objectives and introductory concepts of HSP16.3 structure and relevant computational approaches in structural biology were discussed too. Protocols used and research findings can be referred to methodology, results and discussion sections respectively.

Chapter three highlights on the second stage of study which focused on modelling of dAbs, docking of dAbs to HSP16.3, binding free energy calculations of docked complexes and assessment of point mutational effects in dAbs on their binding affinities against HSP16.3. An overview on antibody design and the successes of computer-aided antibody design were provided in introduction. Specific objectives can also be found. Detailed protocol and research outcomes can be found in methodology, results and discussion sections respectively.

Chapter four concludes the overall research findings, acknowledges limitations in this study and suggests directions for future research.

CHAPTER 2

PREDICTING THE STRUCTURE OF HSP16.3

2.1 Introduction

2.1.1 Structural and Functional Studies of HSP16.3

In times of cellular stress, HSPs (as the stress proteins) are often upregulated as part of cellular defensive response. The vital role of HSPs in stress tolerance makes them highly conserved and ubiquitous across species. By distinguishing HSPs on the basis of molecular mass, it results in six major classes i.e. HSP100, HSP90, HSP70, HSP60, HSP40 and sHSPs (Bakthisaran et al., 2015). The present study focused on a sHSP, namely HSP16.3 (Verbon et al., 1992), a membrane protein which facilitated the persistence of Mtb (Cunningham & Spreadbury, 1998).

In general, sHSPs have molecular mass ranging from 12-43 kDa, that agglomerates into diversified oligomers consisting of 4-42 subunits (Schumann, 2006). The domain architecture of sHSP family constitutes of a highly conserved α-crystallin domain (ACD) that is placed between variable N-terminal and C-terminal regions, as discussed in a review (Hilton et al., 2013). The C-terminal region can be further divided into C-terminal tail, IXI motif and C-terminal extension. Acting as ATP-independent chaperones, sHSPs bind denatured proteins to prevent unwanted aggregation (Jakob et al., 1993). It is postulated that these bound proteins are subjected to refolding, being degraded to smaller peptides or spontaneous release.

The identified 16 kDa HSP16.3 is consistent with the known size range of sHSPs (Lee et al., 1992). The chaperone-like activity of HSP16.3 is evidenced from an early study whereby it had successfully inhibited thermal aggregation of citrate

synthase at elevated temperature (Chang et al., 1996). In addition, novel function of HSP16.3 has been unveiled in a recent discovery that reported its ability to prevent thermal inactivation of enzymes (Panda et al., 2017).

The critical role of the N-terminal and C-terminal region of HSP16.3 has been extensively highlighted in an early study (Fu et al., 2005). A few observations had been noticed. The absence of first 35 N-terminal residues was found to dissociate the oligomeric assembly, abolished its substrate binding capability and chaperoning activity. Apart from that, removal of C-terminal extension (nine residues) could lead to oligomeric dissociation and enhanced its chaperoning activity. Moreover, varying C-terminal truncated forms of HSP16.3 exhibited different strengths of chaperone function (Panda et al., 2017). These results thus suggested the importance of C-terminal extension in chaperone function, oligomerisation and its dynamics.

Due to the diagnostic potential of HSP16.3 (Davidow et al., 2005; Kashyap et al., 2011; Rabahi et al., 2007; Shekhawat et al., 2016; Silva et al., 2014; Zhang et al., 2015), the research community is intrigued to determine its protein structure. Protein structure determination often plays a pivotal role as it holds the key to understanding the protein functions better. By unearthing new pieces of structural information, it would be helpful to gain further insights into the immunodominant antigen of Mtb.

Nevertheless, elucidating the macromolecular assembly of HSP16.3 is an exceptionally challenging task. Previously, the oligomeric form of HSP16.3 was proposed as a nonamer that formed by trimers (Abulimiti et al., 2003; Chang et al., 1996; Gu et al., 2002). The early findings were then contradicted by later research. There is conclusive evidence that showed HSP16.3 as a dodecamer built from six homodimers arranged in a tetrahedral assembly (Kennaway et al., 2005). The atomic coordinates of a wheat HSP16.9 dimer was fitted to the electron microscopy (EM)

density map to construct the HSP16.3 dodecameric assembly. Although it offered new perspectives on molecular architecture of HSP16.3, there is missing structural information on N-terminal region and C-terminal tail. Due to observed sequence divergence between HSP16.3 and template in the above-mentioned regions, they were not included in the 3D reconstruction study. Considering the significance of the N-terminal region, *in silico* approaches were thus employed to predict the full structure of HSP16.3 in this study. The following sections introduce and briefly discuss about relevant computational approaches in predicting and refining theoretical protein structure.

2.1.2 Employing *in Silico* Approaches in Structural Biology

As the key determinant of its 3D structure, protein sequence provides clues about its protein functionality. By obtaining structural information of the target protein, structural biologists might be able to infer its protein functions. Therefore, numerous studies are dedicated towards determining protein structures to annotate their unknown functions (Kennaway et al., 2005; van Montfort et al., 2001).

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are commonly employed in structural studies of proteins. However, conventional experimental methods often face inevitable hurdles such as lengthy time, laborious process and high experimental cost. Furthermore, novel protein sequences are being rapidly discovered at a pace that far exceeds the amount of known structures. To address the knowledge gap, *in silico* approaches offer an attractive alternative for protein structure prediction, thus serve as an essential complement to existing experimental methods.

2.1.3 Protein Structure Prediction Using Comparative Modelling

Computational protein structure prediction is classified into two major approaches, namely template-based modelling (TBM) and *ab initio* modelling. The TBM approach is based on the principle that higher possibility of predicting the model folding correctly when its sequence identity is above 30% (Sánchez and Sali, 1998). Therefore, it is possible to predict structure of a given target sequence by aligning it to an experimentally solved homologous protein structure. This is commonly known as comparative modelling. However, it is later discovered that distantly related protein sequences also share similar structural patterns (Bowie et al., 1991), thus contributed to the concept of threading. In threading, target sequence is aligned to known protein structures to predict its tertiary folds by matching residue environments. On the other hand, *ab initio* modelling predicts native-like conformations for target sequence in the absence of known structure homologues (Lee et al., 2009). The most thermodynamically stable model is selected by identifying the conformation that is close to the global free-energy minimum from the pool of structural decoys.

Since comparative modelling is a well-established approach, various comparative protein modelling tools have been developed and made available. Some examples of the popular modelling tools include SWISS-MODEL (Arnold et al., 2006) and MODELLER (Šali & Blundell, 1993). Regardless of the choice of tools, conventional comparative modelling protocol usually consists of a few sequential steps. It begins searching for suitable templates, aligning target sequence to selected templates, building the model and lastly evaluating the reliability of the built model. These sequential steps will be further elaborated using MODELLER as the modelling tool.

As a preliminary step, template selection is non-trivial in comparative modelling. There are several things to consider before deciding on the most suitable template for modelling. Other than prioritising sequence similarity, quality of the potential template is equally important (Fiser, 2010). Structural divergence and poor template quality could usually affect the modelling accuracy. A minimum of 30% template sequence identity and a good template-sequence alignment, are likely to build a reliable model that overlaps the actual structure about 75-90% (Forrest et al., 2006; Sánchez & Sali, 1998). In the case of multiple potential templates that share high sequence similarity with target sequence, the template determined at better resolution shall be selected (Sánchez & Šali, 2000). To further improve modelling accuracy, multiple templates are commonly used as they can provide adequate structural information for matching protein regions (Webb & Sali, 2014).

After identifying ideal template, the next step is sequence alignment. Likewise, caution shall be exercised while performing an alignment. Misalignment can introduce unrecoverable errors during modelling. Thus, numerous alignment algorithms have been devised to optimise the alignment accuracy and improve the protein structure prediction. For instance, variable gap penalty (VGP) algorithm in MODELLER (Madhusudhan et al., 2006). The algorithm is based on a global dynamic programming algorithm, alternatively known as the Needleman-Wunsch algorithm after the names of developers (Needleman & Wunsch, 1970). Basically, the Needleman-Wunsch algorithm describes how to find the best global pairwise alignment that holds the optimal alignment score. The use of a score matrix enables scores to be assigned for observed amino acid substitutions during alignment. Next, a traceback matrix is employed to keep track of the maximum score value and to deduce the best alignment along the traceback path. The VGP algorithm distinguishes itself from conventional

alignment approaches by including structural information from selected template (Madhusudhan et al., 2006). To improve alignment accuracy, gap penalties are given when gaps are placed within secondary structure elements, straight backbone segments that are solvent inaccessible and residues that are not within close spatial proximity.

The theoretical model can now be built using the sequence-template alignment. Guided by spatial restraints derived from alignment, MODELLER can predict the most probable conformation of target sequence (Eswar et al., 2007; Šali & Blundell, 1993). The spatial information can be retrieved from homology-derived restraints and stereochemical restraints. The homology-derived restraints regarding distances and dihedral angles in target sequence are inferred from its structural alignment with homologous proteins. On the other hand, the stereochemical restraints (bond length and bond angle preferences) are retrieved from the Chemistry at Harvard Macromolecular Mechanics force field (CHARMM22) (MacKerell et al., 1998) and statistical data of dihedral angles and non-bonded atomic distances from databases of known protein structures. Optimisation of the built model is then performed using the variable target function method (VTFM) (Braun & Go, 1985) with conjugate gradients (CG), followed by molecular dynamics (MD) to minimise violations of the spatial restraints.

Upon completion of model building, it is necessary to evaluate the model quality. The Discrete Optimised Protein Energy (DOPE) is a scoring function implemented in MODELLER to rank and to discriminate the most native-like model among a pool of decoys (Shen & Sali, 2006). As an atomic distance-dependent statistical potential, DOPE is based on physical reference state that describes the finite and spherical shape of native proteins. Model with more negative DOPE score assigned can be considered well correlated to native-like model. Additionally,

evaluation of model accuracy can also be made using the GA341 method in MODELLER (John & Sali, 2003; Melo et al., 2002). The GA341 score is calculated by assessing structural compactness, combined statistical potential Z-score of model and sequence identity of the alignment used for modelling. The reliability of the model increases as the GA341 score close to 1.0.

For more robust examination on the stereochemistry of comparative model, a Ramachandran plot (Ramachandran et al., 1963) can be used to highlight potential errors in model regions that require further optimisation. There are many programs such as PROCHECK (Laskowski et al., 1993) and RAMPAGE (Lovell et al., 2003), that include Ramachandran plot analysis for model assessment. First described by Ramachandran (1963), the plot defines the range of allowed phi and psi dihedral angles in a protein that avoid steric collisions between atoms. It is generally accepted that a good model should have above 90% of its residues located in the most favoured regions (Laskowski et al., 1993). Other than that, the spatial properties of built models can be assessed using Verify3D (Bowie et al., 1991; Luthy et al., 1992) and ProSA-web (Sippl, 1993; Wiederstein & Sippl, 2007). In Verify3D, the compatibility of the predicted model structure with its amino acid sequence is verified by its 3D profile. Therefore, a modelled segment with low profile score indicates an incorrect structure. On the other hand, ProSA calculates Z-score of protein model which indicates its overall quality and compares it with scores observed for known protein structures. Besides, regional model quality is examined by measuring energies as a function of amino acid sequence position. Regions with positive energy values hint at erroneous parts of the model.

Although comparative modelling has made great strides in solving 3D structures of target proteins, there are certain limitations to overcome, such as template

identification and alignment accuracy. For these reasons, model assessment becomes an essential step as it ultimately determines the reliability of the comparative model to be used for addressing biological questions. The quality of a comparative model can always be improved by iterating the abovementioned modelling steps until satisfying result is achieved.

2.1.4 Protein-Ligand Docking

Molecular docking is crucial for elucidating interactions between biomolecules at the atomic level to understand the underlying fundamental biochemical processes. Docking studies that involve protein-ligand interactions and protein-protein interactions have garnered considerable research interest given their pharmaceutical and therapeutic significance. Furthermore, the computer-aided molecular docking offers an attractive alternative to determine binding modes and affinities in molecular recognition, in contrary to conventional experimental approaches which are laborious and costly (Huang & Zou, 2010). The subject of this section will be focusing on protein-ligand docking.

Prior to docking, it is necessary to identify putative binding region in protein to increase docking efficiency (McConkey et al., 2002). The binding region can be identified from literatures or with the help of binding site prediction tools. This binding site information helps to narrow down the conformational sampling space and allows intensive conformational sampling of ligand on the predefined protein region. Finally, scoring function is used to discriminate near native docked pose from decoys. In short, the docking performance is closely associated with sampling or scoring.

To tackle docking problem, a variety of docking tools and programs based on different algorithms have been developed. Although some of the newly released

docking programs proclaim to have better accuracy and speed, there are doubts on the validity of results and their performance (Wang et al., 2016). In contrary, traditional docking program such as AutoDock suite is well validated and its predictions are reliable. Besides that, the availability of its interactive graphical user interface simplifies the setup and docking analysis (Morris et al., 2008).

AutoDock suite performs ligand sampling by implementing the stochastic search method. It generates an ensemble of ligand binding poses by randomly changing the translational and rotational position of ligand, together with its torsion angles (Forli et al., 2016). The method is based on genetic algorithm (GA), a popular class of evolutionary algorithm to guide the prediction of optimal ligand binding conformations (Huang & Zou, 2010). In GA, the global minimum energy conformation is searched using user defined rates of cross-over and random mutations. It works in similar manner as biological evolution; binding mode with lowest energy score is ultimately selected. Meanwhile, introduction of Lamarckian genetic algorithm (LGA), an improved version of traditional GA, performs search in local conformational space to identify local minima which will be passed on to its offspring (Morris et al., 1998).

In AutoDock 4, a semiempirical free energy force field is implemented to rank docked decoys (Huey et al., 2007). Unlike its predecessor, the force field incorporates intramolecular energies in binding free energy estimation. This makes the scoring function for docking comparable to function for binding free energy prediction. Hence, it successfully solves the common problem encountered in empirical free energy force fields. There are two aspects evaluated for receptor-ligand binding: intramolecular energetics for both receptor and ligand in their apo and holo forms, intermolecular energetics of receptor-ligand complex. In addition, a novel charge-based desolvation

method with defined atom types and charges is incorporated as well.

To date, numerous molecular interactions have been successfully elucidated using docking approach. It has identified natural inhibitors against the primary targets for cervical cancer (Kumar et al., 2014), helped in developing pheromone trap for rodent pest management by evaluating binding affinity between pheromone compounds and its targeted carrier protein (Rajesh et al., 2016), studied the interactions of new antifungal drugs against an essential enzyme in fungi (Saha et al., 2012) and used in a wide range of applications.

2.1.5 Molecular Dynamics (MD)

MD simulation serves as a powerful *in silico* approach to study molecular motions as a function of time. The physical movement of particles in a system can be deduced by solving Newton's second equation of motion (Leach, 2001). Thus, it generates a trajectory that describes time-dependent behaviour of particles.

Numerous force field models have been developed to simulate different types of biomolecules. Basically, a force field is a mathematical expression that describes the dependence of the energy of a system on its particle coordinates (González, 2011). It can be represented by the following expression:

$$U = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} \lfloor 1 + \cos(n\emptyset - \delta) \rfloor$$

$$+ \sum_{improper} V_{imp} + \sum_{LJ} 4 \in_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}} \qquad \text{(Eq. 2.1)}$$

The first four terms in expression describe bonded term contributions (bond stretching, angle bending, dihedral and improper torsions) to the total energy whereas the last two terms describe non-bonded terms, namely repulsive and van der Waals (vdW) interactions and Coulombic interactions. In simple terms, a force field provides

parameterisation for the energy surface of protein (Guvench & MacKerell, 2008). These force field parameters are derived from quantum mechanical calculations or experimental studies (Adcock & McCammon, 2006). Examples of some popular protein force field models are AMBER (Cornell et al., 1995), CHARMM (MacKerell et al., 1998) and GROMOS (Oostenbrink et al., 2004).

In practice, a standard MD simulation is performed under isothermal-isobaric conditions (Nurisso et al., 2012). Prior simulation, an initial structure (experimentally determined or computationally predicted) is prepared by fixing structural errors such as missing atoms, assigning atom types and charges and adding solvent molecules. The system is simulated in the presence of explicit solvent. The computational cost is saved by applying periodic boundary conditions to minimise non-essential calculations of solvent degree of freedom (Wassenaar & Mark, 2006). Energy minimisation is then carried out to relax the initial structure. After heating up the system to the desired temperature, the system is equilibrated to achieve stability which is usually assessed in terms of energy, density, temperature and pressure. Next, the equilibrated system is subjected to production phase which collects structural and energetic data versus time. Finally, the sampled trajectories are analysed in terms of thermodynamic, structural and dynamical properties.

The applicability of MD has been described for refining comparative models (Nurisso et al., 2012), its role in drug discovery (Durrant & McCammon, 2011) and protein design (Childers & Daggett, 2017). On top of that, it is also frequently used as complementary with docking method because it offers more realistic energy prediction of a bound complex than computational docking (Forli et al., 2016). Given the advancement in computing power, it is now possible to perform an all atom MD simulation involving larger system that consists millions of atoms on an extended time

scale up to millisecond. The development of theories and algorithms that mimic experimental conditions allow to perform realistic MD simulation (Adcock & McCammon, 2006). This enable MD to serve as a computational microscope that illustrates biochemical processes in atomic detail which is difficult to examine in experiment.

2.1.6 Energetic Characterisation of Protein-Ligand Interaction Surface

Binding free energy is main indicator for complex stability, which is central to all biomolecular binding events (Adcock & McCammon, 2006). Using MD, the binding conformations and corresponding binding free energies can be predicted for protein-ligand complexes.

Molecular mechanics generalised Born and Poisson-Boltzmann surface area continuum solvation methods (MM-GBSA/PBSA) are commonly employed in free energy calculations of protein-ligand system. Alternatively, they are known as end-point methods as binding free energy is computed using the bound and unbound states of system. Therefore, it is computationally less demanding.

In general, the binding free energy for protein-ligand complex formation can be obtained as below (Pearlman, 2005):

$$\Delta G_{bind} = G\,(\text{protein-ligand complex}) - G\,(\text{protein}) - G\,(\text{ligand}) \qquad \text{(Eq. 2.2)}$$

On the other hand, the binding free energy for each molecular system (protein, ligand and protein-ligand complex) can be expressed as summation over three components:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} - TS_{solute} \qquad \text{(Eq. 2.3)}$$

In equation 2.3, ΔE_{MM} is molecular mechanics energy in gas phase, ΔG_{solv} is solvation free energy and S_{solute} is the solute entropy. The molecular mechanics term can be further decomposed to bonded and non-bonded energy whereas the solvation free

energy term can be further divided into polar contribution (electrostatic part from solvation model) and non-polar contribution (solvent-accessible surface area, also known as SASA). Due to high computational cost associated with solute entropy term estimation and its lack of conformational information, the term is often ignored in calculation (Genheden & Ryde, 2015).

The reliable performance of MM-GBSA/PBSA approaches and their ability to predict binding free energies comparable to experimental values have been reported previously (Genheden & Ryde, 2012; Rastelli et al., 2010). Despite the success of MM-GBSA/PBSA in predicting ligand binding affinities, there are certain underlying limitations. For example, questionable entropy contributions and data accuracy that heavily depended on conformational space sampled and parameters assigned (Weis et al., 2006). Nevertheless, both approaches are still appealing as they offer fast prediction of receptor-ligand binding thermodynamics at computational level.

2.2 Specific Objectives

i) To predict the three-dimensional (3D) structure of HSP16.3

ii) To evaluate the predicted 3D structure of HSP16.3

2.3 Methodology

The following flow chart outlines the overall research design in predicting HSP16.3 structure (Figure 2.1):

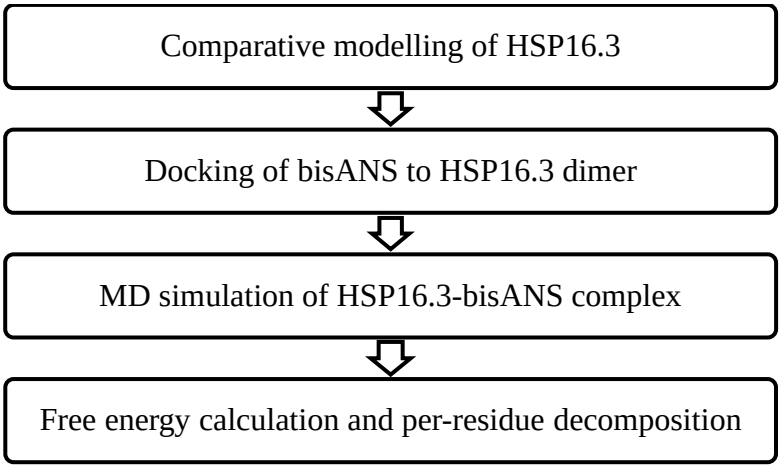

Figure 2.1 The overall methodology flowchart of HSP16.3 structure prediction

2.3.1 Comparative Modelling of HSP16.3

The target sequence of HSP16.3 (accession ID: P9WMK1) was retrieved from the UniProt Knowledgebase (UniProtKB) (The Uniprot Consortium, 2017). Prediction servers were used to perform preliminary sequence analysis by predicting its secondary structures (APSSP2 (Raghava, 2002), Jpred4 (Drozdetskiy et al., 2015), SSpro (Cheng, Randall, et al., 2005), PORTER (Pollastri & McLysaght, 2005), PredictProtein (Rost et al., 2004) and PSIPRED v3.3 (Buchan et al., 2013; Jones, 1999)), disordered regions (DISpro (Cheng, Sweredoski, et al., 2005), IUPred (Dosztányi et al., 2005), Meta-Disorder (Kozlowski & Bujnicki, 2012), PONDR (Li et al., 1999; Romero et al., 2001), PrDOS (Ishida & Kinoshita, 2007) and RONN v3.2 (Yang et al., 2005)) and functional sites (Consurf server (Ashkenazy et al., 2010; Berezin et al., 2004; Celniker et al., 2013)). Suitable candidate templates were identified by Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) against Protein Data Bank (PDB) proteins (Berman et al., 2000), followed by aligning the target and template sequences using MODELLER 9.14 program (Šali & Blundell,

1993). During modelling, symmetry restraints were applied on the HSP16.3 dodecamer to achieve symmetry on each chain. From the pool of comparative models, the best model which was defined as the model with the lowest DOPE score (Shen & Sali, 2006), was selected for further loop refinement. Lastly, the quality of the model was examined using model evaluation servers such as Verify3D (Bowie et al., 1991; Luthy et al., 1992), PROCHECK Ramachandran (Laskowski et al., 1993) and ProSA-web (Sippl, 1993; Wiederstein & Sippl, 2007).

2.3.2 Docking Simulation of bisANS to HSP16.3 Dimer

An early study has shown that HSP16.3 exists in dimeric form after standard HSP isolation procedures (Srivastava et al., 2013). Therefore, docking simulation was performed on HSP16.3 dimer (chain A and B). To evaluate binding properties of HSP16.3 dimer, a known hydrophobic probe for HSP, bisANS was chosen for the task. The AutoDock 4.2.6 software (Morris et al., 2009) and its graphical front-end, AutoDockTools were used in docking simulation. The ligand coordinate file for bisANS was prepared using ChemDraw Professional 15.0. Next, Gasteiger charges were calculated and hydrogen atoms were added for both HSP16.3 dimer and bisANS. The HSP16.3 dimer was regarded as rigid entity whereas flexibility of bisANS was allowed up to seven torsional degrees of freedom. The grid size was defined as 126 × 126 × 126 points, centered at -2.444, -0.152 and 13.965 with grid spacing of 0.375 Å. The docking search parameter chosen was Lamarckian genetic algorithm. The population size was set to 150; maximum number of energy evaluations at 2,500,000; maximum number of generations at 27,000. Default values were used for remaining parameters. A total of 1000 docking runs was performed. All conformations generated were clustered using root mean square deviation (RMSD) tolerance of 2.0 Å.

2.3.3 MD Simulation of HSP16.3-bisANS Complex

MD simulation was performed with AMBER 12 (Case et al., 2012) to evaluate stability of the docked HSP16.3-bisANS complex and to study its dynamics. Since the system involved a protein-ligand complex, the geometry of ligand (bisANS) was first optimised using Antechamber tool (Wang et al., 2006). The partial charges of bisANS were assigned using AMC-B11 charge model (Jakalian et al., 2000; Jakalian et al., 2002) and ligand parameters were defined by general AMBER force field (GAFF) which is specific for small organic molecules (Wang et al., 2004). Next, the system was set up by preparing complex topology and coordinates files required for simulation. The AMBER force field FF14SB (Maier et al., 2015), an significant improvement over its predecessor in terms of optimised dihedral parameters for protein backbone and side chains, was applied along with GAFF. Counterions (14 sodium ions) were added to neutralise the system. For newer force fields such as FF14SB, the ion parameters for TIP3P water (Jorgensen et al., 1983) were required to be sourced. After loading necessary ion parameters, a truncated octahedral TIP3P water box of 12.0 Å was added to solvate the docked complex. The total number of atoms in system were 106, 909. Prior to MD, the solvated complex was subjected to two rounds of minimisation. During the first minimisation stage, 1000 minimisation cycles were performed with positional restraints (force constant of 100 kcal/mol/$Å^2$) to fix the solvated complex at reference position. The use of restraints could be useful to prevent structural distortions during beginning of energy minimisation (Struthers *et. al.*, as cited in Greer, 1991). The subsequent minimisation stage which involved 2000 minimisation cycles was carried out without positional restraints. The minimisation method changed from steepest descent to conjugate gradient upon reaching half of the maximum number of minimisation cycles. After minimisation, the solvated system was heated up to 300 K

within 100 ps. The system temperature was regulated by Langevin dynamics with the collision frequency of 2.0 ps^{-1}. Besides, the SHAKE algorithm was applied to constrain bonds involving hydrogen (Ryckaert et al., 1977). The solvated system was maintained at constant volume. Next, the solvated system was equilibrated for 900 ps to relax the water box before running 10 ns of production. Both equilibration stage and production stage were simulated at constant pressure, controlled by Berendsen at constant temperature to mimic laboratory conditions. The MD simulation was performed under periodic boundary condition with Particle Mesh Ewald (PME) cutoff at 8.0 Å. The time step in simulation was 2 fs. The system equilibrium was monitored in terms of energy, temperature, pressure, volume, density and RMSD. A MD representative structure was extracted from the equilibrated trajectory (10^{th} ns) via clustering using MMTSB toolset (Michael Feig et al., 2004). All protein structure visualisations were generated using PyMOL (Delano, 2002).

2.3.4 Free Energy Calculations and Per-Residue Decomposition

The binding free energy of HSP16.3-bisANS complex was calculated using MM-GBSA/PBSA. The calculations were performed using snapshots extracted at 10 ps intervals from single MD trajectory of complex at the 10^{th} ns MD simulation trajectory. In MM-GBSA, the polar solvation free energy was approximated by applying the modified GB^{OBC} (II) model (Onufriev et al., 2004) (igb=5) using mbondi2 radii set. The Linear Combination of Pairwise Overlaps (LCPO) method (Weiser et al., 1999) was used for determining the non-polar solvation free energy. The surface tension parameter used was 0.0072 kcal/mol/A^2. In MM-PBSA, the polar solvation free energy was calculated by solving the PB equation. The non-polar solvation free energy was calculated by classical model (inp=1) using surface tension parameter that

was set at 0.005 kcal/mol/A^2. The radii from topology files (radiopt=0) were used for PB calculation and its non-polar calculations. By using the same set of radii for both MM-GBSA/PBSA approach, it ensures consistency between the calculations for comparison purpose. The interior and exterior dielectric constant was set to 1.0 and 80.0 in both MM-GBSA/PBSA calculations. Due to surface hydrophobicity of HSP16.3, lower interior dielectric constant was employed in this study. It has been reported that low solute dielectric constant (ε_{in} = 1) is generally applicable for hydrophobic environment whereas high solute dielectric constant (ε_{in} > 1) is suitable for charged environment (Hou et al., 2011a). The interaction energy between HSP16.3 and bisANS was also studied by decomposing the total binding free energy into its individual contributions on a per-residue basis.

2.4 Results

2.4.1 Sequence Analysis of HSP16.3

Secondary structure predictions showed HSP16.3 was overrepresented in beta strands (Table 2.1). Based on the consensus prediction, beta strands were found distributing among residues 42-46, 50-56, 66-71, 74-80, 90-94, 96-103, 113-117, 121-127 and 137-141. Besides, the N-terminal (residues 15-23) was predicted to adopt helical conformation. The proposed secondary structures of HSP16.3 is consistent with protein disordered region prediction results. The protein residues which presumably involved in forming secondary structure element were also predicted to fall into ordered regions of HSP16.3 (Table 2.2). The results showed that residues 1-7, 56, 79-84 and 142-144 were disordered regions, as agreed by prediction servers. Therefore, these identified regions are likely unstable and do not have regular structures.

Table 2.1 Secondary structure prediction by APSSP2 (Raghava, 2002), Jpred4 (Drozdetskiy et al., 2015), SSpro (Cheng, Randall, et al., 2005), PORTER (Pollastri & McLysaght, 2005), PredictProtein (Rost et al., 2004) and PSIPRED v3.3 (Buchan et al., 2013; Jones, 1999). Indicator: H – alpha helix; B – beta strand

	1	11	21	31	41	51
HSP16.3 sequence	MATTLPVQRH	PRSLFPEFSE	LFAAFPSFAG	LRPTFDTRLM	RLEDEMKEGR	YEVRAELPGV
APSSP2	----------	HHHHHHHHHH	HHHHH-----	----------	---BBB---B	BBBBBB----
Jpred4	----------	------HHHH	HH--------	----------	-BBBBB---B	BBBBBBB---
SSpro	------H---	---HHHHHHH	HH--------	----------	-BBBBB---B	BBBBBB----
PORTER	----------	----HHHHHH	HHH-------	----------	-BBBBB---B	BBBBBB----
PredictProtein	----------	HHHHHHHHHH	HHHH------	----------	-BBBBB---B	BBBBBBB---
PSIPRED v3.3	----------	-----HHHHH	HHH-------	---------B	BBBBBBB--B	BBBBBB----
Consensus	----------	----HHHHHH	HHH-------	----------	-BBBBB---B	BBBBBB----
	61	71	81	91	101	111
	DPDKDVDIMV	RDGQLTIKAE	RTEQKDFDGR	SEFAYGSFVR	TVSLPVGADE	DDIKATYDKG
APSSP2	----BBBBBB	B--BBBBBBB	B-------BB	BBBBB-BBBB	BBB------H	HHBBBBB---
Jpred4	----BBBBBB	B--BBBBBBB	--------BB	BBBBBBBBBB	BB--------	--BBBBB---
SSpro	-----BBBBB	---BBBBBBB	----------	BBBB---BBB	BBB------H	HHBBBBB---
PORTER	------BBBB	B--BBBBBBB	----------	-----BBBBB	BBB------H	HH-BBBB---
PredictProtein	---BBBBBBB	B--BBBBBBB	----------	-----BBBBB	BBB-------	-BBBBBBB--
PSIPRED v3.3	-----BBBBB	B--BBBBBBB	BBBBB----B	BBBBBBBBBB	BBB-------	---BBBB---
Consensus	-----BBBBB	B--BBBBBBB	---------B	BBBB-BBBBB	BBB-------	--BBBBB---
	121	131	141			
	ILTVSVAVSE	GKPTEKHIQI	RSTN			
APSSP2	BBBBBBB---	-----BBBBB	B---			
Jpred4	BBBBBBBB--	------BBBB	B---			
SSpro	BBBBBB----	------BBBB	B---			
PORTER	BBBBBB----	-----BBBBB	BB--			
PredictProtein	BBBBBBB---	-----BBBBB	B---			
PSIPRED v3.3	BBBBBBB---	------BBBB	BB--			
Consensus	BBBBBBB---	------BBBB	B---			

Table 2.2 Prediction of protein disordered regions in HSP16.3 by DISpro (Cheng, Sweredoski, et al., 2005), IUPred (Dosztányi et al., 2005), Meta-Disorder (Kozlowski & Bujnicki, 2012), PONDR (Li et al., 1999; Romero et al., 2001), PrDOS (Ishida & Kinoshita, 2007) and RONN v3.2 (Yang et al., 2005).

	1	11	21	31	41	51
HSP16.3 sequence	MATTLPVQRH	PRSLFPEFSE	LFAAFPSFAG	LRPTFDTRLM	RLEDEMKEGR	YEVRAELPGV
DISpro	**DDDDDDDDD**O	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
IUPred	**DDD**OO**D**OOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	**DDDDDD**OOOO
Meta-Disorder	**DDDDDDDDDD**	**DDDDDDD**O**DD**	**DDDDDDDDDD**	**DDDDDDDDDD**	**DDDDDDDDDD**	OOOOO**DDDDD**
PONDR	**DDDDDDD**OOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	O**DDDDDDDDD**	**DDDDDDDDDD**
PrDOS	**DDDDDDDDDD**	**DD**OOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
RONN v3.2	**DDDDDD**OOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOO**DDDDDDD**	**DDDDDDDD**OO
Consensus	**DDDDDDD**OOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOO**D**OOOO
	61	71	81	91	101	111
	DPDKDVDIMV	RDGQLTIKAE	RTEQKDFDGR	SEFAYGSFVR	TVSLPVGADE	DDIKATYDKG
DISpro	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
IUPred	OOOOOOOOOO	**DD**OOOOOO**DD**	**DD**OOOOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
Meta-Disorder	**DDDDD**OOOOO	O**DDDDDDDDD**	**DDDDDDDDDD**	**DDDDDD**OOOO	O**DDDDDDDDD**	**DDDDD**OOO**D**O
PONDR	**DDDDDDDDDD**	**DDDDDDDDDD**	**DDDD**OOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
PrDOS	OOOOOOOOOO	OOOOOOOOOO	OO**DDDDDDD**	**D**OOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
RONN v3.2	OOOOOOOOOO	OOOO**DDDDDD**	**DDDDDD**OOOO	OOOOOOOOOO	OOO**DDDDDDD**	**DDDDDDDD**OO
Consensus	OOOOOOOOOO	OOOOOOOO**DD**	**DDDD**OOOOOO	OOOOOOOOOO	OOOOOOOOOO	OOOOOOOOOO
	121	131	141			
	ILTVSVAVSE	GKPTEKHIQI	RSTN			
DISpro	OOOOOOOOOO	OOOOOOOOOO	OO**DD**			
IUPred	OOOOOOOOOO	OO**DDDDDDDD**	**DDDD**			
Meta-Disorder	OOOO**DDDDDD**	**DDDDDDDDDD**	**DDDD**			
PONDR	OOOOOOOOOO	OOOO**D**OOOOO	OOO**D**			
PrDOS	OOOOOOOOOO	OOOOO**DDDDD**	**DDDD**			
RONN v3.2	OOOOOOOOOO	OOOOOOOOOO	O**DDD**			
Consensus	OOOOOOOOOO	OOOOOOOOOO	O**DDD**			

Indicator: **D** – disordered region; O – ordered region