STRUCTURAL AND FUNCTIONAL PREDICTION OF HYPOTHETICAL PROTEINS FROM *KLEBSIELLA PNEUMONIAE* MGH78578: MOLECULAR MODELLING STUDIES

by

CHOI SY BING

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy of Science

OCTOBER 2011

ACKNOWLEDGEMENT

I would like to acknowledge the unconditioned supports and helps of those who helped and supported me. Without the support, this piece of work cannot be completed. Many people did indeed give in a big influence to me during this process.

First of all, I would like to express my sincere thanks to my supervisor and co-supervisor, Professor Dr Habibah A Wahab and Dr Normi M Yahaya for their training and thoughtful advice through my PhD study. I would like to thank them for providing me unlimited resources, insightful suggestion as well as encouragement and moral support to make me motivated during my research. They had given me such a great opportunity for me to go to Japan to broaden my research perspective. It was an invaluable experience to me. Once again thank you to both of you for guiding me through this rich and rewarding journey.

Next, thanks to my lab members. In the first few year of my study, Dr Wai Keat and Yee Siew were in the lab and both helped me to adapt to the research environments and also make my life in the lab much enjoyable. Wai Keat had encouraged me when things were not going on smoothly, she was also the person who saw me crying and get disappointed and depressed when things failed. Yee Siew (Dr Choong), she is not only my labmate, lunchmate and also jogging mate. Daily jogging routine with Wai Keat and Yee Siew, many happy and unhappy moments we spent together and I appreciate that.

Further, thanks go to others member and ex-members of our lab Dr Nurul, Belal, Imtiaz, Sue, Nani, Fatihah, Pak Muctharidi, Adliah, Lim, Lee and Ban Hong.

Without them, the lab atmosphere will not be fun and motivated. I’m very fortunate to be in such a friendly and cooperative group. I’m also grateful to the Ministry of Science, Technology and Innovation for the PASCA Siswazah Scholarship and USM Fellowship for funding me throughout my study.

I would also like to thank Pooh Koon, for his absolute love, patience and sacrifice. I’m happy that we will continue our journey through life together soon.

At last, my eternal gratitude to my families: my parent, my brother, my sister in law and both my cute little nephew and niece. They are always there for me, standing by me all the time. And last but not least to whom I had forgotten their name in my forgetful mind, without all of you, this would not be possible.

Sy Bing, March, 2011

TABLE OF CONTENT

CHAPTER TWO: METHODOLOGY

CHAPTER THRE: RESULT AND DISCUSSION

CHAPTER FOUR: CONCLUSIONS AND FUTURE STUDY

LIST OF TABLES

LIST OF FIGURES

Multiple sequence alignment of KPN00728 and KPN00729 with seven other members of the Enterobacterociae family. Ser27 and Arg31 in KPN00728 and Tyr83 in KPN00729 are highly conserved among the other seven enterobacters. Possible interactions with UQ are found among these residues. His30, His84 and His91 from KPN00728 and His71 from KPN00729 are also highly conserved. These residues were located closely with the heme group of SDH. These residues are postulated to be either directly or indirectly involved in the binding of heme group to KPN00728 and KPN00729. * indicated as conserved regions.

LIST OF ABBREVIATIONS

SdhC	Succinate dehydrogenase chain C
SdhD	Succinate dehydrogenase chain D
UQ	Ubiquinone
SDH	Succinate dehydrogenase
MD	Molecular dynamics
Ser	Serine
Arg	Arginine
Tyr	Tyrosine
Glu	Glutamic acids
Asp	Aspartic acids
DNA	Deoxyribonucleic acid
UTI	Urinary tract infection
ESBL	Extended-spectrum beta-lactameses
MDR	multidrug–resistant
kDa	Kilo Dalton
NMR	Nuclear Magnetic Resonance
BLAST	Basic local alignment search tool
PDB	Protein Data Bank
E-value	Expected value
3D	Three Dimensional
CHARMM	Chemistry at Harvard Molecular Mechanics
DOPE	Discrete optimized potential energy
3D-PSSM	Three Dimensional position-specific scoring matrix
PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool
PSSM	Position-specific scoring matrix
HMM	Hidden Markov Model
CASP	Critical Assessment of Techniques for Protein Structure Prediction
N-terninal	Amine terminal
C-terminal	Carboxyl terminal
GROMACS	Groningen Machine for Chemical Simulations
NAMD	Not (just) Another Molecular Dynamics program
LJ	Lennard-Jones
HP	Hypothetical protein
PC	Phosphatidylcholine
PA	Phosphaditic acid
PG	Phosphatidyl glycerol
PE	Phophatidyletanolamines
POPC	Palmitoyl oleoyl phosphatidyl choline
POPS	Palmitoyl oleoyl phosphatidylserine
POPG	Palmitoyl oleoyl phosphatidylglycerol
DPPC	Dipalmitoyl Phosphatidylcholine
DMPC	Dimyristoylphosphatidylcholine
POPA	Palmitoyl oleoyl phosphatidyl acid
POPE	Palmitoyl oleoyl Palmitoyl oleoyl
DMTAP	Dimyristoyltrimethylammonium propane
VMD	Visual Molecular Dynamics
SPC	Single point charge

SD	Steepest Descent
CG	Conjugate gradient
NVT	Canonical ensemble
NPT	Isothermal-isobaric ensemble
PME	Particle Mesh Ewald
PBC	Periodic boundary condition
E.coli	*Escherichia coli*
FAD and	Flavin Adenine Dinucleotide
FADH$_2$	Reduced form of FAD (1,5 dihydro-FAD)
TM	Transmembrane
RMSD	Root mean square deviation
TST	transition state theory
KIE	kinetic isotope effect
SLO-1	soybean lipoxygenase-1
QH$_2$	Ubiquinol
Cys	Cysteine
His	Histidine
SI	Sequence identity
S$_{cd}$	Deuterium tail order parameter
RMSF	Root mean square fluctuation
RDF	Radial distribution function
HB	Hydrogen bonding

LIST OF PUBLICATION SEMINAR AND CONFERENCES

1 Choi SB, Normi YM and Wahab HA (2009) Why Hypothetical Protein KPN00728 of *Klebsiella pneumoniae* should be classified as Chain C of Succinate dehydrogenase? Protein J 28:415-427.

2 Choi SB, Normi YM and Wahab HA (2011). Revealing the functionality of Hypothetical Protein KPN00728 from *Klebsiella pneumoniae* MGH78578: Molecular Dynamics Simulation Approches InCoB/ISCB-Asia 2011 BMC Bioinformatics 12 Suppl 11, S11.

3 Kuan CS, Wong MT, Choi SB, Chang CC, Yee YH, Wahab HA, Normi YM, See Too WC and Few LL (2011) *Klebsiella pneumoniae* yggG gene product is a zinc dependent metalloprotease Int. J. Mol. Sci. 12(7), 4441-4455.

4 Choi SB, Wahab, HA and Chan, HY; Preliminary Study of Distant Homology Protein Structure Prediction. 4th USM-Life Sciences Postgraduate Conference, Pulau Pinang, Malaysia, 18-20 June 2008.

5 Choi SB, Normi YM and Wahab HA. Medical News Article on *Klebsiella pneumoniae*. 1st Feb 2010. http://www.newsrx.com/newsletters/Proteomics-Weekly/2010-02-01/3602012010422PW.html

6 Choi SB, Normi YM and Wahab HA; Structural and functional prediction using computational method on hypothetical proteins: A case study in *Klebsiella pneumoniae* MGH78578 pathogen. Second Collaborative Conference UNAIR-USM 2009, Surabaya, Indonesia, 10-11 February 2009. (Best Poster Award)

7 Choi SB, Normi YM and Wahab HA; Sequence Analysis and Molecular Docking Simulation Approaches to Predict the Function of Hypothetical Protein from *Klebsiella pneumoniae*; 13th Annual Symposium on Computational Sciences and Engineer (ANSCSE13), Bangkok, Thailand, 25-27 March 2009.

8 Choi SB, Normi MY and Wahab HA; From Structure Prediction to Molecular Dynamics Simulation: A Case Study on Functional prediction of Hypothetical Protein from *Klebsiella pneumoniae* MGH 78578 Pathogen; Second National Seminar and Workshop on Computer Aided Drug Design, Pulau Pinang, Malaysia, 8-11 December 2009.

9 Choi SB, Normi YM and Wahab HA; Membrane Protein Simulation: A case study on Selected Hypothetical protein from *Klebsiella pneumoniae* MGH78578: 14th Annual Symposium on Computational Sciences and Engineer (ANSCSE14), Chieng Rai, Thailand, 23-26 March 2010.

10 Choi SB, Normi YM and Wahab HA Solvation effect on binding of postulated Succinate dehydrogenase with UQ from *Klebsiella pneumoniae* MGH78578; 1st International Conference on Computation for Science and Technology (ICCST-I) Chiang Mai, Thailand, 4-6 August 2010.

11 Ueta D, Choi SB, Matsuda H and Wahab HA; Virtual screening: A method for improvement of parallelization job submission using Autodock 3.05. International Seminar and Expo on Jamu 2010, Bandung, Indonesia. 5-7 Nov 2010.

12 Teh BA, Choi SB, Najimuddin MNM, Wahab HA and Normi YM. Deciphering the Function of YcbK Lipoprotein-like Hypothetical protein *from Klebsiella pneumonia* MGH78578. International Symposium on Women in Science and Engineering (WISE 2011), Kuala Lumpur, Malaysia, 29-30th September 2011.

RAMALAN STRUKTUR DAN FUNGSI PROTEIN HIPOTETIKAL PADA *KLEBSIELLA PNEUMONIAE* MGH78578: KAJIAN PEMODELAN MOLEKUL

ABSTRAK

Dua puluh peratus gen daripada MGH78578 *Klebsiella pneumonaie* mengkod protein hipotetikal. Dua protein hipotetikal KPN00728 dan KPN00729 telah dikenalpasti dengan menggunakan pendekatan bioinformatik. Kedua-dua rangka bacaan terbuka menunjukkan homologi jujukan tinggi kepada suksinat dehidrogenase rantai C (SdhC) dan D (SdhD) daripada *Escherichia coli.* KPN00729 dikenalpastikan sebagai SdhD pada Mei 2008. Malah penyelidikan terhadap KPN00728 tetap tidak diketahui kerana tidak ada anotasi bagi gen SDHC dalam jujukan genom lengkap daripada *Klebsiella pneumoniae* MGH78578. Dalam kajian ini, KPN00728 mempunyai kawasan hilang yang mengandungi residu yang penting bagi ikatan ubiquinone (UQ) dan kumpulan Heme. Fungsi KPN00728 dengan gabungan analisis struktur sekunder dan topologi transmembran menunjukkan KPN00728 terima guna SDH-struktur (C subunit). Bagi mengkaji fungsinya dengan lebih mendalam, UQ telah didokkan pada model yang dibina (terdiri daripada KPN00728 dan KPN00729) dan pembentukan ikatan hidrogen antara UQ dengan Ser27, Arg31 (KPN00728) dengan Tyr83 (KPN00729) lebih menguatkan dan menyokong bahawa KPN00728 adalah suksinat dihidrogenase (SDH). Namun demikian, keterbatasan dalam simulasi megedok gagal untuk memberikan pemahaman mendalam tentang interaksi SDH yang berada pada trans-membran mitokondria. Simulasi dinamik molekul (MD) KPN00728 dan rantai D dalam membran dilakukan bagi melihat peranan molekul SDH. Kestabilan struktur telah ditunjukkan dalam pengiraan pada kawasan lipid, susunan parameter ekor, ketebalan lipid dan sifat struktur sekunder. Menariknya, molekul air yang ditemui mungkin

lebih menyebabkan penyimpangan interaksi UQ dengan SDH di Ser27 dan Arg31 dibandingkan dengan kajian pendokan sebelumnya. Residu polar seperti Asp95 dan Glu101 (SDH rantai C), Asp15 dan Glu78 (SDH rantai D) mungkin telah menyumbangkan penciptaan lingkungan polar yang sangat penting bagi rantai pengangkutan elektron dalam kitaran Krebs. Walaupun terdapat perbandingan kestabilan struktur, interaksi dinamik telah banyak membuktikan bahawa interaksi KPN00728 sebagai SDH adalah lestari dan juga menepati dengan postulasi kami sebelum ini.

STRUCTURAL AND FUNCTIONAL PREDICTION OF HYPOTHETICAL PROTEINS FROM *KLEBSIELLA PNEUMONIAE* MGH78578: MOLECULAR MODELLING STUDIES

ABSTRACT

Twenty percent of the genes from *Klebsiella pneumonaie* MGH78578 coded for hypothetical protein. Two particular hypothetical proteins KPN00728 and KPN00729 were identified using bioinformatics approaches. Both open reading frames showed high sequence homology to succinate dehydrogenase Chain C (SdhC) and D (SdhD) from *Escherichia coli* KPN00729 was annotated as SdhD in May 2008. Thus, investigation on KPN00728 remained as no annotation for SdhC gene in the complete genome sequence of *Klebsiella pneumoniae* MGH78578. In this study, KPN00728 has a missing region with conserved residues which is important for ubiquinone (UQ) and heme group binding. Structure and function prediction of KPN00728 coupled with secondary structure analysis and transmembrane topology showed KPN00728 adopts SDH-(subunit C)-like structure. To further probe its functionality, UQ was docked on the built model (consisting KPN00728 and KPN00729) and formation of hydrogen bonds between UQ and Ser27, Arg31 (KPN00728) and Tyr83 (KPN00729) further reinforced and supported that KPN00728 is indeed succinate dehydrogenase (SDH). However, limitation in docking simulation failed to provide in depth understanding of the SDH interaction that occurs in the trans-membrane of mitochondria. For more insight into its molecular role as SDH, molecular dynamics (MD) simulation of KPN00728 and Chain D in a membrane was performed. Structural stability was demonstrated in the calculation in area per lipid, tail order parameter, thickness of lipid and secondary structural properties. Interestingly, water molecules were found to be highly possible for the deviation of interaction of UQ with SDH in Ser27 and Arg31 as compared

with earlier docking study. Polar residues such as Asp95 and Glu101 (SDH chain C), Asp15 and Glu78 (SDH chain D) might have contributed in the creation of a polar environment which is essential for the electron transport chain in Krebs cycle. Despite the structural stability comparability, the dynamics of the interaction had further proved that the interaction of KPN00728 as SDH is preserved and well agreed with our postulation earlier.

CHAPTER 1

INTRODUCTION

1.1 Statement of problem

Klebsiella pneumoniae is a Gram negative, non motile and rod-shaped bacterium. It is named after a German microbiologist Edwin Klebs in 19th century (Figure 1.1). The genus Klebsiella belongs to the tribe Klebsiellae and it is a member of the family Enterobacteriaceae which has a prominent polysaccharide capsule (Philippon *et al.*, 1989). The resistance mechanisms against most hosts come from this capsule which encases the entire cell surface (Tsay *et al.*, 2002). Classification of *Klebsiella* is based on the structural variability of the antigens which are expressed on their cell surface. There are two types of antigens, the first is lipopolysaccharide and the other is a capsular polysaccharide (Philippon *et al.*, 1989). Both antigens are pathogenic. There are about 77 capsular antigens and 9 lipoplysaccharide identified to which exist till date (Orskov and Mfife-Asbury, 1977; Toivanen *et al.*, 1999).

At present, 7 species of klebsiella are known which had shown DNA homology. These are *Klebsiella pneumoniae, Klebsiella ozaenae, Klebsiella planticola, Klebsiella rhinoscleromatis, Klebsiella oxytoca, Klebsiella terrigena,* and *Klebsiella ornithinolytica. Klebsiella pneumoniae* is the most medically important species of the group which is responsible in most human infections (Ko *et al.*, 2002).

Klebsiella is known as an opportunistic pathogen found in the environment and specifically in mammalian mucosal surfaces. They appeared as normal flora of the intestinal tract but usually low in number as compared to *Escherichia coli.* Generally, *Klebsiella* infections tend to occur in patient with a weakened immune system and

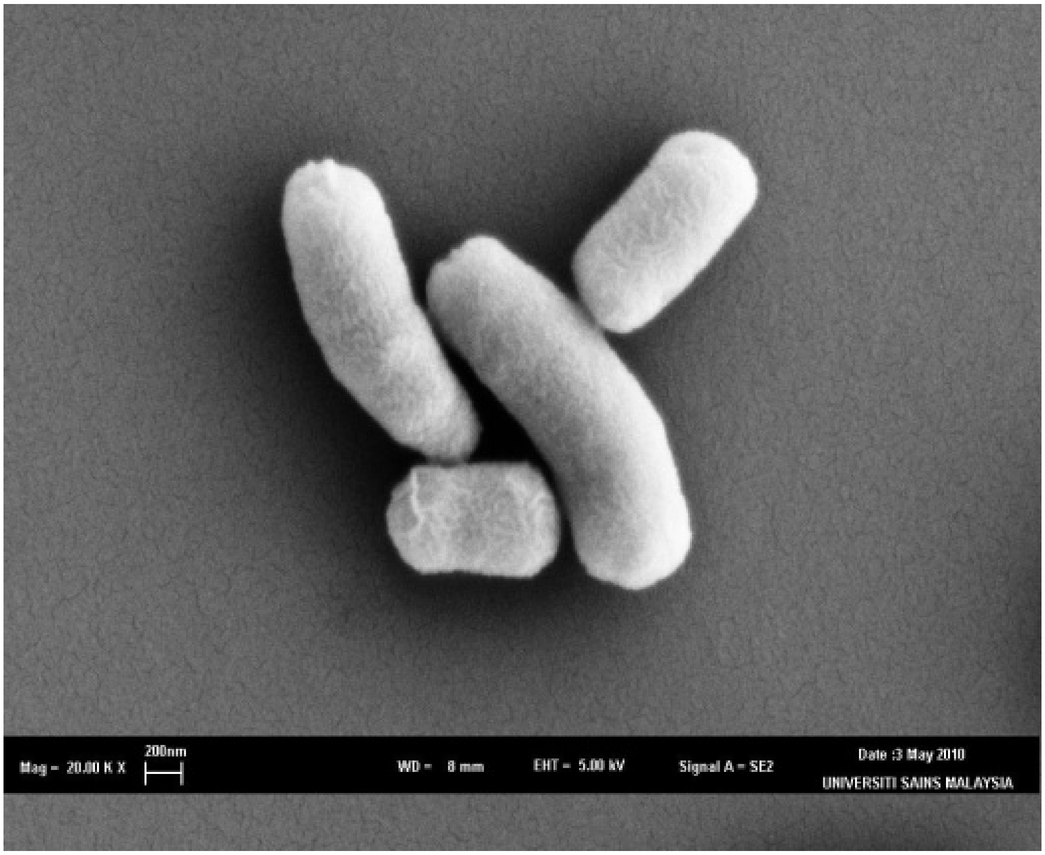

Figure 1.1 Scanning electron microscopy of wild type *Klebsiella pneumonia* MGH78578 (with the permission of Mr Teh Boon Aun).

people with underlying diseases (Kawai, 2006). The principal pathogenic reservoirs of infection are the gastrointestinal tract of patients and the hands of hospital personnel (Marshall, 1991; Obiamiwe and Leonard, 2006). It can spread rapidly, often leading to nosocomial outbreaks. Infections of *Klebsiella* often occur at urinary tract, respiratory tact, biliary tract, and surgical wound sites (Osazuwa *et al.*, 2010; Obiamiwe and Leonard, 2006). Common clinical symptoms include pneumonia, bacteremia (Yinnon *et al.*, 1996), thrombophlebitis, urinary tract infection (UTI)(one of the most common infections (Okadeinde *et al.*, 2011)), cholecystitis, diarrhea, upper respiratory tract infection, wound infection, osteomyelitis, and meningitis. Studies conducted in Asia (Japan and Malaysia) estimated that the incidence rate in elderly persons to be 15-40% (Obiamiwe and Leonard, 2006), which is equal to, if not greater than, that of *Haemophilus influenzae*. The occurrences are likely to be far more common in Asia than elsewhere (Ko *et al.*, 2002). The emergence of multi-drug resistance as in extended-spectrum beta-lactamases (ESBL) in *K. pneumoaniae* has also been reported in the past decade (Paterson *et al.*, 2004) as this become a major concern clinically. Although the incident of community acquired *K. pneumoniae* has apparently decreased, the mortality rate remains twice higher (Kang *et al.*, 2006) as a result of the underlying disease that's tends to be present in affected patients (Wiener *et al.*, 1999; Carpenter, 1990). These rapid boosted incidences deserved to be investigated and delineated.

Recently, genome sequence determination for the complete genome of *K. pneumoniae* had been accomplished in the middle of year 2007 by Genome Research Center of Washington University of St. Louise (NCBI, 2007). It consists of about 5 million of nucleotides and this complete genome map of *klebsiella sp* has enabled us to identify the important part of the genome, eg. Regulatory regions which control the regulatory mechanisms can be identified from turning on or off at a particular gene. However, the major challenge of biomedical research currently is to characterise the properties and biological functions not only from the genes but also from the proteins. There are a total of 4894 genes out of which 4776 genes are encoding proteins in *K. pneumoniae*. Further analysis showed that from the 4776 protein coding genes, there is about twenty percent of the genes is annotated poorly and is classified as hypothetical gene. A hypothetical gene nevertheless will eventually be translated theoretically into a protein sequence which in turn will be identified as a hypothetical protein. Majority of the functional aspect of these proteins are not known and hence, deserving an investigation as they represent a rather large part of the bacterial proteins and they might play important roles towards improved understanding of biological functions.

In this project, the goal is to study the hypothetical protein of *K. pneumoniae* using bioinformatics approaches with two specific aims: To identify novel structure and to characterize the functional and structural features of the hypothetical protein. In order to gain deeper understanding on the functional aspect of the hypothetical proteins, the first approach is to predict its structure. Different methodologies such as comparative genomics, homology modeling and fold recognition could be adopted in line to produce highly accurate structure of which the function of these proteins can be postulated. Once the protein structure is known, many computational modelling approaches can be used for better understand on aspects such as ligand binding, protein-protein interactions, receptor activation, or effects of structure and activity. This information can then act as a platform in establishing the mechanisms of the hypothetical protein and the pathogenecity of *K. pneumoniae* in turn can be further understood in the future. With that in mind, the specific objectives of this research are:

1. To select a hypothetical protein with important biological function in *K. pneumoniae* using computational/bioinformatics approach.
2. To predict the structure of the selected hypothetical protein by comparative genomics and homology modeling.
3. To study the function of hypothetical protein using molecular docking and molecular dynamics simulation approaches.

1.2 Literature Review

1.2.1 K. pneumoniae infection

The non-motile and gram negative bacteria, *Klebsiella pneumonaie* is known as the most common species among the family that associated with human disease was found in 19th century in Germany (Figure 1.1). Although the bacteria is known for over hundreds years, there are still many unanswered question for scientists to reveal.

Klebsiella sp. can be found naturally as a normal flora in gastrointestinal tract or in biliary tract of human and animal (Marshall, 1991). They may colonize skin, pharynx or gastrointestinal tract (Marshall, 1991). They may also colonize sterile wounds and urine (Obiamiwe and Leonard, 2006). It is an opportunistic pathogen; when the immunity is low in the body, the *Klebsiella* infections could occur. The most common infection caused by these bacteria is pneumoniae and it usually occurs in middle age and older men with underlying diseases such as alcoholism, diabetes (Chen *et al.*, 2000) and lung diseases (Marrie and File, 2010; Montgomerie and Ota, 1980). Infection with *Klebsiella* organisms frequently occurs in the lungs, where they cause destructive changes (Osazuwa *et al.*, 2010). Necrosis, inflammation, and hemorrhage occurred within lung tissues, sometimes produce thick and bloody mucous sputum (also described as currant jelly sputum) (Obiamiwe and Leonard, 2006). Mortality rate of this infection is 20 to 50% (Cryz *et al.*, 1985; Montgomerie and Ota, 1980) but can reach up to almost 100% in alcoholic patient that suffer from bacteremia.

Pneumonia that caused by *Klebsiella* is usually indistinguishable from the normal streptococcal pneumonia in term of the associated symptoms such as high fever, chills flu-like symptoms body aches and productive cough with a great deal of sputum (Brook, 2007). However, a patient with normal streptococcal pneumonia can recover without any complication but this is not the case of pneumonia that caused by *Klebsiella sp*, where lung tissues destruction and abscesses are always found in the patient. *Klebsiella* infection also has been identified to be one of the common infections found in neonatal intensive care units (Podschun and Ullmann, 1998), thus it becomes a major concern in infections among pre-mature infants in pediatric wards.

Klebsiellae have also been incriminated in nosocomial infections (Tsay *et al.*, 2002; Obiamiwe and Leonard, 2006). Common sites include the urinary tract, lower respiratory tract, biliary tract, and surgical wound sites. The spectrum of clinical syndromes includes pneumonia, bacteremia, thrombophlebitis, urinary tract infection (UTI), cholecystitis, diarrhea, upper respiratory tract infection, wound infection, osteomyelitis and meningitis. The presence of invasive devices, contamination of respiratory support equipment, use of urinary catheter, and use of antibiotics are factors that increase the likelihood of nosocomial infection with *Klebsiella* species. Sepsis and septic shock may follow entry of organisms into the blood from a focal source.

Symptoms such as UTI, rhinoscleroma and ozena which cause by some other species of klebsiella have also been reported. *Klebsiella sp* is increasingly isolated in patients that have invasive devices such as catheter, feeding tube on. Both rhinoscleroma and ozena are known to be caused by *K. oxtoca* and *K. ozaenae*. Rhinoscleroma is a chronic granulamatous infection on nose which was found to be endemic in several countries such as Egypt and San Salvador (North *et al.*, 1982; Paul *et al.*, 1993; Shum *et al.*, 1982). As for ozane, it also attacks the nose. The occurrence of both the diseases however is rare and is not fatal (Goldstein *et al.*, 1978).

1.2.2 Treatment for Klebsiella pneumoniae

The general treatment of *Klebsiella* in the early days is with the beta-lactam antimicrobials such penicillin, ampicillin and amoxillin. Nevertheless, the extensive use of these broad-spectrum antibiotics in hospitalized patients has led to both increased infections of Klebsiellae and, subsequently, the development of multidrug-resistant strains that produce extended-spectrum beta-lactamase (ESBL) (Philippon *et al.*, 1989). Outbreaks of *Klebsiella sp* where the resistant strain were found had been reported by many (Bradford, 2001; Livermore *et al.*, 2007; Ben-Hamouda *et al.*, 2003; Haryani *et al.*, 2007). ESBL enzyme which consists of capsular type K55 is capable of destroying cephalosporins by cleaving the beta-lactam ring in the antibiotics. These multidrug strains are highly virulent and have an extraordinary ability to spread (Obiamiwe and Leonard, 2006; Kumar and Talwar, 2010). Most outbreaks are due to a single clone or single gene; the major site of colonization with infection of the urinary tract, respiratory tract and wounds appears in bowel (Obiamiwe and Leonard, 2006; Kumar and Talwar, 2010). Bacteremia infection in blood namely bacteremia significant increased mortality has also resulted from infection with these species (Kumar and Talwar, 2010).

Prior to antibiotic use, the presence of invasive medical apparatus in a pateint such as indwelling catheter, feeding tubes, poor health status as well as an intensive care patient are significantly increases the risk factors for infection and treatment (Obiamiwe and Leonard, 2006; Ben-Hamouda *et al.*, 2003). Acquisition of these species has become a major problem in most hospitals because of resistance to multiple antibiotics and potential transfer of plasmids to other organisms.

In Malaysia, *Klebsiella pneumoniae* is one of the high ranking community-acquired pneumonia among patient in local hospital (Loh *et al.*, 2007; Loh *et al.*, 2004). Navaratnam and coworkers (Palasubramaniam *et al.*, 2005) had reported an outbreak caused by *K. pneumonia* in a local hospital. They had isolated an imipenem-resistant strain of *K. pneumoniae* and believed that to be an association of ESBL SHV-5. Characterization of multidrug–resistant (MDR) and extended-spectrum β-lactamase-producing *K. pneumoniae* strains from Malaysia hospitals has been carried out in 2009 (Lim *et al.*, 2009) where more than 50% of *K. pneumoniae* strains found was MDR. This is also well correlated with an earlier study (Loh *et al.*, 2007) carried out between the year of 2002-2007. In the study, they screened through of 1,581 cases of *K. pneumonia*e infections and found that 52.8% of the isolates were resistant to one class of antibiotics while 48.2% were to two classes of antibiotics. It was also noted that the numbers of resistant isolates increased throughout the year of research (Loh *et al.*, 2007).

Due to the rapid emerging of the resistant strain in the *Klebsiella sp*., determination of structure and function of hypothetical protein in *Klebsiella pneumoniae* may provide us an opportunity to find potential target for new antibiotic. The understanding of the structure of these hypothetical proteins might in turn be instrumental in the structure-based drug design strategy for discovering novel and effective antibiotics.

1.2.3 The Genome of Klebsiella pneumoniae

Complete genome sequence of *Klebsiella pneumoniae* was published and can be accessible in NCBI. It comprised a total of 5,315,120 million nucleotides and a total of 4894 coding genes. Out of that, 4,776 (about 85%) genes encode proteins. Further analysis showed that from the 4,776 protein coding genes, there are about 20% of the genes which are annotated poorly and are classified as hypothetical genes. In theory, these hypothetical genes (nucleic acid sequence) are eventually translated into proteins known as hypothetical proteins. It occupied a total number of 1004 protein of the 4776 protein (Table 1.1). Hypothetical protein deserved to be investigated in view of the fact that the hypothetical protein coded by quite a large percentage of genes in the genome of *K. pneumoniae,* and perhaps they might provide an important clue as what would be the best drug target for the bacteria.

Table 1.1 Distribution of all the hypothetical proteins from *Klebsiella pneumoniae* according to the number of amino acid residues.

Size (Number of amino acid residues)	Number of hypothetical protein
0-100	254
101-200	343
201-300	195
301-400	87
>400	125

1.2.4 Hypothetical proteins of Klebsiella pneumonaie and the importance

Approximately 20% of the *K. pneumoniae* protein coding genes are classified as hypothetical genes. Translation of these hypothetical genes into amino acid sequence will give rise to hypothetical proteins. However to date, there is no proper definitions for hypothetical proteins. In general, hypothetical proteins are predicted protein sequence which translated directly from nucleic acids sequences (Galperin, 2001; Lubec *et al.*, 2005; Pawlowski, 2008). The existence of these proteins is not shown in laboratory experiments. In some cases, these proteins have low identity to known annotated protein (Lubec *et al.*, 2005).

Functional characterization of the hypothetical protein(s) of *K. pneumoniae* using computational approaches is a great challenge and is quite difficult due to the fact that the presence of these hypothetical proteins in the organism is unknown. However it is worth attempting to predict hypothetical protein as it might give new protein motif or domain. In more opportune situation, one might also reveal new biochemical pathway or mechanisms which may influence our understanding in protein-protein interaction which is important in selecting proteins as drug targets.

1.2.5 Protein structure prediction

Most of the molecular mechanisms of the cells are realised by decoding the functions of the protein in an organism. Thousands of protein sequences have been determined over the years, and thousand of the associated protein structures have been resolved as well (Rose *et al.*, 2010). However, the experimental determination of the function of protein from known sequence still remains a challenging mission. Fortunately, there are number of computational techniques that can be exploited to assign function to experimentally uncharacterized proteins.

The experimental methods most commonly used to determine a protein's structure are x-ray crystallography and nuclear magnetic resonance (NMR) (Goodsell, 2010). In x-ray crystallography, scientists determine protein structure by measuring the directions and intensities of x-ray beams diffracted from high-quality crystals of a purified protein molecule. NMR uses high magnetic fields and radio-frequency pulses to manipulate the spin states of nuclei. The positions and intensities of the peaks on the resulting spectrum reflect the chemical environment and nucleic positions within the molecule. Scientists have been working to solve the protein-folding mystery for decades. In research that received the 1972 Nobel Prize in Chemistry (Anfinsen, 1973), Christian Anfinsen showed that a completely unfolded protein could fold spontaneously to its biologically active state, indicating that a sequence of amino acids contains all of the information needed to specify its 3D structure (Anfinsen, 1973). Promising results can be developed using both methodology (Kawamura *et al.*, 2011; Medina *et al.*, 2011; Hwang and Hilty, 2011; Sanders *et al.*, 2011). However, both methods are expensive and time consuming, and some proteins are not amenable to these techniques.

During the last decade, the integration for computational biology in protein research has become very essential. Bioinformatics tools have been widely used in predicting the structure of proteins and identifying their function homologue (Rigden, 2009). One of the goals of structural bioinformatics is to determine the three dimensional (3D) structure of all major protein families throughout the tree of life. Computer based 3D structure offer some advantage over experimental characterization: they are faster and less expensive. This will permit a deeper understanding of the relatedness of protein domain and its catalytic functions. In addition, it also enables us to identify the function to many proteins. Hence, to predict the hypothetical protein structure and functional characterization from primary sequence to a complete three dimensional structure point of view using computational methods remains one of the most popular and cost effective routine in structural bioinformatics (Nan *et al.*, 2009; Hernandez *et al.*, 2009; Hoskeri *et al.*, 2010).

1.2.5.1 Computational Protein Structure prediction

The prediction of three-dimensional structures of a protein from its primary sequence is a fundamental and well-studied area in structural bioinformatics (Sali and Kuriyan, 1999; Bourne and Weissig, 2003). There are three main directions in search of better structure prediction including homology or comparative modeling, fold recognition and *ab initio* prediction (Sali and Kuriyan, 1999) (Figure 1.2)(Table 1.2). In the first step of comparative modeling (which is also known as homology modeling), one of the several template proteins of high sequence similarities with the target is identified. This category of protein is known as high homology protein. Comparative modelling provides a great promise in protein structure prediction because small deviation in the amino acid sequence usually results in insignificant changes in term of its 3D structure (Chothia and Lesk, 1986; Marti-Renom *et al.*, 2000). On the other hand, if no unambiguous templates are found, fold recognition is attempted. Typically, the sequence-structure alignment (known also as threading) (Bowie *et al.*, 1991; Lemer *et al.*, 1995) is performed between the target and the template using both the sequence and structure information to identify the fold of which the target is most likely adopt.

Both approaches mentioned rely very much on similarity of sequence found on the target sequence and at least one known 3D protein structure. If no templates can be identified with confidence, *ab initio* methods are used to predict the target structure explicitly using templates sequence that do not align with the sequence of the template), as well as the details of side-chain positions (Zhang, 2008). This approach is aimed to predict the structure of protein on protein sequence alone with no similar amino acid sequence and it does not depend on any known protein structure. Although there are substantial progress seen in particularly in *ab initio* structure prediction (Koehl and Levitt, 1999), comparative modelling remains the most accurate method (Marti-Renom *et al.*, 2000). This approach can be applied to any proteins that have more than 40% sequence identity to the proteins with known structures in the PDB. Thus when a new protein sequence is found e.g. hypothetical protein in the *Klebsiella pneumoniae* which belong to a structure recognizable protein family, and 3D structures are already available for one or more members of that family, an atomic model can be built by comparison with those structures. There are many computer aided tools available in the web such as MODELLER (Sali and Blundell, 1993), SWISS-MODEL (Arnold *et al.*, 2006), and EsyPred3D (Lambert *et al.*, 2002) which manage to generate a reliable model prior to selection of proper template.

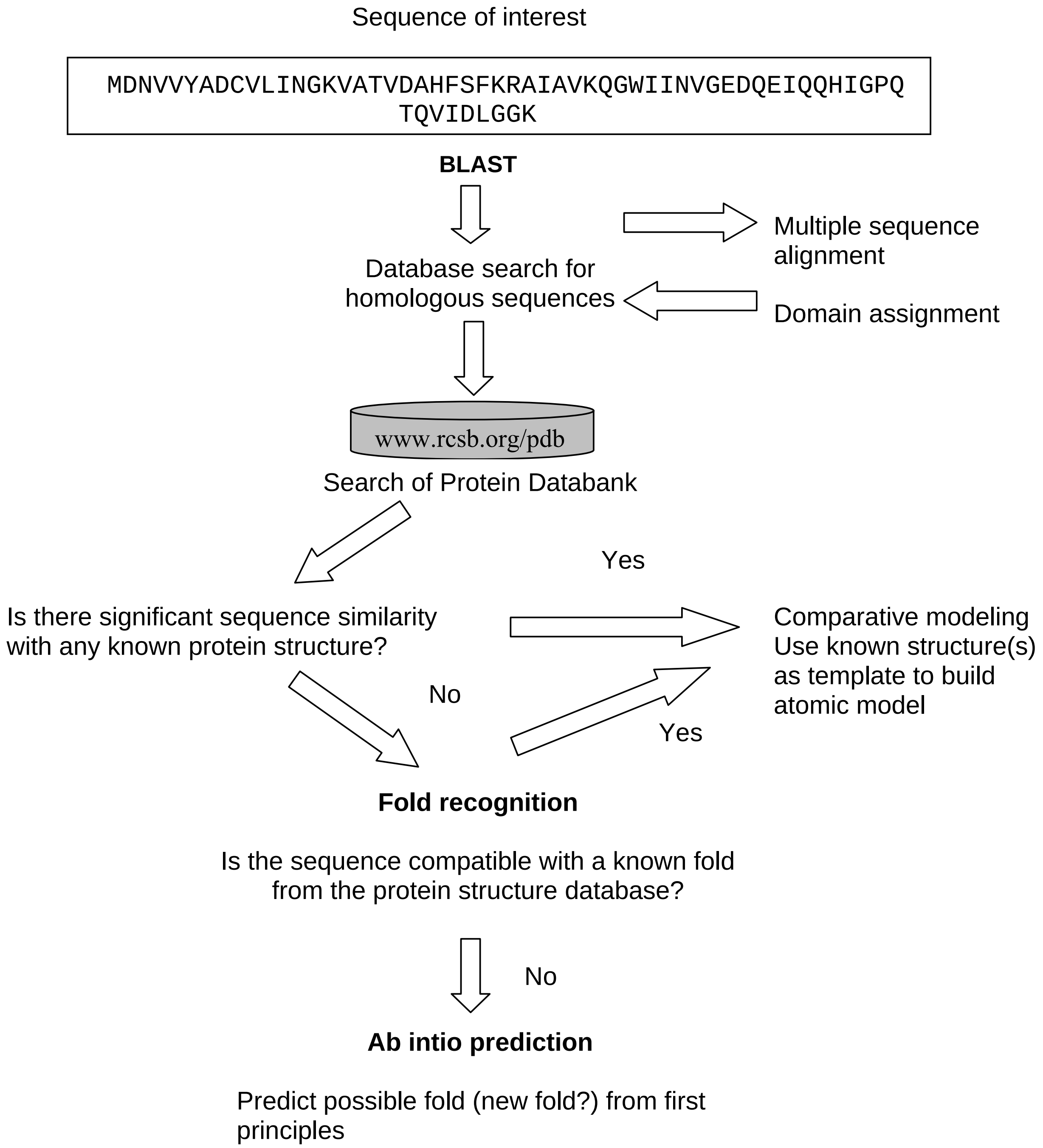
Sequence of interest
MDNVVYADCVLINGKVATVDAHFSFKRAIAVKQGWIINVGEDQEIQQHIGPQ
TQVIDLGGK
BLAST
Multiple sequence alignment
Database search for homologous sequences
Domain assignment
www.rcsb.org/pdb
Search of Protein Databank
Yes
Is there significant sequence similarity with any known protein structure?
Comparative modeling Use known structure(s) as template to build atomic model
No
Yes
Fold recognition
Is the sequence compatible with a known fold from the protein structure database?
No
Ab intio prediction
Predict possible fold (new fold?) from first principles

Figure 1.2 Flow chart showing the various steps and option for prediction of protein.

Table 1.2 Summary of the four main approaches to structure predictions. Note that there are overlaps between nearly al categories.

Method	% Sequence similarity	Knowledge	Approach	Difficulty	Usefulness	Accuracy
NMR, X-ray	-	Magnetic field and radio frequency , X-ray	Measure the directions and intensities of x-ray beams diffracted from high-quality crystals of a purified protein molecule. NMR uses high magnetic fields and radio-frequency pulses to manipulate the spin states of nuclei.	Medium	Very, for X-ray not all the protein can be crystallized. Crystallization of the protein may affect the conformation if the native protein. For NMR, the protein molecule must be a soluble protein and relatively small in size ~30 mg/ml.	High, ~ 1 Å
Comparative modeling (homology modeling)	More than 40	Protein of known structure	Identify related structure with sequence methods, copy 3D coordination and modify and where necessary.	Relatively Easy	Very, if sequence identity > 40% ⟶ drug design	High, ~ 1.5 Å
Fold recognition	Less than 30	Proteins of known structure	Same as above, but use more sophisticated methods to find related structure.	Medium	Limited due to poor models.	Medium, ~3.5 Å
ab initio tertiary structure prediction	Insignificant sequence similarity	Energy functions, statistics	Simulate folding, or generate lots of structures and try to pick the correct one.	Very hard	Not really.	Low, ~ 4-8 Å 80a.a,

1.2.5.2 Homology modelling

Homology modeling aims to predict the protein structures by exploiting the fact that evolutionarily related proteins with sequence similarity (Kaczanowski and Zielenkiewicz, 2010), as measured by the percentage of identical residues at each position based on an optimal structural superposition, share similar structures. Thus, if a new protein sequence is found (by sequence alignment) to belong to a recognizable protein family, and 3D structures are already available for one or more members of that family, an atomic model can be built by comparison with those structures (Tramontano and Morea, 2003). This approach can be applied to any proteins that have more than 40% sequence identity to the proteins with known structures in the PDB. In practice, the homology modeling is a multi-step process that can be summarized in seven steps:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model evaluation

1.2.5.3 Homology modeling by MODELLER

In this project, structure prediction using homology modeling approach was done using MODELLER 9. This software tool is developed by Andre Sali and coworker (Sali and Blundell, 1993). MODELLER is an automated tool adopting spatial restraint approach in homology modelling (Eswar *et al.*, 2007). Sequence alignment is the core process prior to the model building where the sequence alignment between the unknown sequences (target) with the known 3D structure (template) is aligned and used as the input of the program (Figure 1.3). Various types of restraints were calculated based on statistical analysis within a database that consist of 105 families with known 3D structure (Sali and Blundell, 1993). From the result of restraint analysis, these restraint conditions then transferred from the template to the target for 3D structure building. Combination of the restraints and CHARMM (Chemistry at HARvard Molecular Mechanics) (MacKerell *et al.*, 1998) energy as an objective function in the model which was generated by optimizing this particular objective function using conjugate gradient and simulated annealing algorithm. The selection of the best model can be ranked according to the discrete optimized potential energy (DOPE) function (Shen and Sali, 2006).

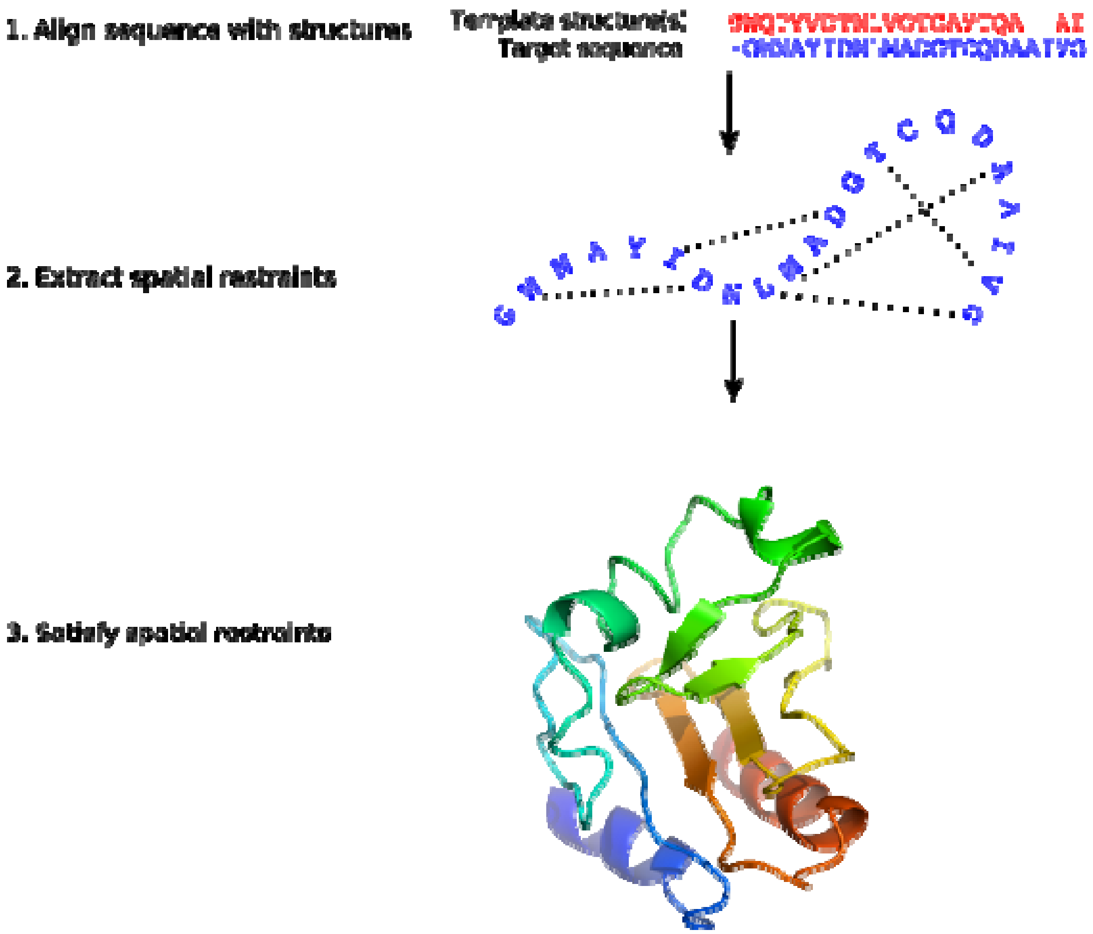
1. Align sequence with structures
Template structure(s)
Target sequence
2. Extract spatial restraints
3. Satisfy spatial restraints

Figure 1.3 A brief process flow of MODELLER automated homology modelling. (Adapted from http://salilab.org/modeller/9v7/manual/node11.html#2071, Date of accessed: 15 Dec 2011)(Sali, 1995)).

1.2.6 Protein function prediction

Functional determination of protein has become the major challenge for scientist recently with the rapid growth of the genomics data in the 20^{th} century (National Research Council, 2009). The main focus is on structural proteomics and how to analyze the protein structure using variety approaches such as computational or bioinformatics analysis. The function of a protein is very much dependent on how the proteins look like. Protein which has a similar structure usually adopts the same function (Brändén and Tooze, 1999). When the structure of the protein is being predicted correctly, with the built model from the prediction, we can use variety of computational approach such as molecular docking and molecular dynamics simulation to further probe or indicate the function of the protein. These two approaches were also used in this project to indicate the postulated function of the selected hypothetical protein from *Klebsiella pneumoniae*.

1.2.6.1 Molecular docking simulation

In the past few decades, computational approaches are used extensively to study the interaction of complexes. Generally, interaction between macromolecule (protein) and small molecule (ligand) can be studied using molecular docking method (Lengauer and Rarey, 1996). The interaction between macromolecule and the small molecules is very much depending on the physical forces and the chemical properties of each other. The binding of these molecules usually exhibit geometrical complimentary and this may also lead to the explanation of the activity or interaction. With the integration of an extensive searching algorithm, the geometrically and energetically best fitted ligand with the binding site of the protein can be determined using molecular docking simulation. Hence this approaches is frequently used to predict the binding affinity which play an important role in drug design (Kitchen *et al.*, 2004).

A large number of molecular docking tools have been developed due to the rapid emerging research in bioinformatics field. The most commonly being used and discussed are DOCK (Ewing and Kuntz, 1997), GOLD (Jones *et al.*, 1997) and Autodock (Morris *et al.*, 1998). DOCK program is developed by University of San Francisco in 1997 and it employed rigid body assumption with graph theoretical searching technique. It is usually used for screening of large database of ligand as it is less expensive computationally (due to the fact that both protein and ligand are treated as rigid body). As for Autodock and GOLD, both tools allowed more flexibility as compared to DOCK. Different variant of genetic algorithm are used in both softwares which enable full range of ligand conformation flexibility of protein and also the ligand. This enhancement in term of flexibility of protein and ligand in the program is one step closer to the fundamental requirement that ligand and protein are bound in the water which allows tremendous flexibility in their binding mode.

Autodock 3.0.5 (Morris *et al.*, 1998) was used in this study. Flexibility of the molecules can be achieved due to randomization on the torsion angle, which is done by exploring translation, rotations and its internal degree of freedom of the ligand. This will lead to the favourable conformation in its binding mode. Lamarkian genetic algorithm scoring function incorporated with Solis and Wet search procedure in this version showed better handling in large ligand and higher accuracy as compared to the previous version.

Docking simulations enabled us to understand the preferable conformation of ligand in the binding mode to form a stable complex but there are limitations. In docking simulation, rigidity of protein and target of docking location is defined by the user. Hence this decreases the degree of freedom of both interacting component during the simulation. Furthermore, results from docking can only provide a single snapshot of the ligand orientation which is lacking in interaction dynamics. Therefore, another more powerful computer simulation technique, namely molecular dynamics was employed in this research to obtain an in-depth understanding about the predicted hypothetical protein structure and function.

1.2.6.2 Molecular dynamics simulation

The dynamics nature of the protein and ligand lead us to further investigate the structural and functional properties using molecular dynamics simulation. Molecular dynamics (MD) simulation is a well-established method for modeling. It provides insight into biomolecular systems in particularly the interaction properties, understanding of protein folding, and interactions specifically in phospholipids membrane bilayer that are difficult to access experimentally (Karplus and McCammon, 2002).

Forty years ago, McCammon and co-worker performed the very first molecular dynamics simulation (McCammon *et al.*, 1977). It was on a small globular protein and the total simulation time is less than 10 ps. Over the years, many promising development on software and hardware enhancement enable us to perform a longer time scale simulation which allows us to gain a better insight. Many different software tools available such as AMBER (Pearlman *et al.*, 1995), CHARMM

(Brooks *et al.*, 1983), NAMD (Phillips *et al.*, 2005), GROMACS (Lindahl *et al.*, 2001). Different variant of stochastic dynamics integration are used in most of these tools. In this study, GROMACS 4.0.5 (Groningen Machine of Chemical Simulation) (Van Der Spoel *et al.*, 2005) a free and efficient software to perform energy minimization and molecular dynamics simulation was used. It is developed by University of Groningen and designed primarily for biological systems such as lipid, protein and nuclei acids. To set up a MD simulation in general, a set of coordinates of the starting structures, information about the interaction such as bonding, torsion and angle, last but not least the MD simulation parameters are needed. The process flow of typical MD run using GROMACS with protein in a box of water is shown in Figure 1.4. GROMACS is commonly use especially with membrane system; this is also the reason that this program is selected to aid the validation.

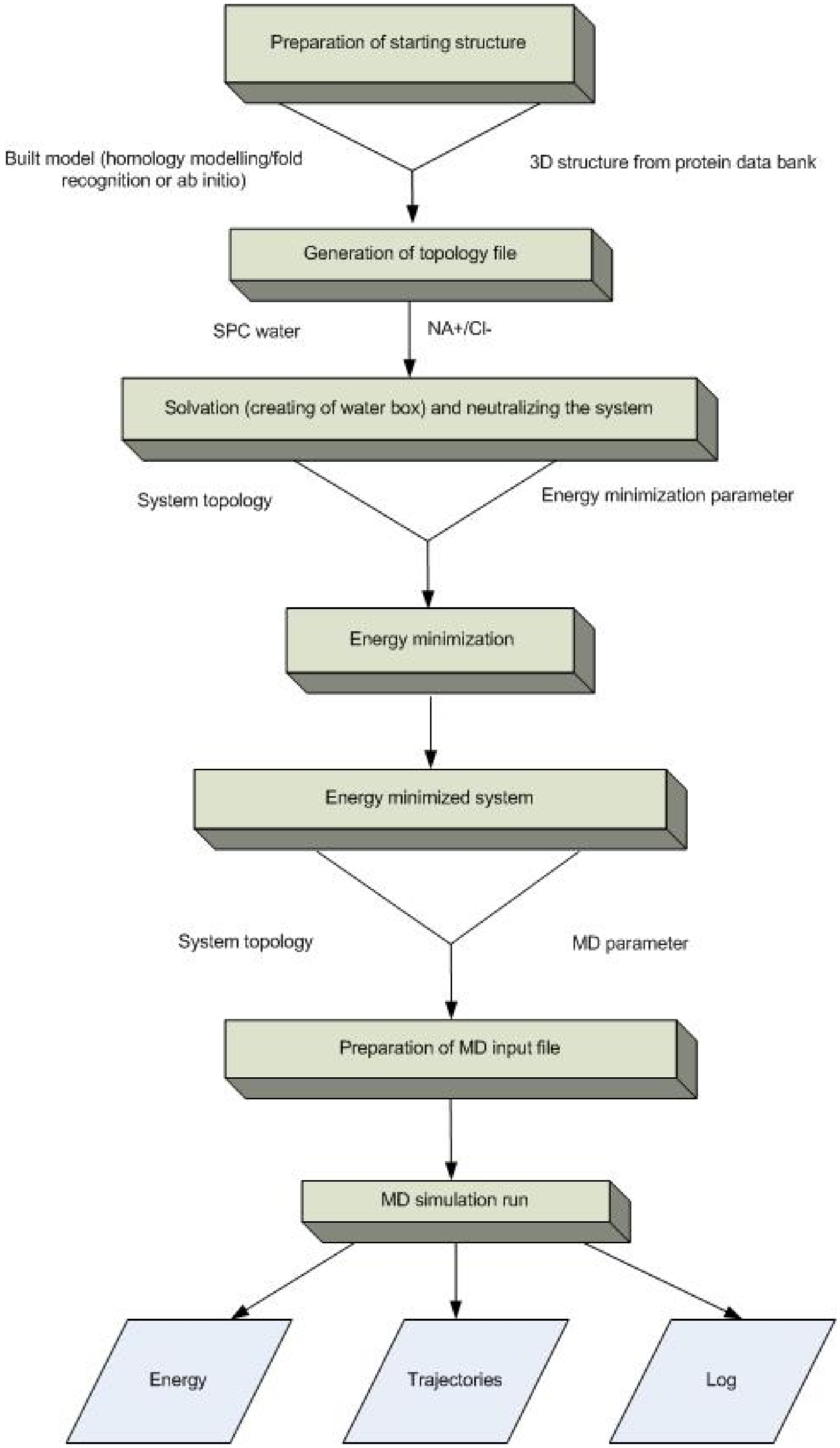
Preparation of starting structure
Built model (homology modelling/fold recognition or ab initio)
3D structure from protein data bank
Generation of topology file
SPC water
NA+/Cl-
Solvation (creating of water box) and neutralizing the system
System topology
Energy minimization parameter
Energy minimization
Energy minimized system
System topology
MD parameter
Preparation of MD input file
MD simulation run
Energy
Trajectories
Log

Figure 1.4 Process flow of a general set-up of the molecular dynamics simulation system aided with GROMACS.