# CHAPTER II
# LITERATURE REVIEW

## *Mycoplasma pneumoniae* and its role extrapulmonary disease

*M. pneumoniae* is one of the simplest self-replicating bacterium capable of cell-free existence, with genome as small as 0.86 Mb and cell wall-less. It is spread both by direct contact between an infected person and a susceptible person and by airborne droplets expelled when an infected person sneezes, coughs, or talks. Disease rates are especially high in young children (Razin *et al.*, 1998; Dandeker *et al.*, 2000; Kannan *et al.*, 2005). The lack of cell wall of *Mycoplasma* is used to define this organism from ordinary bacteria and to include them in a separate class named *Mollicutes*. *M. pneumoniae* are usually living as host-associated parasite. This organism multiplies within host and persistence for long period of time (Biberfeld and Biberfeld, 1970; Wall *et al.*, 1983). Serious infections requiring hospitalization occur in both young adults and children (Rastawicki *et al.*, 1998; Wattanathum *et al.*, 2003; Bitnun *et al.*, 2001). This pathogen infects not only pulmonary but also multiple organ systems. Extrapulmonary complications can occur in association with *M. pneumoniae* infection as a result of direct invasion especially central nervous system (CNS) as encephalitis (Bitnun *et al.*, 2001; Sotgio *et al.*, 2003; Coelho *et al.*, 2004). Encephalitis manifestations are greater severity and clinical importance than the primary respiratory infection (Bitnun *et al.*, 2001). It has been known that *M. pneumonaie* is assigned to be a surface parasite by that remain attached to the surface of either endothelial or epithelial cells that observed by scanning electron microscope (Rottem *et al.*, 2003). In addition this organism has advance mechanisms to enter host cells that are not physically phagocytic (Yavlovich *et al.*, 2004). There have been reports on binding to plasminogen of *M. fermentans* and *M. pneumoniae* (Yavlovich *et al.*, 2004). In case of *M. fermentans*, the binding facilitates the invasion of HeLa cells by the bacterium, suggesting that the ability of the organism to invade host cells. An increase in the invasive capacity of *M. fermentans*, which arise from the potential of this organism binding to zymogen plasminogen and activated by urokinase to plasmin, has been recently observed (Yavlovich *et al.*, 2004).

## *Mycoplasma* virulence factors

It has been reported that many microorganisms are able to adhere to host tissues for colonization and infection including *M. pneumoniae* (Seto *et al.*, 2003; Shin *et al.*, 2005). The *Mycoplasma* organisms enter an appropriate host in which they multiply and survive for long period of time. Several infection of *Mycoplasma* depends on adhesion to host tissues for colonization and infection. This adherence is the major virulence factor and adherence-deficient mutant are avirulent (Dallo *et al.*, 1990; Dallo *et al.*, 1996). Most study of adherence systems are those of *M. pneumoniae*, the causative agent of primary atypical pneumonia in human. Adhesin and its receptor system are extensively studied with respect to *M. pneumoniae*. Adhesin, a surface 169-kDa protein designated P1 is densely clustered at the tip organelle of virulent *M. pneumoniae* (Wall *et al.*, 1983; Dallo *et al.*, 1988; Seto *et al.*, 2001). The P1 protein localizes at the tip organelle is considered to be important for the attachment, providing a critical concentration for primary association with receptor molecule on the host cell (Dorigo-Zetsma *et al* 2001; Layh-Schmitt *et al.*, 2000). This finding suggests a possible role of P1 as a virulence factor that facilitates invasion of *M. pneumoniae* through hBBB by initiating bacterial binding to host cell membrane, which might further generate an uptake signal induced invasion into cytoplasm. In summary, the most bacterial pathogens express surface factors that either mediates direct binding to host cells or indirectly attach host adhesion factors (Niemann *et al.*, 2004). However in the revision of bacterial invasion, it is essential to differentiate between microorganisms adhering to a host cell and that with have penetrated the cell. It is to be expected that surface molecules either proteins or lipids that might well assist the adhesion process. Invasion is correlated with adhesin and its receptors on host cell surface that mediate interaction of the microorganism with the host cell. Nevertheless, adherence to the surface of host cell is not sufficient to elicit events that lead to invasion. The most recent study microorganism invasion is based on the ability of several organisms to bind plasminogen (Fox *et al.*, 2001; Jong *et al.*, 2003). Plasminogen binding activity was detected in *M. fermentans* with increase in the invasive capacity of this organism. The invasive process arises from the prospective of this organism to bind plasminogen and activated by urokinase plasminogen activator, to plasmin, has been recently described. The increase in the

invasive capacity of *M. fermentans*, which take place by plasmin, a protease with broad substrate activity and by this means promotes its invasion (Yavlovich *et al.*, 2004).

Furthermore another potential virulence factor is α-enolase, a surface protein on *M. pneumoniae*. Because, it was reported that α-enolase has a role as cell-surface plasminogen-binding site (Miles *et al.*, 1991; Redlitz *et al.*, 1995; Pancholi *et al.*, 1998; Ehinger *et al.*, 2004). Recent studies have shown an increase in the invasive capacity of microorganism, which arises from α-enolase binding to human plasminogen kringle domain 2 (Rios-Steiner *et al.*,2001; Jong *et al.*, 2001; Jong *et al.*, 2003). Alpha enolase has been reported as an important mediator of tissue pathology in infectious disease (Redlitz *et al.*, 1995). This enzyme, with 45 kDa, is basically cytoplasmic enzyme in glycolytic pathway, an important step in the process of ATP production, by means of the reversible conversion of 2-phosphoglycerate into phosphoenolpyruvate in the presence of $Mg^{2+}$ (Redlitz *et al.*, 1995; Paist *et al.*, 2005). Remarkably, the enolase gene is not a housekeeping gene because its expression and varies according to the pathophysiological (McAlister *et al.*, 1982). Enolase was found as a strong plasminogen-receptor by its expression on the surface of variety of both eukaryotic and prokaryotic cells (Miles *et al.*, 1991; Nakajima *et al.*, 1994; Pancholi *et al.*, 1998). Group A *Streptococci* was the first reported organism with the presence of surface enolase for prokaryote with demonstrated function of its binding to plasminogen (Pancholi *et al.*, 1998). The localization of α-enolase on cell surface of several *Streptococcus* species has been recently identified to be a virulence factor (Ehinger *et al.*, 2004; Ge *et al.*, 2004). Moreover other host invasive bacteria such as *Candida albican* and *Streptococcus* species were also found the ability of α-enolase reassociate to surface membrane and complex formation with human plasminogen which then bound to human brain microvascular endothelial cell (HBMEC) (Jong *et al.*, 2003; Pancholi *et al.*, 1998; Rios-Steiner *et al.*, 2001; Bergmann *et al.*, 2001). However the mechanism of enolase reassociates to cell membrane is remaining ambiguous. It has been known that most bacterial pathogens express surface factors that either mediate direct binding to host cells or indirectly attach host adhesion factors (Niemann *et al.*, 2004). Microorganism binds to host cell would facilitate cell invasion. The increasing ability of organisms to cross an in *vitro* blood brain barrier is facilitated by binding with plasminogen in the presence of either tissue plasminogen

activator or urokinase plasminogen activator to plasmin, has been recently reported (Jong *et al.*, 2003; Yavlovich *et al.*, 2001).

Plasminogen is one of the components of proteolytic or fibrinolytic system that binds to variety of blood cells (Herren *et al.*, 2001). Plasmin is a serine protease with broad substrate specificity that behaves to alter organism-cell surface proteins and thereby promotes its invasion. For instance, *Yersinia pestis* plasminogen was activated by its plasminogen activator which then leads to degrade bacterial outer membrane proteins triggering virulence (Sodeinde *et al.*, 1992). Notably the U937 monocytoid cell was demonstrated as a model system expressing α-enolase as a candidate plasminogen receptor on nucleated blood cells (Miles *et al.*, 1991). Remarkably, enolase is presently known to be exposed to the surface of hematopoietic cells such as monocytes, T cells and B cells, neuronal cells, and endothelial cells (Miles *et al.*, 1991; Dudanai *et al.*, 1993; Redlitz *et al.*, 1995). Plasminogen is usually containing kringle domain which recognizes carboxy-terminal lysine (Rios-Steiner *et al.*, 2001; Kim *et al.*, 2003; Derbise *et al.*, 2004). Therefore, cell surface protein containing carboxy-terminal lysyl residues might play a functional role of plasminogen receptor. Accordingly, α-enolase possesses a naturally existing carboxy-terminal lysine, a common structure of plasminogen receptor. Recent study has been reported that α-enolase is a strong plasminogen binding protein (Pancholi and Fischetti, 1998). Moreover, α-enolase-plasminogen complex have clearly shown that receptor-bound plasminogen is more readily activated to plasmin than free plasminogen (Miles *et al.*, 1991; Redlitz *et al.*, 1995; Fox and Smulian, 2001). In addition, host plasminogen is also bound with microorganism enolase by means of interlinking molecule to facilitate invasion of organism into HBMEC (Jong *et al.*, 2003). Accordingly, the complex of host plasminogen binding microorganism enolase is considered as a possible, important virulence characteristic which may contribute to tissue invasion. Moreover, α-enolase has been identified on a surface area of *Streptococcusl* species (Bergmann *et al.*, 2001; Derbise *et al.*, 2004) and the fungal pathogen *Candida albican* (Jong *et al.*, 2001; Jong *et al.*, 2003) which involves in penetration through the brain microvascular endothelial cells (BMEC). The existence of enolase on *M. pneumoniae* genome may perhaps play a possible role of interaction with plasminogen which might increase the ability of *M. pneumoniae* to penetrate across hBBB. In addition, *M. pneumoniae* enolase (MpnE) is a protein

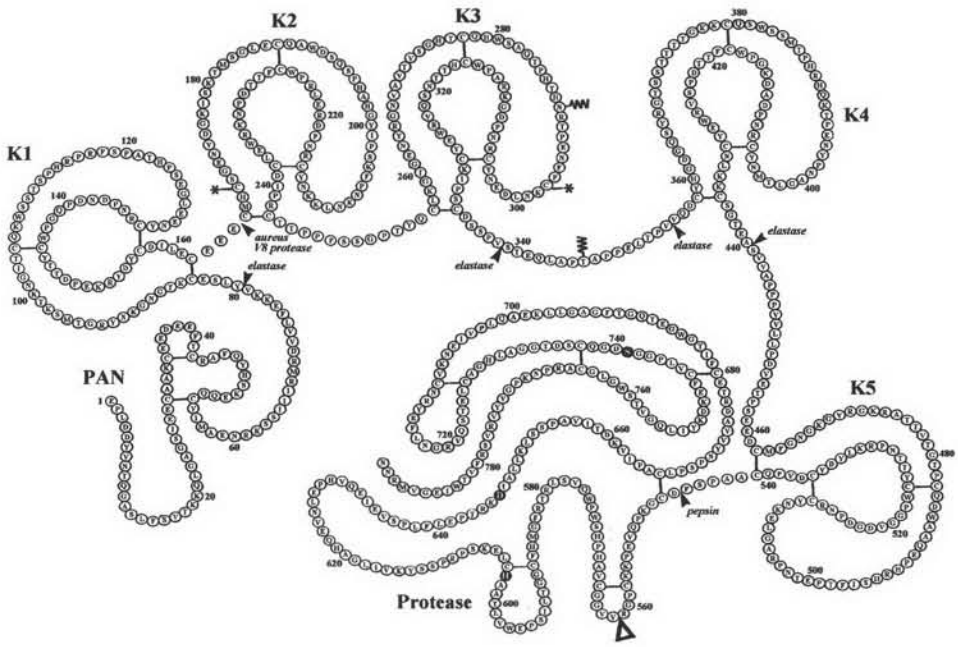
with significant similarity to known structure in PDB reported by Berkeley Structural Genomics Center (BSGC) at http://www.strgen.org/3/05/06.

## Human plasminogen

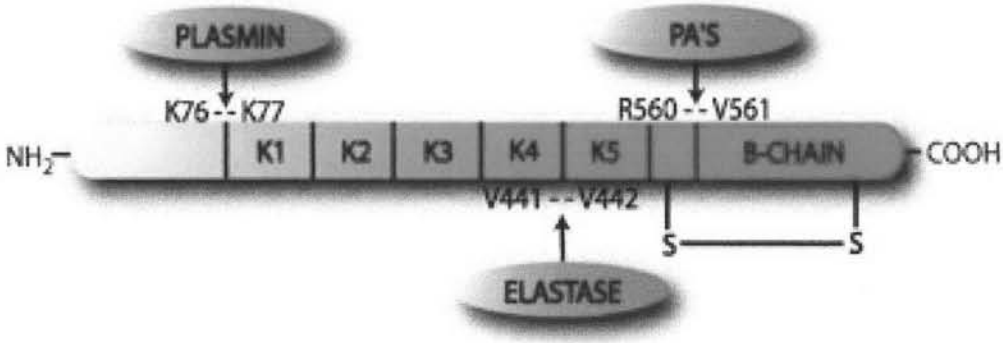
Human plasminogen is composed of seven independently folded domains. At the N-terminus is a growth factor-like domain, followed by five kringle domains and C-terminal trypsin like catalytic domains and a serine protease domain (Figure 1). The Glu1-Val79 N-terminal domain of human plasminogen (Plg) is followed by a tandem array of five kringle domain of 9 kDa of each. K1, K2, K4, and K5 each contains lysine binding site (LBS). The kringle domains provide fibrin binding sites. In the absence of fibrin, plasminogen adopts a "closed" conformation, as a consequence of intra-molecular interactions. It is converted to an "open" conformation when bound to fibrin and consequently has a much enhanced activation rate. Thus, domain interactions control overall conformation of plasminogen, which in turn provides the site-specificity of activation. The structure of human plasmin catalytic domain has been determined in complexes with streptokinase (Wang *et al.*, 1998). C-terminal domain of streptokinase is assumed (Wang, *et al.*, 1998) to activate the catalytic device of the bound plasminogen in an activator complex.

The structural features associated with the transition from a zymogen to an active serine protease are well known owing to crystal structures of zymogen/ enzyme pairs. Most serine protease zymogens have in their catalytic units an activation loop, where upon activation, a peptide bond is proteolytically cleaved ensuing in conformational changes, including the formation of a new solvent-inaccessible salt-bridge bond and consequently the formation of an active site. Such a structural restraint on the site of proteolytic activation cleavage is shown to bring about unique properties to plasminogen activation. Plasminogen activation is also known as a pivotal role for many physiological and pathological processes involved pathogen invasion (Coleman *et al.*, 1999; Jong *et al.*, 2003; Leytus *et al.*, 1981; Monroy *et al.*, 2000; Pancholi *et al.*, 1998). Under physiological conditions, plasminogen is activated by limited proteolysis mediated by two highly specific proteases, tissue-type plasminogen activator (tPA) and urokinase plasminogen activator (uPA) (Gething *et al.*, 1988; Gong *et al.*, 1998; Leytus *et al.*, 1981). Plasmin, plasminogen active form,

**(A)**




**(B)**




http://www.haemtech.com/Enzymes/Plasmin.htm/27/07/06

**Figure 1** Primary structure of human plasminogen (Rajante and Llinas, 1994) (A) Human plasminogen contains 7 main domains. (B)The first 5 kringle domains domain structure of human plasminogen is represented K1-K5. B-CHAIN is catalytic domain of plasmin, and the arrows indicate the sites, a single arginine-valine bond, of proteolytic cleavage (Gong *et al.*, 2001) by plasminogen activators (PA'S).

Plasmin, plasminogen active form, has broad specificity but can not directly catalyze the proteolytic activation of plasminogen. Native tissue plasminogen activator converts the zymogen plasminogen to the active form, plasmin (Gething *et al.*, 1988). In order to the invasive scheme, some pathogen produces plasminogen activators (Wang *et al*, 1998) that elicit plasmin production and proteolysis which then penetrate the host tissue.

## Bioinformatics approach for identifying virulence factor

Bioinformatics and computational methods have become a potential approach for life science research in genomic era. With the explosion of sequence and structural information available to public domains, the field of bioinformatics is playing and increasing considerable role in the study of fundamental biological problems. Moreover, the advent of bioinformatics in conjunction with genomic sequence conjunction with genomics sequences information of pathogenic organisms provide the conception of new prophylactic and therapeutic interventions and offers also great potential in term of drug development. Currently, genomes of more than two hundred bacterial strains have already been sequenced and available including that of *M. pneumoniae*. Therefore, it is possible to apply bioinformatics techniques to hasten the hunt for virulence factor identification this research work. The obvious advantage of using computer-aided screening technique to scan putative virulence factor is beginning with virulence factor coding protein; adhesin, invasin, toxins, and protein secretion system (Chen *et al.*, 2004), computational tools can be used to identify them from *M. pneumoniae* genome sequence. It has been known that, in the pathogenic strains, some of the G/C content carries virulence genes which code for adhesins or other virulence factors. Based on the knowledge that these genes are located on Pathogenicity Islands (*PAIs*) in the genome (Yoon *et al.*, 2005), *PAIs* of *M. pneumoniae* genome must be identified. *PAIs* is a part of genomic region with atypical G+C content, bias of codon usage pattern, carriage of mobile sequence element, and/or association with *tRNA* genes (Mantri and William, 2004). Some computational tools for *PAIs* identification have been available. *IslandPath* (Hsiao *et al.*, 2003) is a web based tool for identifying genomic islands in prokaryote genomes.

An advantage of this tool is that it is a single computational application for finding genomic islands of interest and integrating multiple features for island detection such as *PAIs* region based on known islands from literature survey in full genome context. Besides, it has an advantage over other type of analyses as it can complement multiple DNA signal analyses with additional annotation feature. In this study, *IslandPath* has been selected to detect *PAIs* in *M. pneumoniae* genome. The genes located on *PAIs* obtained from *IslandPath* will be screened and selected for further study.

## Biological sequence analysis

Because *M. pneumoniae* genome has not been completely annotated and functions of many genes in the genome are still unknown, the endeavor to describe function of un-annotate genes that found in *PAIs* is required. Therefore the sequence analysis tools were reviewed as following.

The sequence analysis tools are based on physicochemical property to compute the hydrophobicity of amino acid. Then the profile of hydrophobicity can be used to predict an interaction site on the surface of globular protein. Because membrane protein structures are difficult to perform from experimental approach, thus the prediction of transmembrane regions provides a very valuable tool to suggest the possible function of membrane proteins in undesirable genes. The preferred genomics sequence annotation tools that used in prokaryote are *PRINTS, BLOCKS*, and *ProDom* which were created to integrate documentation resource of protein family, domain and functional site, the major protein signature database by InterPro web server (Bru *et al.*, 2004; Möller *et al.*, 2001; Mulder *et al.*, 2005). Most databases are derived from comparisons of either microbial genomes or protein sequences. Comparative genomics allocate the similarities between known sequences and a new dataset to be analyzed, and classify families of related protein. In addition, the usages of several web servers to detect the signal site are based on the special patterns at the boundary site in signaling protein. *SignalP* web server is widely used to predict signal peptides in secretory protein by special pattern of cleavage site (Weinstock, 2000). The *SOSUI* web server provides specialize secondary structure prediction particularly on membrane proteins with high accuracy (Möller *et al.*, 2001). The evaluation of method for prediction of membrane region TMHMM is currently

the best program to perform transmembrane identification (Möller *et al.*, 2001). However, up to now, the statistical significance of a match from each tool is frequently low (Mulder *et al.*, 2005). Therefore, the search results should be used as a guide for protein characteristic feasibility. After the virulence factor is characterized by primary sequence analysis, it will be undertaken to study in detail of its 3D-structure. The most advanced tool for generating 3D-structure by sequence itself is homology or comparative modeling approach that will be described in detail as following.

## Comparative modeling

It has been known that experimental methods to determine protein structure are difficult, time consuming and requiring expensive equipment. Determining a protein structure experimentally using nuclear magnetic resonance (NMR) or X-ray crystallography method can not be applied to every protein since many of them are difficult to be crystallized or too large for NMR studies. Besides, these attempts are not always successful in case of membrane proteins due to the intrinsic difficulties involved in growing crystals of membrane proteins. Therefore, so many scientists are interested in finding an algorithm, a method, to predict the native structure of protein given just its sequence. The most advance strategies for protein structure prediction is comparative modeling. Comparative modeling, homology modeling, is by far the most reliable scheme to predict 3D protein structures. By comparative modeling the structure of interesting protein is conducted by comparing its sequence with the sequences of known structural protein (Marti-Renom *et al.*, 2002). If high sequence identity is found then it can be assumed that the proteins have similar overall structures. On the other hand, if no strong identities have found the comparative modeling can not be employed. The assumption is that proteins with identity sequences have almost similar structure (Šali and Blundell, 1993; Eswar *et al.*, 2003). It has been noted that two proteins have the same fold if they display amino sequence identity above 30% (Marti-Renom *et al.*, 2000; McGuffin and Jones, 2003). The most reliable tool of this approach is validated by Critical Assessment in Structure Prediction (CASP) competition at http://predictioncenter.llnl.gov. This experiment is carried out every 2 years, this present year, 2006, is CASP7 event. By CASP,

crystallographer and NMR spectroscopists are available to the prediction community the sequences of as yet unpublished structures. Each group of research endeavors is to predict these protein structures and deposit their predictions before the structure are revealed to publish. A few months further of this experiment, all the predictions are collated and validated by a number of independent assessors. By this scheme, the present state of art in protein structure prediction can be certified without any possibility of either bias or cheating. MODELLER is the most preferred comparative modeling package that has been shown to do consistently well by CASP experiment. MODELLER uses a simulated annealing approach (Šali and Blundell, 1993) that freely available for academic use. The initial stage to conduct model is identified one or more template and generating an alignment between template and the target sequence. This package contains single step to assemble structurally conserved regions (SCRs) and structural variable region (SVRs) by MD approach. SCRs are known as core region by means of conserve region with template. SVRs are the structural regions that differ from template. A conventional MD force-field is used and Newton's motion laws are also integrated in package. Moreover, additional spatial-restraints are imposed in the force-field known as probability density function (PDF). This PDF is weighted such that regions those are mostly conserved in structure have stronger restraints whereas region that vary more in structure have weaker restraints. The best structural model from MODELLER is selected by lowest objective function.

The preparatory point to model protein structure by comparative modeling is to categorize all protein structures related to the target sequence, which then be chosen those that will be employed as template. This step begins by finding all sequences clearly related to the sequence to performed the target sequence profile. Additionally, the similar profiles are constructed for entire known protein structures. THREADER, Phyre, and 3D-PSSM are web server that used to determine the optimal template for comparative modeling. Probable of templates are then hit upon comparison the target sequence profile which each of the sequence profile for known structure. The multiple sequence scheme for fold identification are most effective fully-automated approaches to identify template when the sequence identity between the target and template less than 25% (Marti-Ranom *et al.*, 2000). The second class of methods is threading or 3D-template matching methods. This method relies on

pair-wise comparison of a protein sequence and a known structure protein. The target sequence is threaded through a library of 3D folds. This procedure is particularly useful when there are no sequences clearly related to the modeling target, and hence the search can not benefit from the increased sensitivity of the sequence profile scheme. Consequently, a list of all related protein structures has been achieved, it is crucial to select those templates. Regularly the highest overall sequence identity between the target and the template sequence yields the most favor template (Sanchez and Sali, 1997).

## Target-template alignment

Once the suit template has been selected, an alignment method will bring to align the target sequence with the template structure. For closely related protein sequences with identity above 40%, the alignment is almost constantly correct. On the other hand the alignment becomes tricky in the "twilight zone" when the overall sequence identity is below 30% (Sanchez and Sali, 1997). By the sequence identity decreases, alignments contain an increasingly huge number of gaps and error of alignments. Highest endeavor to perform the most accurate alignment is needed because no current comparative modeling method can recover from incorrect alignment. However, the information from structures helps to avoid gaps in secondary structure element, in buried regions, or between two residues that are far in space. Secondary structure predictions for the target sequence and its profile are also regularly useful to perform a more accurate alignment (Aloy *et al.*, 2000). However, evaluating the corresponding models and picking the best model according to the 3D-model evaluation is more reliable than the alignment score (Sanchez and Sali, 1997).

## Model building

Once an initial target-template alignment has been built; a variety of methods can be used to generate a 3D model for target protein. The original method is modeling by rigid-body assembly (Greer, 1990). Another method uses segment matching for modeling relies on an approximate position of conserved atoms in the templates (Claessens *et al.*, 1989; Jones and Thirub, 1986). The third group is the

most advanced method, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints performed by alignment (Šali and Blundell, 1993).

## Model building by satisfaction of spatial restraints

This method produces many constraints or restraints on the structure of the target sequence, by means of the guidance of its alignment to related protein structure. The restraints are usually acquired by assuming that the corresponding distances and angles between aligned residues from template to target structure are similar. These comparative derived restraints are generally complemented by stereo-chemical restraints on bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts derived from a molecular mechanics force field. The model is further optimized by minimizing the violations of all the restraints. This can be achieved either by distance geometry or real-space optimization. A geometry distance approach constructs all-atom models between lower and upper bounds on distances and dihedral angles (Havel and Snow, 1991). A real space optimization method, which implement in MODELLER (Šali and Blundell, 1993) initiates by generating the model using the distance and dihedral angle restraints on the target sequence obtained from its alignment with template 3D structures. Afterwards the spatial restraints and the CHARMM22 force field terms that enforce proper stereo-chemistry (Mackerell *et al.*, 2004; Patel *et al.*, 2004) are combined into an objective function. Ultimately the model is constructed by optimizing the objective function in Cartesian space. Outstandingly the satisfaction of spatial restraints can use several diverse types of information about the target sequence; it is the strongest promising of all comparative modeling schemes.

## Loop modeling

An accuracy of loop modeling is a key factor determining the usefulness of comparative models in investigating protein-protein interaction. The prediction of loop modeling by optimization is theoretically applicable to simultaneous modeling of several loops. A variety of optimizing based on unified atom, all non-hydrogen

atoms, non hydrogen and 'polar' hydrogen atoms. The optimal degrees of freedom include Cartesian coordinates and internal coordinates, for instance dihedral angles, optimize in continuous or discrete spaces. Loop prediction by database search consists of finding the segment of main chain that compatible with two stem region of a loop. The search is generated throughout variety known protein databases. Typically, different alternative segments that fit the main stem residues are performed and probability sorted according to geometric criteria or sequence similarity from template to target loop sequences. The top quality of segment are selected and further superimposed and annealed on the stem regions. Then an automated energy optimization is generated to refine the initial crude model. Nevertheless, the database searches are edged by the exponential increase in the number of geometrically possible conformations as a function of loop length. Referring to only segment of 7 residues or less has most of their conceivable conformation exist in known structure databases (Fidelis *et al.*, 1994).

## Model quality evaluation

Reliable computational methods for 3D-structural prediction are important tools as they provide the basis for further experimental analysis. Consequently, a 3D-structure prediction requires an accuracy method. Model evaluation is important to validate 3D model obtained from molecular modeling. The different types of errors can occur due to backbone connectivity error misalignments or mis-registrations of residues, and misplacement of side chain. Several methods of model evaluation have been developed to assess stereo-chemical quality. PROCHECK is the one of validation tool to check the quality of the conformations of the polypeptide backbone and side chains. Most largely mis-folded structures can be identified in this tool (Pontius *et al.*, 1996). VERIFY3D is a tool for assessment of the environment of the side chain (Kirton, Baxter, and Sutcliffe, 2002). ERRAT is a another tool which determine the compatibility between the amino acid sequence and the environment of the amino acid side chains in the model by assessing the distribution of different types of atoms with respect to one another in the protein models. ERRAT is also a protein structure verification algorithm which is especially well-suited for assessing the progress of crystallographic model building and refinement. The program works by

analyzing the statistics of non-bonded interactions between different atom types. Through the comparison with statistics from highly refined structures, the error values have been calibrated to provide confidence limits. This is extremely useful in making decisions about reliability (Colovos and Yeates, 1993).

## Molecular docking

The determination of protein-protein interaction is important in understanding biological processes at the molecular level. By knowledge of 3D protein structure permits the feasible of intervention and manipulation of molecular interaction via structure-based drug design, and protein engineering. Therefore molecular docking method becomes important tool to simulate potential protein-protein complex (Smith *et al.*, 2002). Computational docking approach is combination of rigid body and torsion angle dynamics (Wojcik *et al.*, 2001; Wang *et al.*, 2005). A protein-protein complex usually has a function consequence as signal transduction and also be responsible to develop pathogenesis process. Macromolecular complexes are essential for a mechanistic description and understanding of cellular process (Russell *et al.*, 2004). Therefore, an investigation of protein-protein interactions is significant because activities of biological systems depend on the specific recognition of proteins. The general protein-protein docking procedure can be made as following. Firstly, a simplified description of the protein, protein surface, is used instead of atomic level detail. Secondly, some form of simple surface area complementarily measure is used to score the 'fitness' of different solution complex. Thirdly, both of the protein molecules are basically consider rigid. Ultimately, search problem is restricted to 6 degrees of freedom, three translations and three rotations (Gray *et al.*, 2003). Rationally, the result from theoretical study of protein-protein interaction can be used to describe physicochemical of molecular interaction level in detail by means of free energy binding, decreasing of accessible surface area. However, the fact that protein complex is required an interaction combination with an all atom flexibility of both partner by further refinement with molecular dynamics simulation (Fan and Mark, 2004).

## Refinement model by molecular dynamics simulation

This computational scheme calculates the time dependent behavior of a molecular system. MD simulations have provided detailed information on the fluctuations and conformational changes of nucleic acids including macromolecule as proteins. These methods are at this moment routinely used to scrutinize the structure, dynamics and thermodynamics of biological molecules and their complexes. They are also used to determine the structures from x-ray crystallography and from NMR experiments (Novotny and Sippl, 1997).

## Application of molecular dynamics simulation

MD simulations are developed to revise complex, dynamic processes that occur in biological systems. These include, for example, stability of protein, conformational changes, protein folding and also provide the mean to carry out studies on drug design, and structure determination based on X-ray and NMR (McCammon, *et al*, 1977). The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) (McCammon, *et al*, 1977). At present in the literature, MD simulations of solvated proteins, protein-DNA complexes as well as lipid systems addressing a variety of issues including the thermodynamics of ligand binding and the folding of small proteins are routinely observed. The number of simulation techniques has greatly expanded; the classical simulations, that are being employed to study enzymatic reactions in the context of the full protein. MD simulation techniques are extensively used in experimental study such as X-ray crystallography and NMR structure determination.

## MD simulation steps

There are three typical stages of a MD simulation system; energy minimization, equilibration, and dynamics.

## Energy minimization

The force field, CHARMM, has been assigned to the atoms in the system it is essential to locate a constant point or a minimum on the potential energy surface in order to initiate dynamics. At a minimum on the potential energy surface the lattice force on each atom vanishes. There will be more than one minimum for biopolymer, or a liquid under periodic boundary circumstances. In theory there may be a global minimum. Minimization provides in sequence that is corresponding to molecular dynamics. Ensembles of structures are practical for calculating thermodynamic averages and estimating entropy. Minimized structure represents an underlying configuration about which fluctuations occur during dynamics. The use of a force field to describe structure is regularly called molecular mechanics (Mackerell *et al.*, 2004).

Constraints are imposed during minimization, as well as during dynamics. These constraints based on data from NMR experiment or they are imposed by a template to find the minimum closest in structure to a target molecule. The area of template forcing is also important for comparative modeling. Based upon it is not possible at present to fold a protein by energy minimization. However, the present study determines protein folding by comparing with a structure that has significant amino acid sequence homology. Given the amount of data generated by Microbial complete genome database as well as the extensive databases that contain protein sequences, this is a precious area for research.

Minimization needs a function which is provided by the force field and a starting guess or set of coordinates. The magnitude of the initial derivative can be used to determine the direction and magnitude of a step that change in the coordinates, required to approach a minimum configuration. The magnitude of the first derivative is also a rigorous way to characterize convergence. A minimum becomes to converge when the derivatives are close to zero. To achieve the minimum the structure must be consecutively updated by changing the coordinates and checking for convergence. Each entire cycle of differentiation and stepping is known as "minimization iteration". Typically thousands of "minimization iterations" are required for large macromolecules, protein complex, to reach convergence (Novothy and Sippl, 1997).

There are three major approaches for minimization: steepest descent, conjugate gradient, and newton-raphson. The efficiency of minimization can be judged by both the number of iterations required to converge and the number function evaluations needed per iteration. Steepest descent, first derivative, is not particularly efficient because it must be combined with a line search to determine the step size. A line search requires several function evaluations in order to determine the optimum step size. This technique is robust and is used to minimize initially when the structure is far from the minimum configuration. Newton-Raphson, second derivative, method is prohibitive for large systems, has a maximum of 200 atoms allowed for Newton-Raphson. Therefore this present study has chosen conjugate gradient protocol that implement with NAMD2 package program for energy minimization procedure (Kal'e *et al.*, 1999). The most advanced minimization protocol can be obtained by conjugate gradients protocol. This conjugate gradient technique has applied information from first derivatives to determine the optimum direction for a line search. Consequently, the gradient information is used to predict where along the gradient the function will change direction (Kal'e *et al.*, 1999).

## Equilibration

MD is useful for solving the motion equations for atoms system. The solution for this system represents the time evolution of the molecular motions and the trajectory. Depending on the temperature at which a simulation is allowed for running MD barrier crossing and exploration of multiple configurations. To assign velocities initially is needed to initiate MD. This is done using a random number generator using the constraint of the Maxwell-Boltzmann distribution. The temperature is delineated by the average kinetic energy of the system according to the kinetic theory of gases. The temperature can be estimated by averaging over the velocities of all of the atoms in the system. Ultimately an initial set of velocities has been generated the Maxwell-Boltzmann distribution that maintains throughout the simulation process. By following the minimization, the temperature is considered as being essential zero Kelvin. To initiate the dynamics system must be brought up to the temperature of interest, 300 degree Kelvin for instance. This is done by assigning velocities at some low temperature and then running dynamics according to the

equations of motion. After a number of iterations of dynamics the temperature is scaled upwards. The most common means of temperature scaling is velocity scaling. This is done systematically during the equilibration (initialization) stage. Generally a typical time step of 1 fs equilibration is run for at least 5 ps (5000 time steps) and often for 10 or 20 ps (MacCommon *et al.*, 1977; Mackerell *et al.*, 2004).

## Dynamics

The dynamics stage is the stage of interest to determine an averaging of thermodynamic or sampling new configurations. The application of this scheme is known as production dynamics. MD solving uses the second law of Newton's motion equations $F_i = m_i a_i$, where $F_i$ is the force, $m_i$ is the mass, and $a_i$ is the acceleration of atom i. The force on atom I can be computed directly from the derivative of the potential energy V with respect to the coordinates $r_i$, $F_i = -dV/dr_i$. Therefore $m_i d^2r_i/dt^2 = -dV/dr_i$. Analytical solutions of the equations of motion are possible only for two particles. Large systems require numerical methods (MacCommon *et al.*, 1977).