CHAPTER 2

LITERATURE REVIEW

2.1 Dengue

Dengue viral infection is amongst the most important mosquito-borne diseases in the world. Dengue virus belongs to the Flaviviridae family and is a widespread human pathogen that can cause hemorrhagic fevers (Kautner *et al.,* 1997). This virus is transmitted by the mosquitoes *Aedes aegypti* (principal domestic mosquito vector) and *Aedes albopictus*. There are four serotypes of the Dengue virus with type 2 being the most prevalent (DEN-2).

2.1.1 History and epidemiology

Gubler (1998) reported that dengue had a wide geographic distribution even before the 18th century. The earliest record of dengue infection found to date is in a Chinese encyclopaedia of disease symptoms and remedies first published in 265 to 420 A.D. The first reports of major epidemics of a possibly dengue illness occurred in Asia, Africa and North America in 1779 and 1780. The global pandemic of dengue began after the World War II, where the ecologic disruption in the Southeast Asia and Pacific during this time offered the ideal conditions for increased transmission of mosquito-borne diseases. The first occurrence of Dengue Haemorrhagic Fever (DHF) was recorded in Manila, Philippines, between 1953 and 1954, and within 20 years, epidemic of the disease had spread throughout Southeast Asia. In the Pacific Islands, dengue was reintroduced in the 1970s after its absence of 25 years. In Africa, epidemic dengue fever caused by all 4 serotypes has increased dramatically since 1980. In Central and South America, geographical distribution of *Aedes aegypti* was wider in 1997 than its distribution during the eradication programme in the 1950s to 1970s. Figure 2.1 shows the world distribution of dengue transmitted countries.

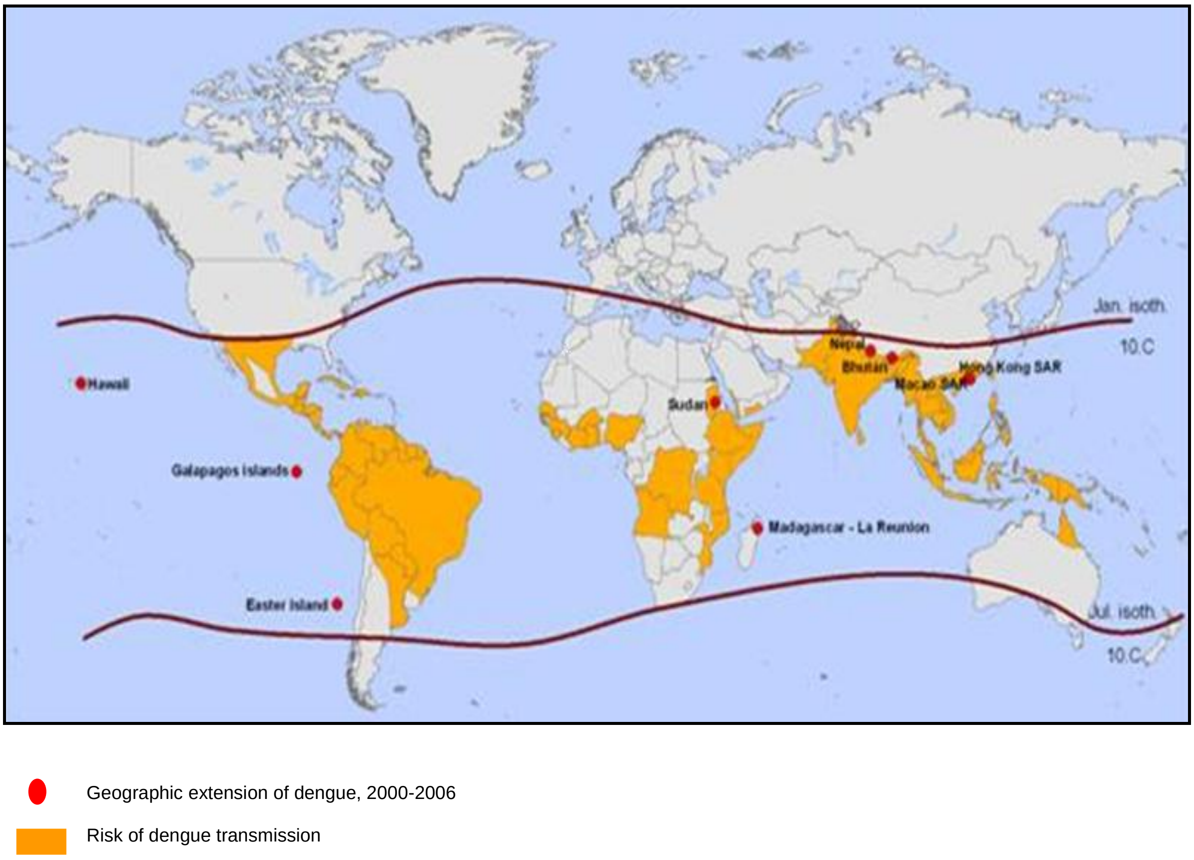


Figure 2.1 World map showing countries / areas at risk of dengue transmission (WHO, 2006; http://www.who.int/csr/disease/dengue/impact/en/index.html).

On 18 May 2002, the WHO General Assembly confirmed dengue fever as a matter of international public health priority through a resolution to strengthen dengue control and research (Guha-Sapir & Schimmer, 2005). Historically, DF and DHF were reported to occur predominantly among urban populations. However, reports have also shown some cases of higher incidence rate in rural than urban areas. Increased in transportation, mobility and spread of urbanisation are the most frequent cited reasons for its occurrence increase.

2.1.2 Dengue syndromes

Dengue infection can cause a spectrum of illness ranging from mild, undifferentiated fever to illness with high fever up to 7 days' duration. In infants, this may be accompanied by macular or maculopapular rash, while older children and adults commonly experience high fever, severe headache, retro-orbital pain, arthralgia and rash, but rarely death (WHO, 1999). Dengue Haemorrhagic Fever (DHF), however, is a deadly complication of dengue fever. Symptoms include haemorrhagic tendencies, thrombocytopenia and plasma leakage. Dengue Shock Syndrome (DSS) includes all the above criteria plus circulatory failure, hypotension for age and low pulse pressure. DHF and DSS are potentially deadly but patients with early diagnosis and appropriate therapy can recover with no sequelae. Case management for Dengue Fever (DF) is mainly symptomatic and supportive. DHF requires continuous monitoring of vital signs and urine output, while DSS is a medical emergency that requires intensive care unit hospitalization.

2.1.3 Dengue virus

There are four serotypes of dengue virus: Dengue virus type 1 (DEN-1), DEN-2 (the most prevalent Dengue serotype), DEN-3 and DEN-4. Dengue is an icosahedral enveloped virus with a diameter of approximately 500 Å and contains a single-stranded RNA of positive polarity (Kuhn *et al.*, 2002). Figure 2.2 illustrates the images of a dengue virus particle determined by cryoelectron microscopy and image reconstruction techniques.

The RNA genome codes for a single polyprotein precursor of 3 391 amino acids, comprising three structural and seven non-structural proteins, arranged in the order C-prM-E-NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5 (Irie *et al.*, 1989). The structural proteins are the capsid (C), premembrane (prM; precursor of membrane, M) and envelope (E). The non-structural proteins include the large, highly conserved NS1, NS3 and NS5 proteins, and four smaller hydrophobic proteins NS2A, NS2B, NS4A and NS4B. Co- and post-translational proteolytic processing of this polyprotein precursor are catalysed by the host cell and virus encoded proteases to yield the mature viral proteins (T. J. Chambers *et al.*, 1990). Of these mature viral proteins, the envelope protein, E, and non-structural proteins NS1, NS3 and NS5, are of immense interest in the design and development of vaccine and therapeutic agent in the effort to fight against dengue virus infections (Wahab *et al.*, 2007). Figure 2.3 illustrates a schematic of the proteolytic processing of the DEN-2 polyprotein.

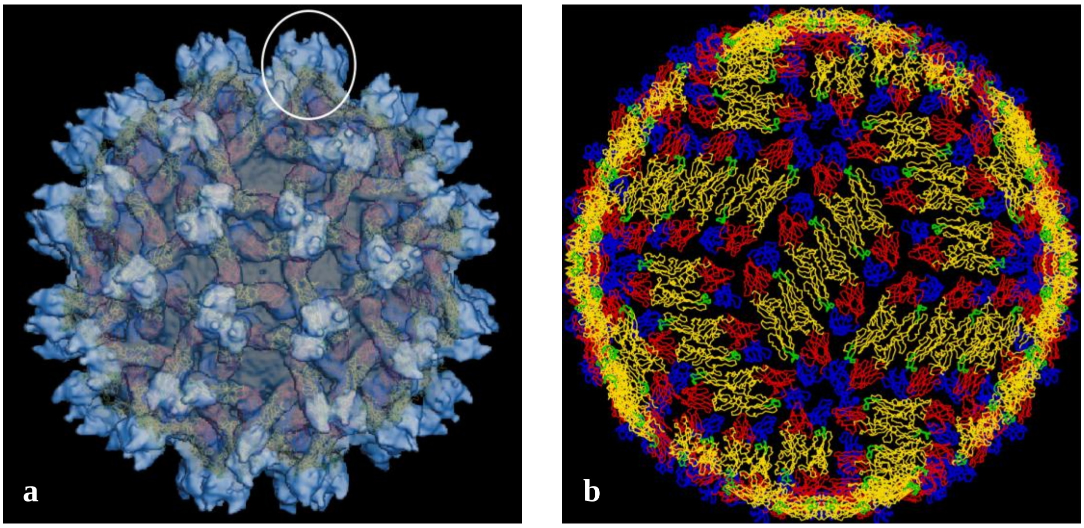


Figure 2.2 Images of a dengue virus particle (denoting a single virus) determined by Michael Rossmann and Richard Kuhn's team using cryoelectron microscopy and image reconstruction techniques (Kuhn *et al*., 2002; Y. Zhang *et al.*, 2003). (a) Image of an immature dengue particle determined to 16 Å resolution. There are 60 icosahedrally organized trimeric spikes on the particle surface making the immature particle far less smooth than the mature form. One such spike is circled for reference, each consisting of three prM:E heterodimers, where E is an envelope glycoprotein and prM is the precursor to the membrane protein M. The spikes cause the immature particle to have a considerably larger diameter (600 Å) than the mature virion. (b) The structure of the mature dengue virus particle determined to 24 Å resolution (diameter of 500 Å). The virus surface is smooth and its membrane is completely enclosed by a protein shell. The protein is color-coded blue, green and yellow to show its three specific domains and its shell serves as a cage for the genetic material inside.

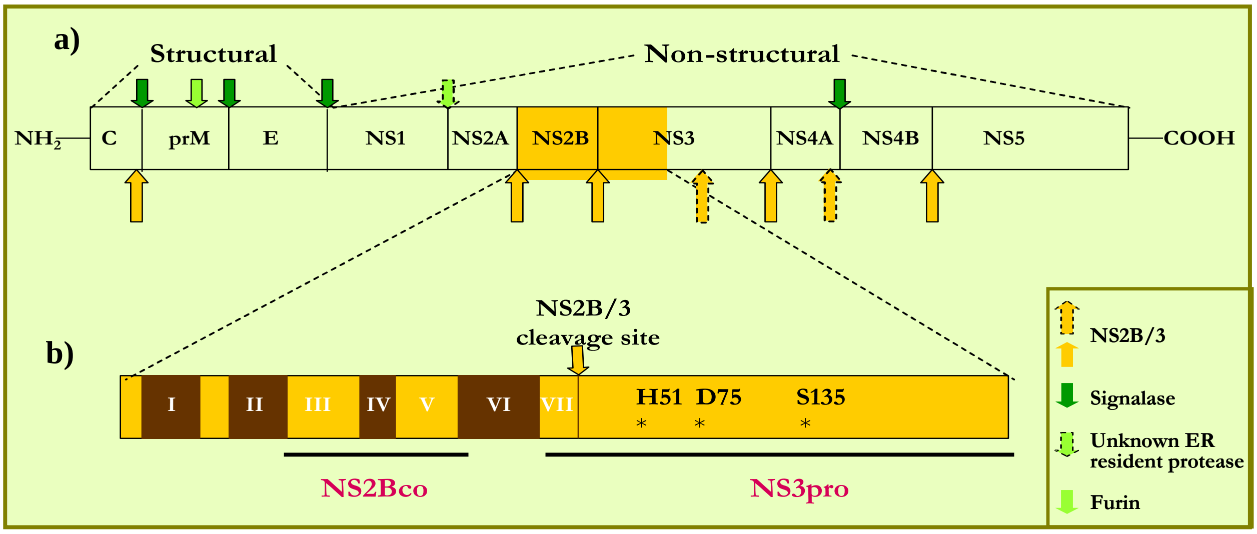


Figure 2.3 Schematic representation of the dengue polyprotein processing (Brinkworth *et al.*, 1999). a) The cleavage sites on the polyprotein cleaved by host-encoded proteases (green arrows above) and the virus-encoded protease complex NS2B-NS3 (yellow arrows below) are shown. Secondary cleavage within individual proteins is shown as dotted arrows. The NS2B cofactor and the protease domain of NS3 (NS3pro) are shaded in yellow. b) Schematic highlighting the key features of the NS2B cofactor and NS3pro. The hydrophobic domains of NS2B are shaded in black; the region underlined with black bar indicates the 40 amino acid hydrophilic region shown to be the minimal requirement for the functional NS2B cofactor involvement in the catalytic activity of NS3. The catalytic triad of NS3 is denoted as asterisks.

2.1.4 NS2B-NS3 protease complex

Dengue replication is dependent upon the correct cleavage of the viral polypeptide which requires both the host cell proteases and a virus-encoded, two-component protease, NS2B-NS3 (Falgout *et al.*, 1991; Yusof *et al.*, 2000). NS3 contains a trypsin-like serine proteinase domain at its N-terminal of 180 amino acid residues thus suggesting its role as the putative viral protease (Bazan & Fletterick, 1989; Gorbalenya *et al.*, 1989). Analysis of virus sequence revealed the catalytic triad of DEN-2 protease to be His51, Asp75 and Ser135 (Brinkworth *et al.*, 1999). The N-termini of several non-structural proteins are produced by cleavage at dibasic sites by NS2B-NS3. The protease activity of NS3 is dependent upon the presence of NS2B for optimal catalytic activity, and biochemical studies and deletion analyses had mapped the required region on the NS2B to a central, hydrophilic 40 amino acid domain (Lys54-Glu93) in an otherwise relatively hydrophobic protein (Arias *et al.*, 1993; Clum *et al.*, 1997; Yusof *et al.*, 2000). Hence, the NS2B-NS3 protease complex serves to be a target for the development of antiviral drugs.

The recognition site for cleavage activity of the protease is the two basic amino acids at the P1 and P2 positions of the cleavage junctions of the non-structural proteins, ie. Lys-Arg, Arg-Arg or Arg-Lys (Gln-Arg for NS2B/NS3 junction), followed by short side-chain amino acids, Gly, Ser or Ala, at P1' position (Preugschat *et al.*, 1990; Yusof et al., 2000; L. Zhang & Padmanabhan, 1993). Figure 2.4 shows a schematic representation of a protein substrate binding to a protease.

The exact mechanism of how the NS2B cofactor enhances the efficiency of proteolytic activity of the protease is yet to be fully determined. Recent

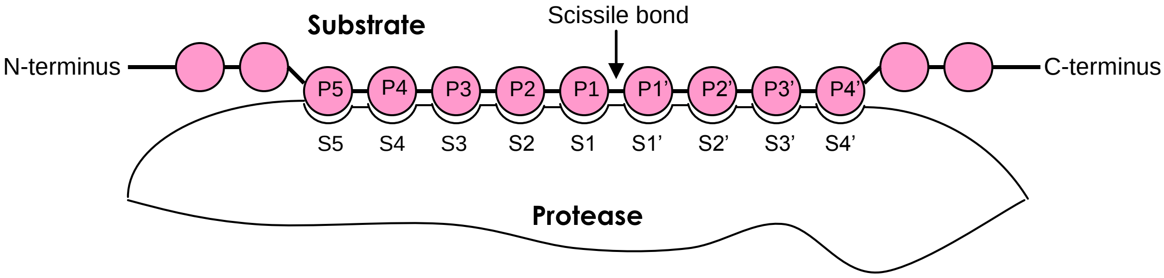


Figure 2.4 Schematic representation of a protein substrate binding to a protease (Turk, 2006). The system of nomenclature to describe the interaction of a substrate with a protease is known as The Schechter and Berger nomenclature (Schechter & Berger, 1967). The protease surface that is able to accommodate a single side-chain of a substrate residue is known as subsite; numbered S1-Sn upwards towards the N-terminus of substrate, and S1'-Sn' towards the C-terminus, beginning from the sites on each side of the scissile bond. The substrate residues that the protease accommodate are numbered P1-Pn and P1'-Pn', respectively. Consequently, the structure of the protease active site determines which substrate residues can bind to specific substrate binding sites of the protease, hence, determining the substrate specificity of a protease.

development in protein crystallography study and nuclear magnetic resonance (NMR) spectroscopic study of the protease complex have shed some light into the function of NS2B cofactor in the activation of the protease complex (D'Arcy *et al.*, 2006; Erbel *et al.*, 2006; Melino *et al.*, 2006).

2.1.5 Dengue vaccine and antiviral drug development

In theory, a dengue vaccine is highly feasible since dengue virus causes only acute infection and the viral replication is effectively controlled after a short period of 3 to 7 days of viraemia. In addition, individuals who recovered from dengue virus infections are immune to rechallenge with the same serotype but not to other serotypes of dengue virus (Chaturvedi *et al.*, 2005). However, if this is the case, an effective vaccine will have to be tetravalent, the formulation of which should retain immunogenicity of all four serotypes. This has proven difficult, requiring the use of more complicated, multiple-dose immunization regimes.

Live attenuated viruses are the most successful viral vaccines comprising about 63% of the vaccines approved for use by the US Food and Drug Administration (Ray & Shi, 2006). World Health Organisation (WHO) has put a high priority in developing an effective vaccine against dengue viruses (Pervikov, 2000). Several efforts have been made to develop these vaccines. The most advanced live attenuated tetravalent vaccine against dengue virus was developed in Mahidol University, Thailand, with the support of WHO’s South-East Asia Regional Office (Thomas J. Chambers *et al.*, 1997; Edelman, 2005; Pervikov, 2000; Sabchareon *et al.*, 2002). The tetravalent vaccine formulations were developed by combining successful monovalent vaccines, and are currently in the advanced phase of clinical trials, under the license of Aventis Pasteur

(Hombach *et al.,* 2005). This may allow this dengue vaccine to make its way as the first effective licensed tetravalent vaccine available to general public. Another promising tetravalent vaccine currently under research is by a group at the Walter Reed Army Institute of Research. The tetravalent vaccines developed by combining successful monovalent vaccine strains, licensed by GlaxoSmithKline, were also found to be acceptably safe and is able to induce an antibody response against the respective dengue serotypes in Phase I and Phase II clinical trials (Edelman 2003, Kanesa-Thasan 2003). The Phase III clinical trials of the vaccines are currently on-going and expected to be completed by 2010 (Halstead, 2008). Although life attenuated tetravalent dengue virus vaccines could provide life-long immunity against all four dengue serotypes, these types of vaccines have potentially harmful or toxic effects should these viruses revert to a virulent form of virus (Wahab *et al.,* 2007). In addition, the development of one is a long, difficult and laborious process.

Another approach in the development of a dengue vaccine is via molecular engineering vaccine design. A chimeric vaccine (ChimeriVax-D2) was developed where a chimeric yellow fever (YF)-dengue type 2 (DEN-2) virus was constructed using a recombinant cDNA infectious clone of a YF vaccine strain as a backbone into which the prM and E genes of DEN-2 virus were inserted (Guirakhoo *et al.,* 2002; Guirakhoo *et al.,* 2000). Tetravalent chimeric vaccines using prM and E genes of each wild-type dengue virus representing serotypes 1 to 4 have also been developed (ChimeriVax-DEN1-4) (Guirakhoo *et al.,* 2004). These chimeric vaccine candidates are currently undergoing early clinical trials (Edelman, 2005).

Besides vaccine development, another approach that has been taken to overcome the problems of dengue infections is the development of anti-dengue therapeutics.

Potentially, these are compounds that could inhibit any process in a viral life cycle critical for its reproduction. Many compounds, either synthetic peptide or peptidomimetics, or from natural product, have been screened for anti-dengue viral activities, directing the activities towards the inhibition of proteolytic processing by NS3 protease. Leung *et al*. (2001) reported the activities of the first substrate-based peptide inhibitors, some of which showed potent inhibitory activity against recombinant CF40.gly.NS3 protease. In another work, Chanprapaph *et al.* (2005) reported the activities of synthetic peptides (ranging from hexa- to dipeptides), as competitive inhibitors of the protease with K_i values ranging from 67 to 12 μM.

Yin *et al*. (2006a; 2006b) designed, synthesized and screened substrate-based tetrapeptide inhibitors with various warheads against the dengue virus NS3 protease. The peptide inhibitors with elecrophilic warheads demonstrated effective inhibitions. Three combinatorial libraries of linear hexapeptides with arginine or lysine at the P1 position synthesized by Teoh *et al.* (2005), however, exhibited weak inhibitory effects. Subsequently, two cyclic peptides were synthesized with the basic amino acids, arginine or lysine, incorporated at the P1 position (Wahab *et al.*, 2007). These cyclic peptides showed more promising and better inhibitory effect compared to all the linear peptide-based compounds screened. Presumably, the more rigid cyclic nature of the peptides helped to improve structure stability, making it a much better probe for the active site of the protein and enhancing the potency of peptide-based inhibitors.

Inhibitors isolated from plant resources have also been reported. Castanospermine, a natural alkaloid derived from the black bean tree (*Castanospermum australae*) was reported to inhibit infection and viral spread of all four serotypes of the dengue virus (Whitby *et al.*, 2005). Two sulphated polysaccharides isolated from the red seaweeds *Gymnogongrus griffithsiae* and *Cryptonemia crenulata* were selective inhibitors of DEN-2 multiplication in Vero cells with IC_{50} values around 1 µg/ml (Talarico *et al.*, 2005). In another work, two cyclohexenyl chalcone derivatives, 4-hydroxypanduratin A and paduratin A, isolated from *Boesenbergia rotunda* exhibited competitive inhibitory activities towards DEN-2 NS3 protease with K_i values of 21 and 25 µM, respectively. In addition, two other compounds isolated from the same plant, namely pinostrobin and cardamonin, were observed to exhibit non-competitive inhibitory activities (Tan *et al.*, 2006).

2.2 Serine protease

Dengue virus protease belongs to the protein fold family of trypsin-like serine-proteases (S07.001; according to MEROPS) (Rawlings *et al.*, 2006), which falls into the all-β proteins class of protein structure (according to SCOP) (Bazan & Fletterick, 1989; Murzin *et al.*, 1995). Proteases, also known as proteolytic enzymes, are enzymes that catalyse the breakdown of proteins by hydrolysis of peptide (amide) bonds. Four major protease families are serine proteases, cysteine proteases, aspartic proteases and metallo proteases. Serine proteases are so named as they have a highly reactive serine residue that attacks the carboxyl group of the substrate (Garrett & Grisham, 1997b). They are divided into two sub-families: the chymotrypsin-family which includes chymotrypsin, trypsin and elastase, and the subtilisin-family which includes bacterial enzymes such as subtilisin. Substrate specificity of serine proteases (chymotrypsin-family) is restricted to the P1 residue. Besides peptide bonds, this family of serine protease also cleaves ester bonds of certain synthetic substrates.

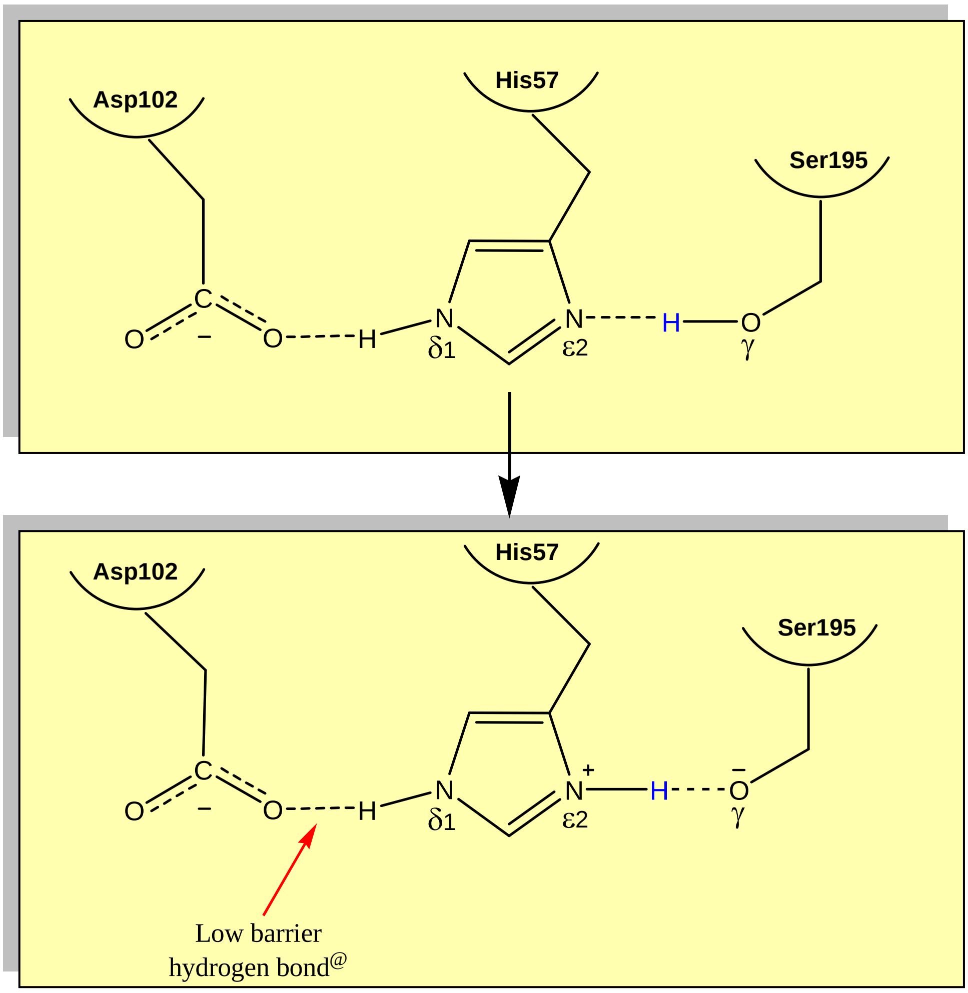
Three residues essential for the catalytic process of serine protease are known as the catalytic triad comprising of His57, Asp102 and Ser195 residues (Hedstrom, 2002). In dengue virus, these residues are His51, Asp75 and Ser135 (Brinkworth *et al*., 1999; Gorbalenya *et al*., 1989). These residues form a charge-transfer relay network (Figure 2.5); His57, polarized by Asp102, acts as a proton shuttle which accepts the proton from Ser195 as the latter makes a nucleophilic attack on the substrate (thus forming a tetrahedral intermediate). The nucleophilicity of Ser195 is increased by the His57 side chain. The imidazolium ion (ring) of His57 does not, however, donate a proton to Asp102 in the intermediate (Bachovchin, 1985). The mechanism of catalysis of peptide hydrolysis involves the acylation and deacylation steps which is illustrated in Figure 2.6:

i. Acylation step:

The nucleophilic Oγ of Ser195 attacks the carbonyl carbon of the scissile bond of the substrate, forming the tetrahedral intermediate. The proton donated from the Oγ atom to His57 is then donated to the N atom of the scissile bond, while cleaving the C-N peptide bond and the C-O ester bond, to produce the amine (which diffuses away) and the reasonably stable acyl-enzyme intermediate.

ii. Deacylation step:

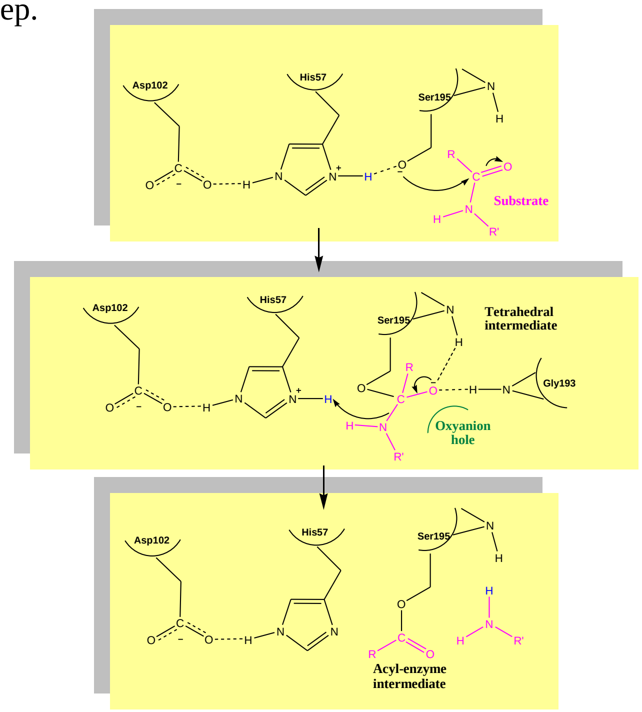
A water molecule loses a proton to His57, and the resulting OH nucleophile attacks the acyl-enzyme intermediate, forming another tetrahedral intermediate. The proton is then donated to the Ser Oγ, releasing the acid product.



@(Cassidy *et al.*, 1997; Frey *et al.*, 1994; Santis & Carloni, 1999)

Figure 2.5 Charge-relay network in the catalytic triad of serine protease (the numbering of the amino acid residues refers to that of the chymotrypsin-family of serine protease).

(i) Acylation step.



Covalent enzyme-substrate complex

(ii) Deacylation step.

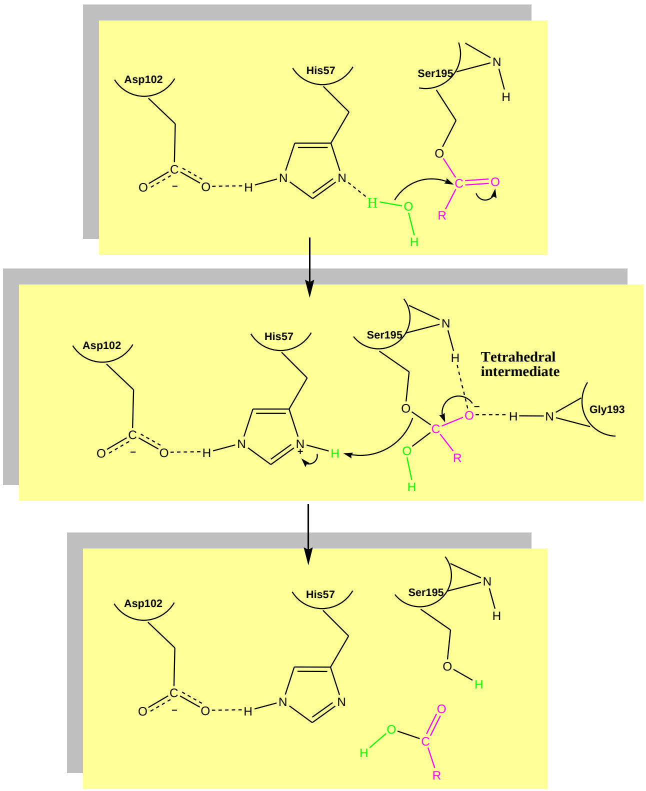


Figure 2.6 Mechanism of peptide hydrolysis by serine protease catalytic triad. (Adapted from Garret & Grisham, 1997b)

2.3 Protein Crystallization

A detailed understanding of three dimensional (3-D) protein structure is important in the design of new drugs. X-ray diffraction is the most powerful method to determine the structure of these large molecules. However, it is only applicable when suitable protein crystals are obtained (Saridakis & Chayen, 2003). Protein crystallization has always been considered a bottleneck to protein structure determination, primarily, due to the many variables involved in the experimental setups, in addition to the uncertain reproducibility of some successful trials. Some even considered ‘dumb luck’ to be one of the ingredients required for a successful crystallization trial (Cudney, 1999).

2.3.1 History and background

When the Protein Data Bank, PDB, (http://www.pdb.org) was founded in 1971, it contained only seven structures (Berman *et al.*, 2000). Initial progress was slow in determining protein shapes due to the difficulty in crystallizing proteins. Historically, the processes involved had been time-consuming and expensive. However, since then, the pace at which protein structures are being discovered has grown exponentially and the PDB currently contains over 46 000 structures. The bottleneck of obtaining suitable protein samples was greatly reduced with the advancement in molecular biology tools developed in the 1980s and 1990s, as well as methods for parallel expression and purification of large numbers of gene products (Hosfield *et al.*, 2003).

Proteins, like many other molecules, can be prompted to form crystals when placed in the appropriate conditions. Many of these macromolecules are still functional in the crystalline state (Rhodes, 2000). In order to crystallize, the purified protein undergoes slow precipitation from an aqueous solution. Consequently, individual protein molecules align themselves in a series of unit cells by adopting a consistent orientation. The crystalline lattice that is formed is held together by non-covalent interactions. The goal of crystallization is usually to produce a well-ordered crystal that is lacking in contaminants and large enough to provide a diffraction pattern when hit with an x-ray beam (Figure 2.7). It is from the reflections of the diffraction that an electron-density map can be created. The macromolecule's known sequence is then fitted to the map for a 3-D protein structure. Difficulty arises because protein crystals are fragile in nature and have irregularly shaped surfaces due to the formation of large channels within the crystals. Hence, non-covalent bonds that hold the lattice together are often formed through several layers of solvent molecules. Some factors to be considered for a successful crystallization are protein purity, pH, protein concentration, temperature and precipitants. The protein solution should usually be at least 97 % pure, in order to gain sufficient homogeneity.

2.3.2 The crystallization phase diagram

Finding favourable conditions for crystallization is usually achieved by screening of the protein solution with numerous crystallizing agents (Chayen, 2005). Optimization of the crystallization conditions involves the manipulation of the crystallization phase diagram with the aim of leading crystal growth in the direction that will produce the desired results. Figure 2.8 illustrates a protein crystallization phase diagram. The phase diagram shows the stable state (liquid, crystalline or amorphous solid) under a variety of crystallization parameters. This provides a means of quantifying the influence of the parameters on the production of crystals.

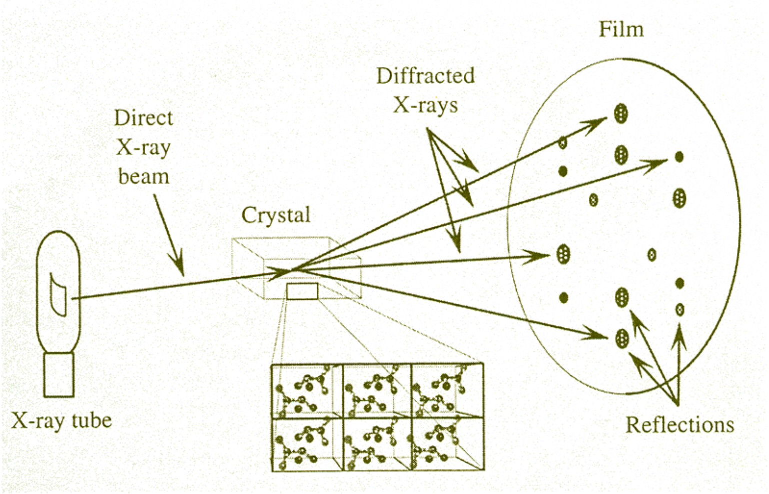


Figure 2.7 Crystallographic data collection. The crystal diffracts the source beam into many discreet beams, each of which produces a distinct spot (reflection) on the film. The positions and intensities of these reflections contain the information needed to determine molecular structures (Rhodes, 2000).

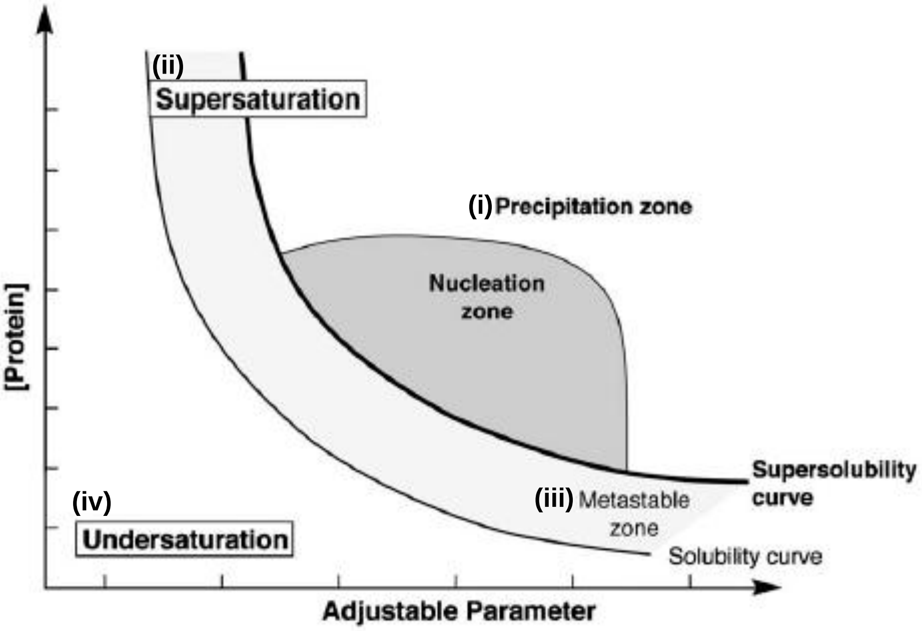


Figure 2.8 Schematic illustration of a typical protein crystallization phase diagram (Chayen *et al.*, 1996). (i) Area of very high supersaturation where protein will precipitate; (ii) area of moderate supersaturation where spontaneous nucleation takes place; (iii) area of lower supersaturation just below the nucleation zone where crystals are stable and may grow but no further nucleation will occur (the metastable zone is thought to contain the best conditions for growth of large well-ordered crystals; (iv) undersaturated area where the protein is fully dissolved and will never crystallize.

Crystallization proceeds in two phases: nucleation and growth. Nucleation is a prerequisite for crystallization, and requires different conditions than those of growth. Once nucleus has formed, growth will spontaneously follow. Ideally, once the nuclei are formed, the protein concentration in the solute will drop, leading the system into the metastable zone where growth should occur, without further formation of nuclei. However, often, excess nucleation occurs resulting in the formation of numerous low-quality crystals.

2.3.3 Techniques in protein crystallization

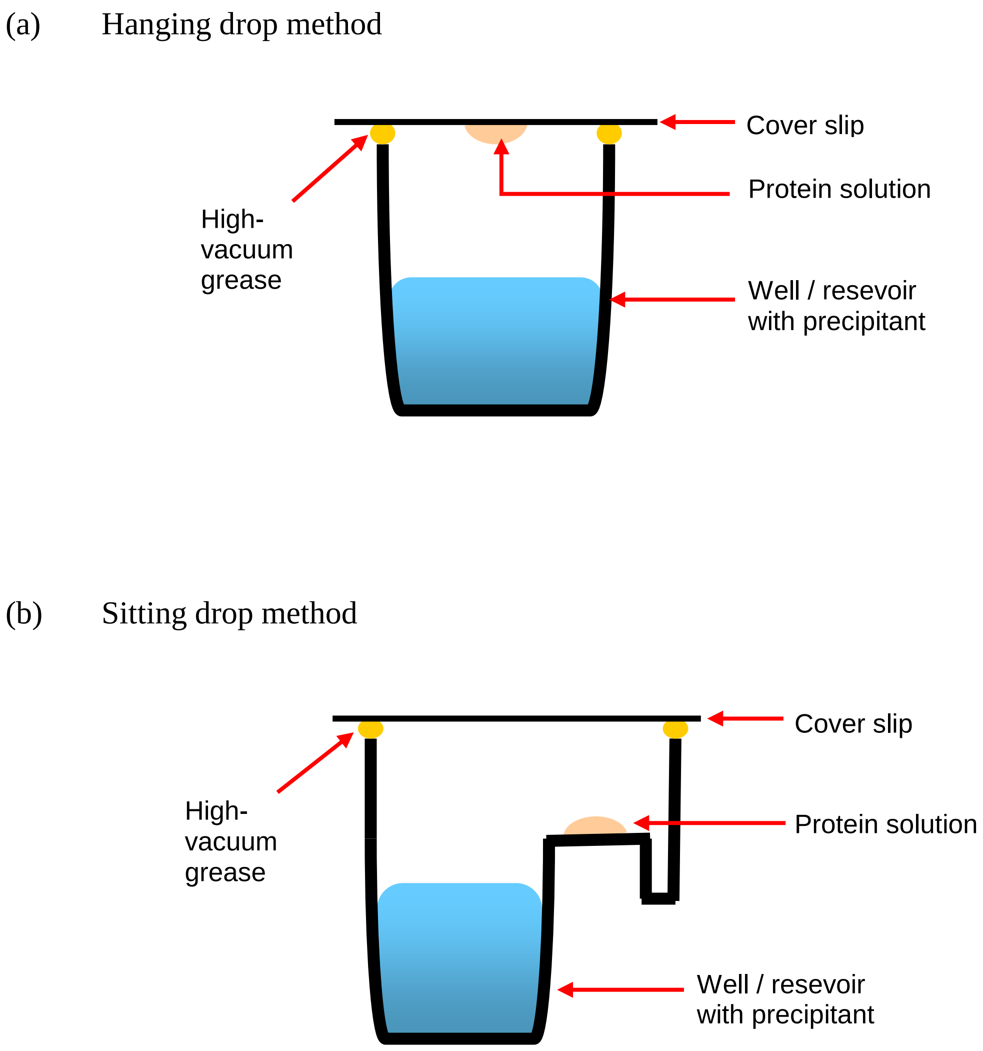
Optimal conditions for crystallization of a protein are difficult to predict, since every protein is unique in its physical and chemical properties, with unique 3-D structure and distinctive surface characteristics. The variables influencing crystal growth are too many to allow an exhaustive search. Hence, a sparse matrix method of trial conditions is used, which is based and selected from known crystallization conditions for macromolecules. Consequently, it is possible to test in a reasonably short time, wide ranges of pH, salts and precipitants using very small sample of macromolecules. Results from initial trials could produce crystals or solubility information. Most common methods for initial crystal trials are the vapour-diffusion methods, dialysis and batch method. In this study, the vapour-diffusion methods were employed.

2.3.3.1 The vapour-diffusion technique

In the vapour-diffusion technique, the protein/precipitant solution is allowed to equilibrate in a closed container with a larger aqueous reservoir whose precipitant concentration is optimal for producing crystals. Two methods of set-up are involved with this technique: the hanging drop method and the sitting drop method.

The hanging drop vapour diffusion technique is the most popular method for crystallization of macromolecules (Adachi, 2003; Hampton Research Corp., 2006a). Figure 2.9A illustrates the set-up for a hanging drop method, where a droplet containing purified protein, buffer, and precipitant is placed onto a siliconized glass cover slip. Siliconized cover slip is used as it provides hydrophobic surface to produce a drop which “stands well” and does not flatten on the glass, and it prevents the adhesion of crystals and precipitates onto the glass surface. The cover slip is then ‘glued’ in an inverted position onto a reservoir containing similar buffer and precipitant in higher concentrations. Initially, the droplet of protein solution will contain an insufficient concentration of precipitant for crystallization. However, as water vaporizes from the drop and transfers to the reservoir, the precipitant concentration in the drop increases to a level optimal for crystallization. Since the closed-system is in equilibrium, these optimum conditions are maintained until the crystallization is completed (Drenth, 1999; Rhodes, 2000).

Another popular method for protein crystallization is the sitting drop vapour diffusion technique (Hampton Research Corp., 2006b). This method is usually preferable if the protein solution has a low surface tension and tends to spread over the cover slip in the hanging drop method. The droplet containing protein, buffer and precipitant is placed on a platform in vapour equilibration with reservoir containing the same crystallization reagent (Figure 2.9B). The disadvantage of this method is that crystals can sometimes adhere to the sitting drop surface making removal difficult.



(a) Hanging drop method

(b) Sitting drop method

Figure 2.9 Vapour diffusion techniques in protein crystallization trial.

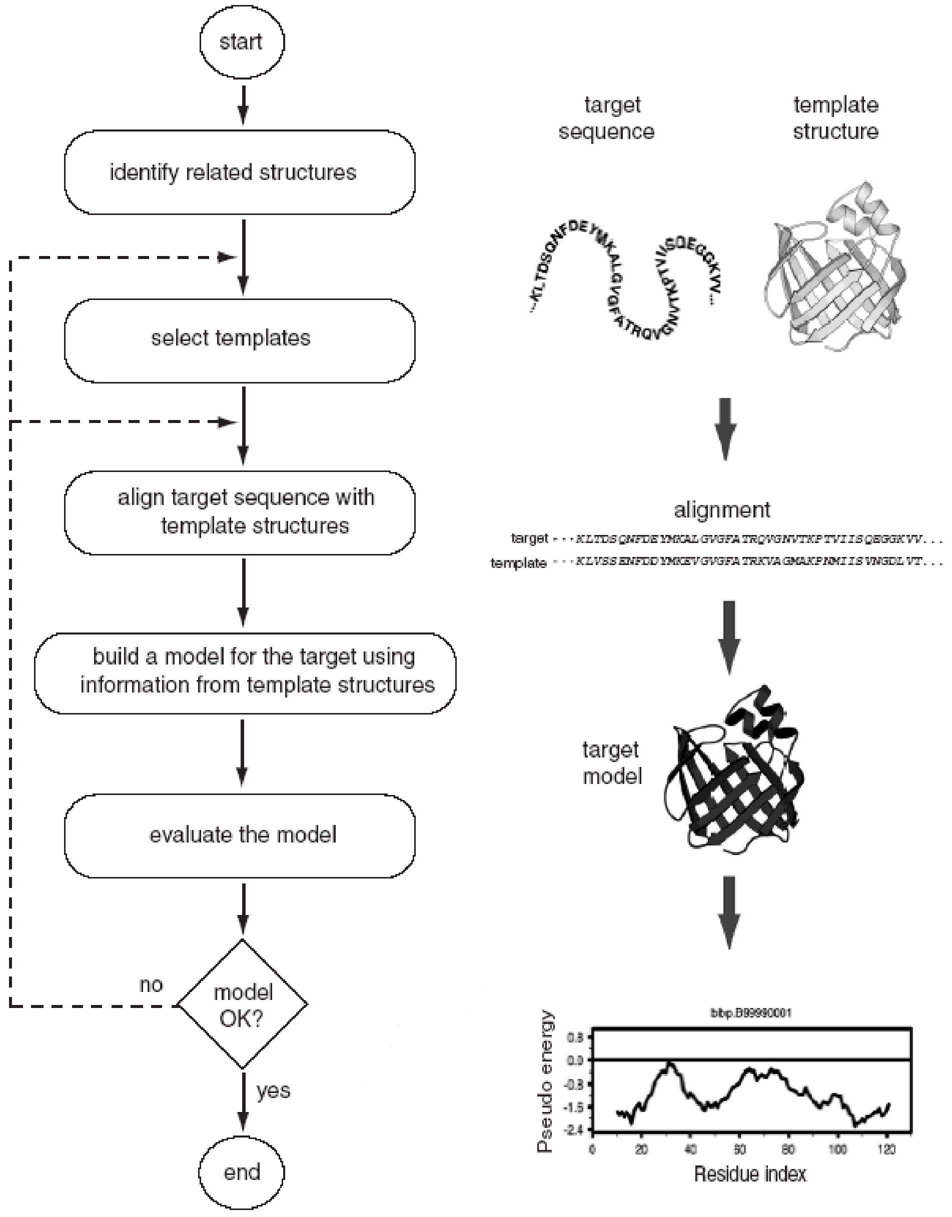
2.4 Molecular modelling studies of protein

Molecular modelling is a method to mimic the behaviour of molecules and molecular systems. It is a collection of techniques for deriving, representing and manipulating the structures and reactions of molecules, and those properties that are dependent on these 3-D structures. Today, molecular modelling is customarily associated with computer modelling which has revolutionised the molecular modelling techniques. The interaction between molecular graphics and their underlying theoretical methods has enhanced the accessibility of molecular modelling methods and assisted in the analysis and interpretation of the calculations made (Leach, 2001b). Amongst the types of biological activities that have been investigated using molecular modelling approach include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA and membrane complexes.

2.4.1 Protein secondary structure prediction and homology modelling

The biological function of a protein is often dependent upon the conformation adopted by the molecule. Knowledge on the 3-D structure of a protein is therefore crucial towards understanding its function. X-ray crystallography and nuclear magnetic resonance (NMR) are the experimental techniques most widely used to provide detailed information about protein structures. Since these methods are often tedious and difficult, computational techniques have become very popular in generating models of proteins. Homology, or comparative modelling remains the only method that can reliably predict the 3-D structure of a protein with accuracy comparable to that of a protein structure resolved at low-resolution via experimental means (Martı-Renom *et al.,* 2000). Figure 2.10 outlines the steps involved in a homology modelling of protein structure. This technique relies upon the alignment of a protein sequence of unknown structure (target) to a homologue of known structure (template). However, potential problems can occur in structural determination when the target protein and template have less than 25 % sequence identity (based on an average domain length of 80 amino acids) (Marti-Renom *et al.*, 2003; Sander & Schneider, 1991). Nevertheless, sufficiently long alignments can still infer structural similarity, even when the sequence similarity is below 25 % (Lund *et al.*, 1997).

With no homologue of known structure from which to make a 3-D model, a logical next step for protein modelling is to predict secondary structure which aims to provide the location of α-helices and β-strands within a protein or a protein family. Once the secondary structure of a protein has been determined, the protein fold recognition can be carried out, followed by a prediction of the tertiary structure. There are many methods (web servers) available to perform secondary structure prediction. In earlier work, prediction success has been rather low. For example, Kabsch and Sander (1984) reported the low level of success of prediction without additional information to be commonly attributed to the neglect of long-range interactions within a protein. However, Pan *et al*. (1999) attributed the low levels in prediction accuracy to the limitation of the available protein database size or prediction algorithm. Nevertheless, the field of secondary structure has achieved a break-through by combining algorithms from artificial intelligence with evolutionary information (Rost, 2003), boosting the current prediction accuracy to around 77 %. Thus, secondary structure prediction can be accurate enough to be taken seriously as a tool to assist in the design of experiments to probe protein structure and function (Barton, 1995).


start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
model OK?
no
yes
end
target sequence
template structure
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
alignment
target - - - KLTDS QNFDE YMKAL GVGFA TRQVG NVTKP TVI IS QEGGK VV . . .
template - - - KLVSS ENFDD YMKEV GVGFA TRKVA GMAKP NMI IS VNGDL VT . . .
target model
bbp.B99990001
Pseudo energy
0.8
0.0
-0.8
-1.5
-2.4
0
20
40
60
80
100
120
Residue index

.

Figure 2.10 Steps in homology modelling of protein structure (Marti-Renom *et al.*, 2003).

2.4.2 Computational approaches to study ligand-protein interactions

Molecular modelling techniques are widely used in the drug discovery area of research. Common methods in molecular modelling include quantum mechanics (*ab initio* calculation, semi-empirical calculation and density functional theory), molecular mechanics, molecular dynamics, Monte Carlo and docking simulations. Each of the modelling method requires different magnitude of computational power depending on the size of the system studied (Figure 2.11).

Discovering and developing any new drug is a long and expensive process (Leach, 2001a). Two key steps in a drug discovery programme are the identification of a hit molecule (a molecule that has some reproducible activity in a biological assay), and identification of lead series (a set of molecules that share some common structural feature but exhibit some variation in the activity as the structure is modified). Computational techniques can play a significant role in the lead optimization stage where the synthesis of a drug candidate with the desired potency and selectivity, lack of toxicity and the appropriate features for it to reach the target *in vivo* could be designed rationally. In rational drug design, computer modelling of enzyme-ligand interactions replaces much of the initial chemical synthesis and clinical pre-screening of potential therapeutic agents, saving much time and effort in drug development (Garrett & Grisham, 1997a).

The basic computational calculations performed in molecular modelling studies are (Foresman & Frisch, 1996):

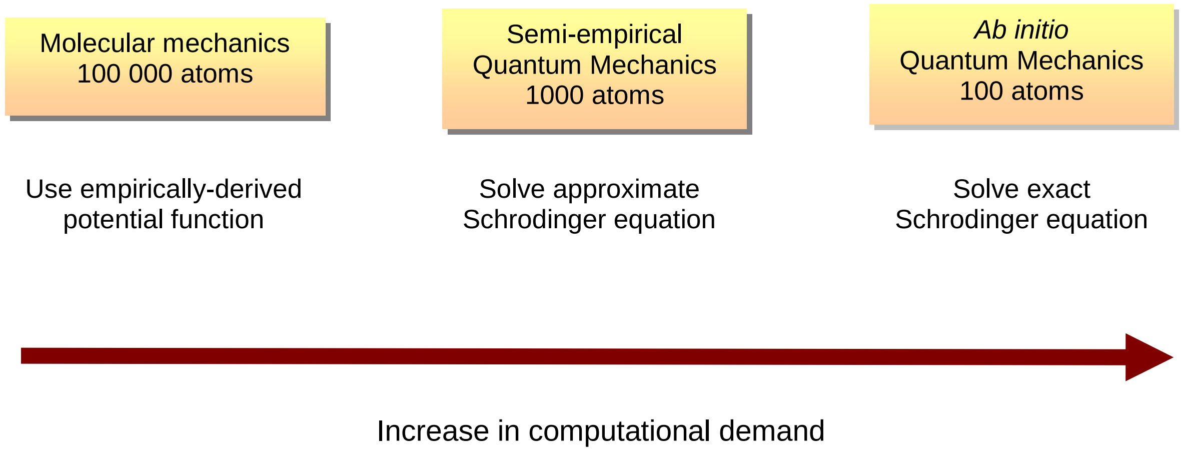

Molecular mechanics
100 000 atoms
Semi-empirical
Quantum Mechanics
1000 atoms
Ab initio
Quantum Mechanics
100 atoms
Use empirically-derived potential function
Solve approximate Schrodinger equation
Solve exact Schrodinger equation
Increase in computational demand

Figure 2.11 Different strengths and computational costs of the different molecular modelling methods.

- computing the energy of a particular molecular structure (spatial arrangement of atoms or nuclei and electrons). Hence, properties related to the energy may also be predicted;
- performing geometry optimizations to locate the lowest energy molecular structure that is in close proximity to the specified starting structure. This depends on the gradient of the energy (the first derivative of the energy with respect to atomic position); and
- computing the vibrational frequencies of molecules resulting from interatomic motion within the molecule. This depends on the second derivative of the energy with respect to atomic structure. Properties which depend on second derivatives may also be predicted.

2.4.2.1 Types of molecular interactions.

Tight-binding ligands (substrates or inhibitors) often have a high degree of complementarity with the target receptors (macromolecules to which the ligands bind). Besides shape complementarity, various inter- and intramolecular interactions also affect the protein folding and protein-ligand binding processes. Covalent interactions involve the formation of permanent bonds between atoms, such as disulphide bridges. Non-covalent interactions include electrostatic interactions, van der Waals interactions, polarization and charge transfer effects. Salt bridges (between ions of opposite charges), and hydrogen bonds (due to polarization effects) are often observed between amino acid residues of proteins and ligands. Hydrogen bonds can make important contributions to the binding energy and stabilization of molecular structures. van der Waals interactions represent electron correlations, and even though these interactions are weak interactions, the total van der Waals interactions in a biological system can consiberably contribute to the stabilization of macromolecular structures and protein-ligand interactions.

Another important driving force in the formation of macromolecular structures and ligand recognition is the hydrophobic interactions. As hydrophobic regions of two interacting molecules come together, there is a favourable increase in the system's entropy as the solvent molecules (previously in an ordered shell around the hydrophobic surface) become disordered. There is also an energetic contribution involved as more favourable apolar-apolar interactions replace the unfavourable apolar-polar interactions (Chothia & Janin, 1975; Jones & Thornton, 1996).

2.4.2.2 Computational docking

The docking process involves the prediction of ligand conformation and orientation (or posing) within a targeted binding site (Kitchen *et al.*, 2004). Generally, two aims of docking studies are to model structures accurately and to correctly predict activities. The binding affinity prediction problem addresses the question of how well the ligands bind to the protein (scoring) (Sousa *et al.*, 2006). Examples of popular docking programs include AutoDock (Huey *et al.*, 2007; Morris *et al.*, 1998), GOLD (Jones *et al.*, 1995; Jones *et al.*, 1997), FlexX (Rarey *et al.*, 1996) and DOCK (Ewing and Kuntz, 1997). In most cases, the binding site is predetermined and the search space is limited to the predetermined region in docking simulations. However, blind docking can also be performed efficiently in which a ligand is docked to a target without prior knowledge of the binding site (Hetényi & Spoel, 2002).

Docking protocols combine a search algorithm, to enumerate and test possible poses of the ligand in the protein's active site, and a scoring function to rank the binding affinity of each candidate ligand. Some common search algorithms are molecular dynamics, Monte Carlo methods, genetic algorithms, fragment-based methods, systematic searches, distance geometry methods, point complementary methods and tabu searches. Some common scoring functions used are force field methods, empirical free energy scoring functions and knowledge-based potential mean force.

2.4.2.3 Quantum mechanic/molecular mechanic (QM/MM) method

In recent years, computational chemists have become interested in chemical reactions of large molecular systems. Molecular dynamics (MD) simulations of large and complex biological systems based on molecular mechanics (MM) force fields have been extensively carried out. Examples are work done by Wahab *et al.* (2009) who performed molecular docking and MD simulation to study the binding of isoniazid onto the active site of *Mycobacterium tuberculosis* enoyl-acyl carrier protein reductase, and by Chipot *et al.* (2005) who applied MD simulation to investigate the structures and the dynamics of lipid aggregates. However, the MM force fields are incapable of describing the changes in the electronic structure of a system due to processes such as bond-breaking and bond-forming, and charge transfer. Modelling of these processes require quantum mechanics (QM) calculations, which are computationally demanding. Hybrid QM/MM methods allow the investigation of the chemistry of large systems with high precision to be possible (Vreven *et al.*, 2006).

A chemical reaction is often a local phenomenon. Hence, an expensive reliable method is required to describe the active region of the system. The atoms in the vicinity of the action centre participate in the reaction indirectly by distorting the electron distribution in the orbitals. This effect can be described by molecular orbital (MO) method to take into account its electronic effects. Atoms farther away from the action centre (the surrounding protein molecules) may provide some interaction, such as long-range electrostatic interaction or non-bonding steric interaction. These can be treated with inexpensive MO method or MM method (Morokuma *et al.*, 2001). Boundary treatment for the covalent bonds between the QM and MM regions is conventionally done by applying link atoms (usually hydrogen atoms, but sometimes halogens or methyl groups) (Singh and Kollmann, 1986). A link atom serves to cap the QM electron density at the QM/MM boundary. The link atom, however, should not distort the MM region, or introduce large interactions in the QM region (Eurenius *et al.*, 1996). Available approaches or codes which implement the hybrid QM/MM methods include ONIOM (Svensson *et al.*, 1996), DYNAMO, CHARMM interfaced with MOPAC, GAMESS-US or GAMESS-UK; and AMBER works with ROAR.