DETERMINATION OF R-3-HYDROXYACYL-ACP-COA TRANSFERASE (PhaG) STRUCTURE

by

HASNI BIN ARSAD

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

March 2010

ACKNOWLEDGEMENT

First of all, I would like to thank my supervisor Prof. Madya Dr. Mohd Razip Samian and my co-supervisor Prof. Madya Dr. Habibah A. Wahab for their guidance and full support during the course of this study.

I would also like to express my sincere gratitude to Prof. Dr. Nazalan Najimudin and Prof. Maqsudul Alam (CCB@USM) for their kind support. To Dr. Irene Newhouse and especially, Dr. James Newhouse my heartfelt thanks for his patient tutoring in Molecular Dynamics, helpful discussion and support, particularly in molecular dynamics simulation and access to bioinformatic tools and the Maui supercomputer at the Universiti of Hawaii.

To Prof. Nor Muhammad Mahadi (MGI), Prof. Aishah A. Latiff (DCC-USM), Prof. Dr. Raja Noor Zaliha Raja Abd. Rahman (UPM), Dr Aida, Dr. Rashidah, Dr. Ahmad Sofiman Osman, Dr. Mustapha Fadzil, Dr. Tengku Sifzizul Tengku Mohamad, Dr. Amirul Ashraf and Dr. Alex; my heartfelt thanks for their advice and permission to use their chemicals and lab apparatus.

I am grateful to the Ministry of Science, Technology and Innovation, Malaysia for providing me the National Science Fellowship. I am also grateful to MGI for support in terms of scientific activities and access to their supercomputer.

To members of Lab 414, Lab 406, Lab 409, Lab 318 and CCB; thank you for your direct or indirect contribution toward the accomplishment of my PhD study.

Last but not least, my heartfelt gratitude to Salina Mokhtar, Ameerah Husna Syafiqah, Ameer Syahir Hafidzi, Ameerah Hanim Syuhada and my parents (Arsad Hasan and Samsiah) whose continuous and unwavering support permit me to pursue my ambitions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURES

Å	Angstrom
α	Alpha
β	Beta
°C	Degree Celsius
%	Percentage
ϕ	Phi
ψ	Psi
3-D	Three dimensional
3HB-co-3HV	3-hydroxybutyrate-co-3-hydroxyvalerate
3HB	3-hydroxybutyrate
3HA	3-hydroxyacyl
ACP	Acyl carrier protein
ATP	Adenosine triphosphate
APS	Ammonium persulfate
CoA	Coenzyme A
DNA	Deoxyribonucleic acid
g	Gram
IEF	Isoelectric focusing
IPTG	Isopropyl β-D-thiogalactopyranoside
LJ	Lennard Jones

lcl-PHA	Long chain length-PHA
LB	Luria Bertani
MCL	Medium chain length
mcl-PHA	Medium chain length polyhydroxyalkanoate
mcl-3HA	Medium chain length 3-hydroxyalkanoates
μm	Micrometer
μl	Microliter
μM	Micromole
μg/ml	Microgram per milliliter
mg/l	Milligram per liter
mM	Millimole
ml	Milliliter
M	Molar
MD	Molecular dynamics
nm	Nanometer
NADPH	Nicotinamide adenine dinucleotide phosphate
TEMED	N,N,N',N'-tetramethylethylenediamine
N	Normality
NMR	Nuclear magnetic resonance
OD	Optical density
ps	Picosecond
PCR	Polymerase chain reaction
PHA	Polyhydroxyalkanoates

PDB	Protein Data Bank
psi	Pounds per square inch
RMSD	Root mean square deviation
rpm	Rotation per minute
scl-PHA	Short chain length-PHA
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
TAE	Tris-acetate/EDTA electrophoresis buffer
UV	Ultra violet
U	Units
V	Volt
V-hr	Volt per hour

PENENTUAN STRUKTUR R-3-HIDROKSIASIL-ACP-COA TRANFERASE (PhaG)

ABSTRAK

R-3-hidroksiasil-ACP-CoA tranferase (PhaG) berfungsi sebagai pemangkin penukaran (R)-3-hidroksiasil-ACP kepada (R)-3-hidroksiasil-CoA, sebagai prekursor utama untuk pempolimeran polyhydroxyalkanoate (PHA) daripada substrat tidak berkaitan. PHA adalah bioplastik yang berpotensi untuk menggantikan termoplastik berasaskan fosil kerana ianya boleh di biourai. Enzim PhaG daripada pencilan strain tempatan, *Pseudomonas* sp. USM 4-55, telah berjaya diklon (nombor akses Genebank EU305558). Pada masa ini struktur 3 dimensi yang sama dengan PhaG masih tidak diketahui. Untuk ekspresi lampau gen *phaG* telah diklon ke dalam vector ekspresi, pQE ke-30 dan telah berjaya diekspresi melalui aruhan menggunakan 0.5 mM IPTG di dalam sel perumah *Escherichia coli* SG 13009. PhaG berjaya dimurnikan dengan menggunakan kaedah kromatografi afiniti logam diikuti oleh kromatografi penurasan gel dan analisis Coomassie Blue SDS-PAGE menunjukkan satu jalur. Protein yang telah dimurnikan itu telah dipastikan sebagai PhaG melalui kaedah penjujukan peptide *de novo*. Nilai pI bagi $PhaG_{P.\ sp\ USM\ 4\text{-}55}$ yang ditentukan oleh IEF adalah 7.45. Penskrinan hablur protein telah dilakukan menggunakan kaedah "microbatch" dan "hanging drop". Malangnya, tidak satu pun daripada kit penghabluran yang digunakan menunjukkan hablur PhaG. Satu model 3-D PhaG dicadangkan dengan kaedah "threading". Protin struktur enzim 2-Hydroxy-6-okso-6-phenylhexa-2 ,4 hydrolase daripada *Rhodococcus* sp. RHA1 (identity PDB: 1C4X) digunakan sebagai template. Modeller9v4 digunakan untuk menghasilkan model $PhaG_{P.\ sp\ USM\ 4\text{-}55}$. Nilai RMS diantara model PhaG model 132 dan template, 1C4X adalah 0.75Å. Menariknya, analisis tapak pengikatan pada

PhaG model 132 yang telah ditentukan menggunakan SiteMap telah berjaya melabuhkan ligan R-3-hydroxyoctanoate-CoA di tapak pengikatan dengan Glide. Seterusnya mekanisme tindakbalas enzim PhaG dicadangkan berdasarkan analisis docking, dan mekanisme enzim malonil CoA-ACP transakilase (FabD).

DETERMINATION OF R-3-HYDROXYACYL-ACP-COA TRANSFERASE (PhaG) STRUCTURE

ABSTRACT

R-3-hydroxyacyl-ACP-CoA transferase (PhaG) catalyzes the conversion of (R)-3-hydroxyacyl-ACP to (R)-3-hydroxyacyl-CoA derivatives, which serve as the ultimate precursor for polyhydroxyalkanoate (PHA) polymerization from unrelated substrates. PHA is a family of bioplastic that has a good potential to replace fossil-based thermoplastics because it is biodegradable. The transferase enzyme PhaG of a locally isolated strain, *Pseudomonas* sp. USM 4-55, was recently cloned (GeneBank accession number EU305558). Currently there is no known 3D structure with high similarity to PhaG. In order to over express, the phaG gene was cloned into expression vector pQE-30 and it was successfully overexpressed by induction with 0.5 mM IPTG in the host *Escherichia coli* strain SG 13009. The PhaG was purified by metal affinity chromatography followed by gel filtration chromatography to a single band under Coomassie Blue SDS-PAGE analysis. The purified protein was confirmed as PhaG by *de novo* peptide sequencing. The pI of $PhaG_{P.\ sp\ USM\ 4\text{-}55}$ is 7.45, as determined by IEF. Protein crystallization screening was then performed by applying microbatch and hanging drop method. However, none of the crystallization screening kit used were able to produce PhaG protein crystal. A 3-D model of PhaG was predicted using threading method. The 2-hydroxy-6-oxo-6-phenylhexa-2,4-dienote hydrolase enzyme from *Rhodococcus* sp. strain RHA1 (pdb id: 1C4X) structure was used as the template. The Modeller9v4 program was used to model the $PhaG_{P.\ sp\ USM\ 4\text{-}55}$. The RMS value of the PhaG model 132, and the template 1C4X is 0.75Å. Interestingly, binding site prediction using SiteMap on the proposed model, showed a binding pocket that can accomodate the ligand, R-3-hydroxyoctanoate-

CoA by using Glide. Consequently, the reaction mechanism of PhaG is proposed based on the docking analysis and the mechanism of malonyl CoA-ACP transacylase (FabD).

CHAPTER 1

1.0 INTRODUCTION

Polyhydroxyalkanoate (PHA) is a kind of bioplastic that has a good potential to replace fossil-based thermoplastics due to its biodegradable properties. PHAs are normally found in various microorganisms, accumulated as intracellular granules, that can serve as carbon and energy storage material under nutrient-limiting conditions (Dawes and Senior, 1973).

PHA can be divided into three groups depending on the number of carbon atoms in the monomer units. The first two groups are short-chain-length PHA (scl-PHA), which consist of 3 – 5 carbon atoms and medium-chain-length PHA (mcl-PHA), which consist of 6 – 14 carbon atoms. The third group, long-chain-length PHA (lcl-PHA), is reserved for polymers containing more than 14 carbon atoms (Steinbüchel *et al.*, 1992).

Steinbüchel (2001) reported that there are approximately 150 different hydroxyalkanoic acids known to be constituents of bacterial storage polyester (PHA). Bacterial polyesters such as the homopolyester poly(3HB), the copolyester poly(3HB-co-3HV) and PHAs consisting of 3-hydroxyoctanoate, 3-hydroxydecanoate and a few other medium-chain-length 3-hydroxyalkanoates (poly(mcl-3HA)) have been manufactured and used in various applications. Byrom (1994) reported the commercial production of poly(3HB) and poly(3HB-co-3HV) as biodegradable bioplastics. PHAs are also used in product applications such as latex paints (van der Walle *et al.*, 2001), medical application such as retard material (Fraser *et al*., 1989) and as scaffolding material for tissue engineering (Williams *et al*., 1999).

The unique and interesting properties of PHAs attracted academic and commercial bodies to embark on extensive research to study this biopolymer. Biosynthesis of PHA, especially in the bacterial cell was investigated by a few groups of scientist. A diverse range of bacteria are able to accumulate PHA, for example almost all *Pseudomonads* synthesize poly(mcl-3HA) when cultured on alkanes, organic acids, glucose and many other carbon sources (Haywood *et al.*, 1990, Timm and Steinbuchel, 1990).

Huijberts *et al.* (1994) and Rehm *et al*. (1998) found that there are at least three different metabolic routes in *P. putida* for the synthesis of 3-hydroxyacyl coenzyme A, the substrate for PHA synthase to synthesize PHA. They are 1) beta-oxidation pathway, 2) fatty acid *de novo* biosynthesis pathway and 3) chain elongation reaction pathway.

A further investigation on PHA synthesis by fatty acid *de novo* revealed that R-3-hydroxyacyl-acyl carrier protein-Coenzyme A transferase (PhaG) was the enzyme responsible to channel down substrates from fatty acid *de novo* biosynthesis pathway to the PHA synthase for PHA accumulation in *P. putida* (Rehm *et al*., 1998). The evidence showed that PhaG catalyzes the conversion of (R)-3-hydroxyacyl-ACP to (R)-3-hydroxyacyl-CoA derivatives which serve as the ultimate precursor for PHA polymerization from unrelated substrates.

Van der Leij and Witholt (1995) were the first to allude to the connection between *de novo* fatty acid biosynthesis and PHA production in plant. They predicted the existence of a transferase enzyme that was needed to convert (R)-hydroxyacyl-ACP available in the plant fatty acid biosynthesis pathway to the (R)-hydroxyacyl-CoA, which is the substrate for PHA synthase.

In this thesis, the PhaG from *Pseudomonas* sp. USM 4-55 ($PhaG_{P.\ sp.\ USM\ 4\text{-}55}$) was used as the subject. This strain was isolated from soil samples taken from Tasek Chini, Pahang (Few, 2001). This isolate was chosen for this thesis because of its rare characteristic; it was able to accumulate two groups of polymer at the same time, namely scl-PHA and mcl-PHA. Previous studies showed that *P.* sp USM 4-55 could accumulate mcl-PHA up to 20 % of cell dry weight when cultured on 2 % glucose (Few, 2001). The result of the study mentioned above indicates that the PhaG in this strain is functional as described by Rehm *et al*., (1998). PhaG plays an important role in channeling the substrates for PHA synthase when glucose or gluconate is used as the carbon source (Rehm *et al.*, 1998). This enzyme is usually present in *Pseudomonads*. The unique and interesting characteristic of PhaG attracted me to embark on an effort to determine its structure and understand the functional mechanism of this enzyme.

1.1 Research Objectives

1. To overexpress and crystallise PhaG
2. To model 3-dimensional structure of PhaG
3. To propose a reaction mechanism of PhaG

1.2 Thesis overview

The thesis is divided into four chapters. It starts of with a literature review of protein structure in general and PhaG in the chapter 1. The second chapter involves the overexpression of $PhaG_{P.\ sp\ USM\ 4\text{-}55}$ in *E. coli*. Details about cloning, expression and purification, pI determination and protein crystal screening will be elaborated.

Chapter three is an analysis to understand the three dimensional structure of this protein by employing bioinformatics tools to build and predict the protein model. In addition, the reaction mechanism of PhaG is also proposed. Chapter four is a general discussion and the suggestion for the future studies.

1.3 Literature review

1.3.1 Protein structure:

Protein has biologically diverse functions in the cell ranging from DNA replication, forming cytoskeletal structures, transporting oxygen around the bodies of multicellular organism to converting one molecule into another. Lately, the studies of protein structures and functions have increased. This is due to the current trend to understand of protein interactions with other biomolecules and their role within different biological systems and their potential manipulations by employing genetic or chemical methods.

Biochemists have classified protein structures into four different classes, namely primary structure, secondary structure, tertiary structure and quartenary structure. The primary protein structure is the amino acid sequence in a linear heteropolymer amino acid chain. Amino acid is consisted of atoms C, O, N, H and S. The basic structure of amino acid is shown in Figure 1.1. A carbon atom (Cα) is bound to an amino group ($-NH_2$), a carboxyl group (-COOH), a hydrogen atom (H) and an organic side group (assigned as R). A statistical survey of x-ray structures in the Cambridge Structure Database found that the average of C-N peptide bond length is 1.33 Å, except for proline residue, which is 1.34 Å; the C-O bond length is 1.23 Å, C_α-C bond length is 1.53 Å except glycine residue, which is and 1.52 Å,

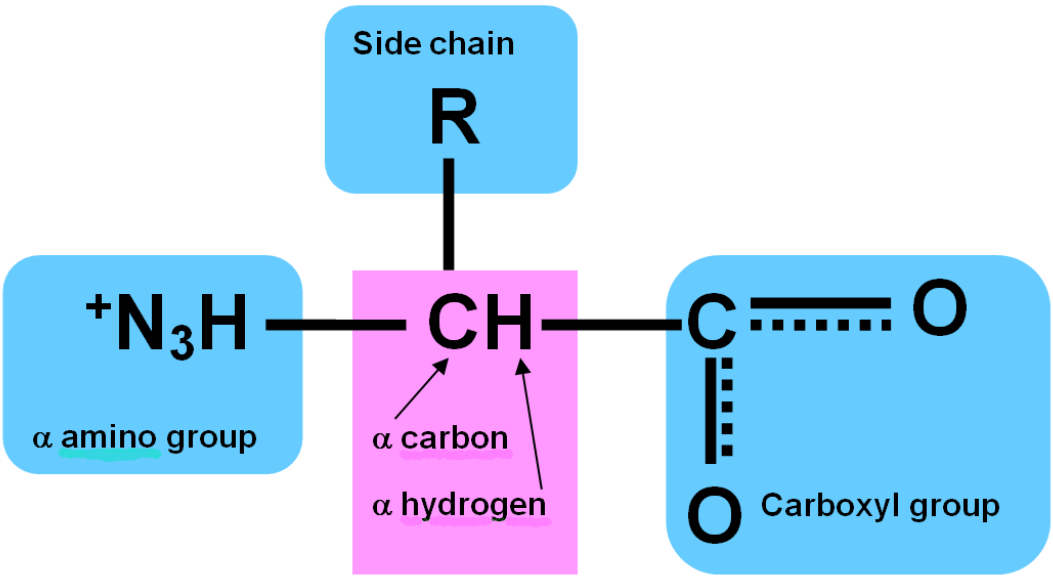

Figure 1.1. Basic structure of an amino acid.

N-C_α bond length is 1.46 Å except for glycine and proline, which is 1.45 Å and 1.47 Å respectively (Engh and Huber, 1991). The average value of the bond lengths is shown in Figure 1.2.

The dihedral angle is also important in the protein conformation analysis. The dihedral angles for the backbone of the peptide are assigned as ϕ and ψ angle for the N - Cα bond and Cα - C bond accordingly (Figure 1.3). Both parameters are used for the protein structure validation especially in the Ramachandran plot analysis as described by Ramachandran *et al*., (1963) (Figure 1.4). Amino acids with torsion angles in the range $-180 < \phi < 0^\circ$, $-100 < \psi < 45^\circ$ are considered to be in the α-helical region. Amino acids with torsion angles in the range $-180 < \phi < -45^\circ$, $45^\circ < \psi < -45^\circ$ are considered as β-sheet region. The turn region are in the range $0^\circ < \phi < 180^\circ$ and $-90^\circ < \psi < 90^\circ$.

The analysis of 237,384 amino acids in 1,042 protein subunit from the PDB by Hovmoller *et al*., (2002) found that the average values of two torsion angles, ϕ and ψ of amino acids in the α-helix region is within $\pm 2^\circ$ of -63.8° (ϕ) and -41.1° (ψ). The average torsion angle is $\phi = -122^\circ$ and $\psi = +136^\circ$ for antiparallel β-sheets and $\phi = -116^\circ$, $\psi = +128^\circ$ for parallel β-sheets (Hovmoller *et al.*, 2002).

The properties of the R group will give the unique physical and chemical properties to each amino acid. The R side chain structure and their three-letter and one-letter code are shown in Table 1.1. Amino acids can be classified according to their physical, chemical and the structure properties. One such classification was reported by Taylor (1986), Figure 1.5.

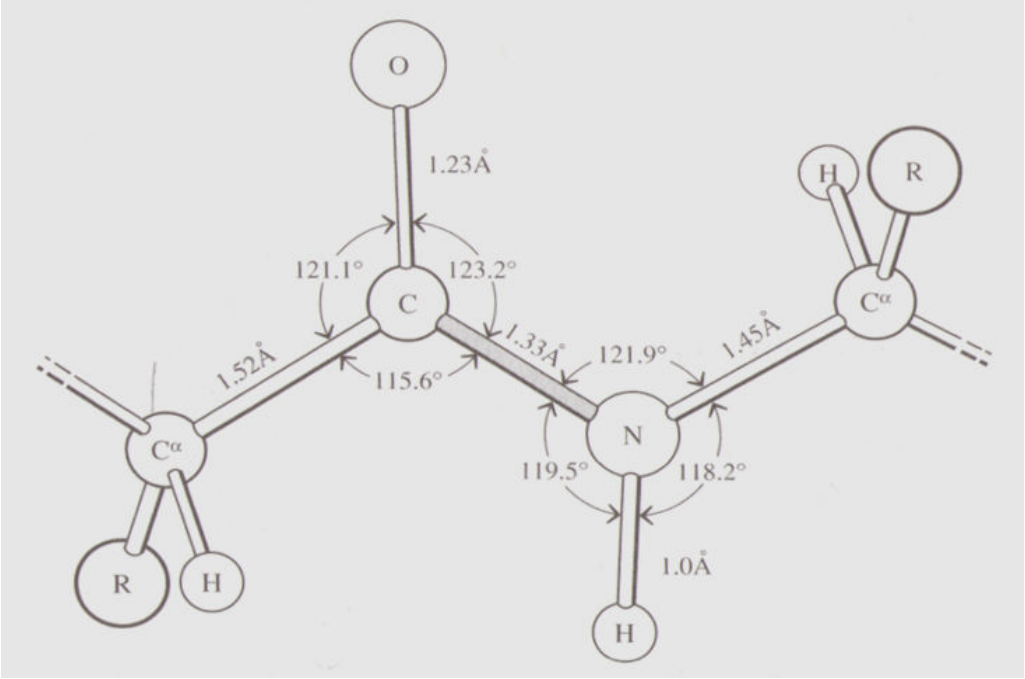

Figure 1.2. Geometry of the peptide backbone. The dimensions given are the averages observed crystallographically in amino acids and small peptides (Creighton, 2002).

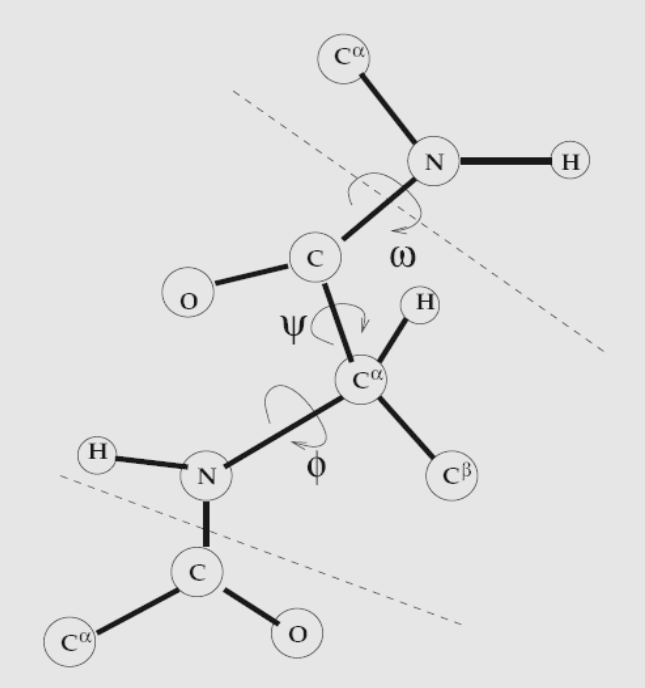

Figure 1.3. Ramachandran plot angles. The ϕ and ψ angle used in the peptide dihedral angle analysis.

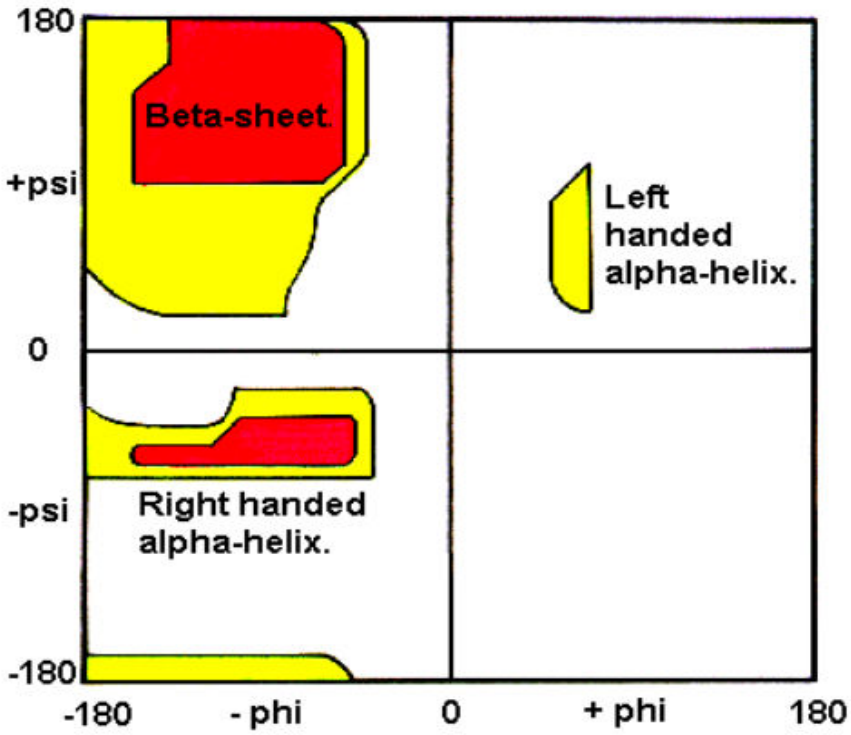

Figure 1.4 The Ramachandran plot showing the dihedral angle at N-Cα-C, i.e. the ϕ and ψ angle for the peptide backbone in the protein structure. The red color represents fully allowed region and the yellow color represents the outer limit of the ϕ and ψ angle.

Table 1.1. Standard amino acid abbreviation and side chain properties

AMINO ACID	3-LETTER	1-LETTER	SIDE CHAIN
Alanine	Ala	A	CH_3-
Arginine	Arg	R	HN=C(NH_2)-NH-(CH_2)$_3$-
Asparagine	Asn	N	H_2N-CO-CH_2-
Aspartic acid	Asp	D	HOOC-CH_2-
Cysteine	Cys	C	HS-CH_2-
Glutamine	Gln	Q	H_2N-CO-(CH_2)$_2$-
Glutamic acid	Glu	E	HOOC-(CH_2)$_2$-
Glycine	Gly	G	H-
Histidine	His	H	N=CH-NH-CH=C-CH_2- |_____|
Isoleucine	Ile	I	CH_3-CH_2-CH(CH_3)-
Leucine	Leu	L	(CH_3)$_2$-CH-CH_2-
Lysine	Lys	K	H_2N-(CH_2)$_4$-
Methionine	Met	M	CH_3-S-(CH_2)$_2$-
Phenylalanine	Phe	F	Phenyl-CH_2-
Proline	Pro	P	-N-(CH_2)$_3$-CH |_____|
Serine	Ser	S	HO-CH_2-
Threonine	Thr	T	CH_3-CH(OH)-
Tryptophan	Trp	W	Phenyl-NH-CH=C-CH_2- |_____|
Tyrosine	Tyr	Y	4-OH-Phenyl-CH_2-
Valine	Val	V	CH_3-CH(CH_2)-

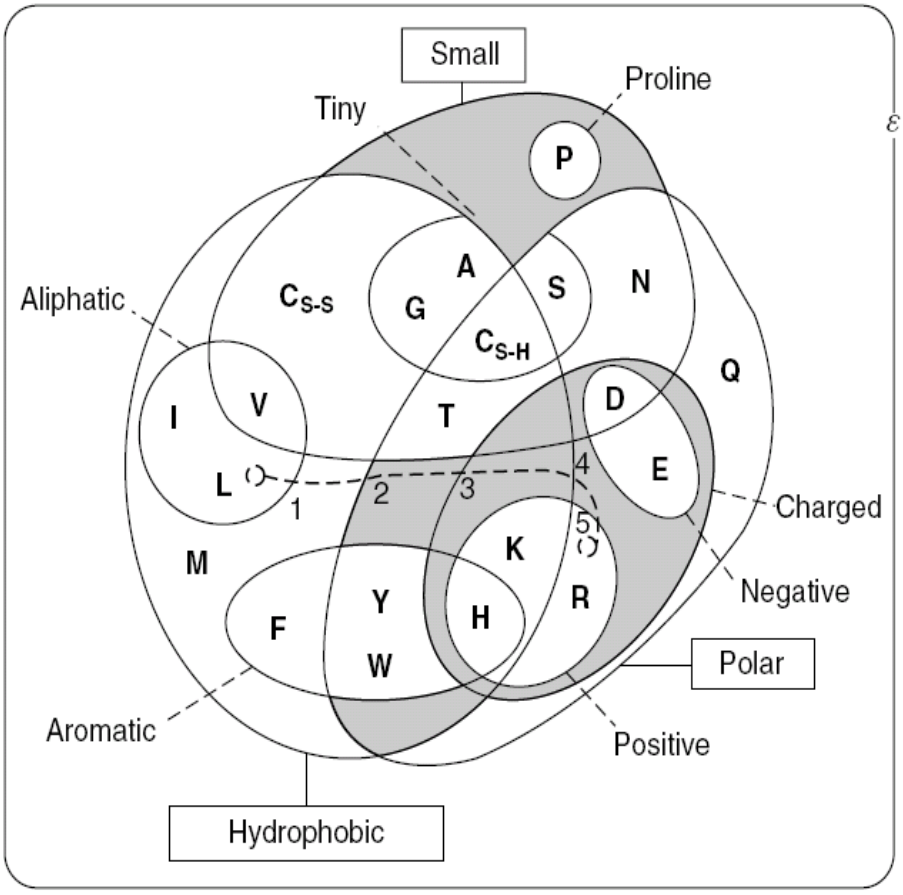

Figure 1.5. A Venn diagram illustrating the properties of amino acids (Taylor, 1986, Betts and Russell, 2003)

Most commonly, amino acids are classified according to their compatibility in aqueous environment – those that are compatible are classified as hydrophilic and those that do not, are classified as hydrophobic. The hydrophobic amino acids can be divided according to their aliphatic or aromatic side chains. Aliphatic side chains are non-reactive and are rarely involved in protein function. The aliphatic side chain amino acids are alanine, isoleucine, leucine, proline, and valine. A side chain is classified as aromatic when it contains an aromatic ring system, which is planar and the electrons are shared over the whole ring structure. The aromatic amino acids are phenylalanine, tryptophan, tyrosine and histidine.

The amino acids that can be surrounded by water have polar side chains and those buried in the protein structure usually form hydrogen bonds with other amino acids, either with other side chain or the main chain. Examples of of this group of amino acids are aspartic acid and glutamic acid are negatively charged. Lysine and arginine are positively charged. Histidine, asparagines, serine, threonine and tyrosine are all neutral.

A group of amino acid is subcategorized as small and tiny, grouped together according to their small side chain (Betts and Russell, 2003). These are alanine, cysteine, glycine, proline, serine and threonine.

The amino acid side chain properties are also used in the qualitative description of each position in a multiple alignment and the information could also be used for protein engineering in identifying the alternative amino acid that could be used as a replacement at any position for specific application. As an example, the substitution of a small side chain for a large side chain can be disastrous to the protein structure (Betts and Russell, 2003).

Betts and Russell (2003) also reported that the polar side chain such as tyrosine, histidine, arginine, lysine, aspartate, glutamate, asparagines, glutamine, serine, threonine and cystine residues are frequently found in the protein active or binding site. This shows that each amino acid, with varying side chain property plays an important role in the protein structure as well as in enzymatic reaction mechanism.

1.3.1.1 Primary structure

The primary structure is the linear order of amino acid residues along the polypeptide chain. The heteropolymer chain may contain any number and combinations of the 20 amino acids. The linear amino acid sequence is formed by covalent linkages of individual amino acids via peptide bonds. Covalent bonding between two amino acids is formed by the carbon atom from the carbonyl group of one amino acid sharing electrons with the nitrogen atom from the amino group of a second amino acid and a molecule of water is formed during this process (Alberts *et al.*, 1994). The diverse number of amino acid combinations in amino acid sequences is the basis of protein structure and function diversity.

Amino acid sequence in the protein molecule plays an important role in protein folding. As we know, the 20 amino acid side chains vary in their physico-chemical properties. Grigoriev and Kim (1999) studied the amino acid properties to identify the fold similarity between two proteins which had 15% amino acid sequence identity. The physical properties of neighboring amino acid residues in sequence at structurally equivalent position of two proteins of similar fold are often correlated even though the amino acid sequences are different (Grigoriev and Kim, 1999). Current bioinformatic tools can predict the protein secondary and tertiary

structure when the amino acid sequences of the proteins are available, because the pattern of amino acid sequence will give specific structure according to their side chain properties (Hung and Samudrala, 2003).

Organisms that live in the different temperature optima (mesophiles and thermophiles) differ in their amino acid composition and have their own protein repertoire (Table 1.2). This is to maintain the protein structure in their specific environment (Jaenicke, 2000). Analyzing amino acid sequences, enables us to understand the protein in its natural environment

1.3.1.2 Secondary structure

The secondary structure is the local folding of the polypeptide chain. The spatial relationship of amino acid residues that are closed together in the primary sequence tend to form specific secondary structures.

In 1951, the secondary structure of protein was clearly understood with the availability of the X-ray protein crystallography data. The protein secondary structure is comprised of helices, beta sheets and extended structures (Pauling and Corey, 1951, Pauling *et al.*, 1951). The interactions of the side chains and main chains result in the formation of secondary structures such as alpha helices (Figure 1.6), beta sheets (Figure 1.7) and turns of the protein.

The α-helix structure is formed by the hydrogen bond between the oxygen of the carbonyl group (C=O) of one turn and the hydrogen of the amide group (N-H) in the neighboring turn. One turn is formed by 3.6 residues. The distance between two turns is 0.54 nm. The helix structure can be right-handed or left-handed.

Table 1.2. Relative amino acid compositions of mesophiles and thermophiles One-letter abbreviations of amino-acid residues in brackets

	Mesophiles	Thermophiles
Charged residues (DEKRH)	24.11%	29.84%
Polar:uncharged residues (GSTNQYC)	31.15%	26.79%
Hydrophobic residues (LMIVWPAF)	44.74%	43.36%

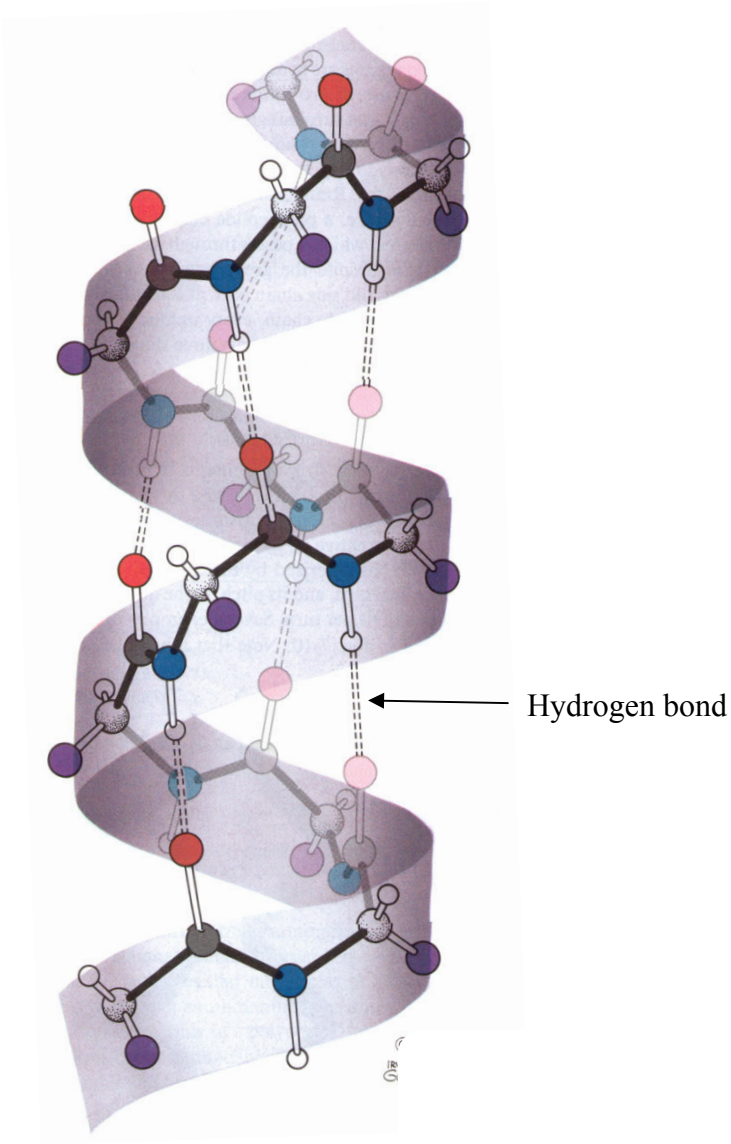

Figure 1.6. A right-handed alpha helix. Hydrogen bonds between the N – H groups and the C=O groups that are four residues back along the polypeptide chain are indicated by dashed lines (Voet and Voet, 1995, Orengo *et al.*, 1997).

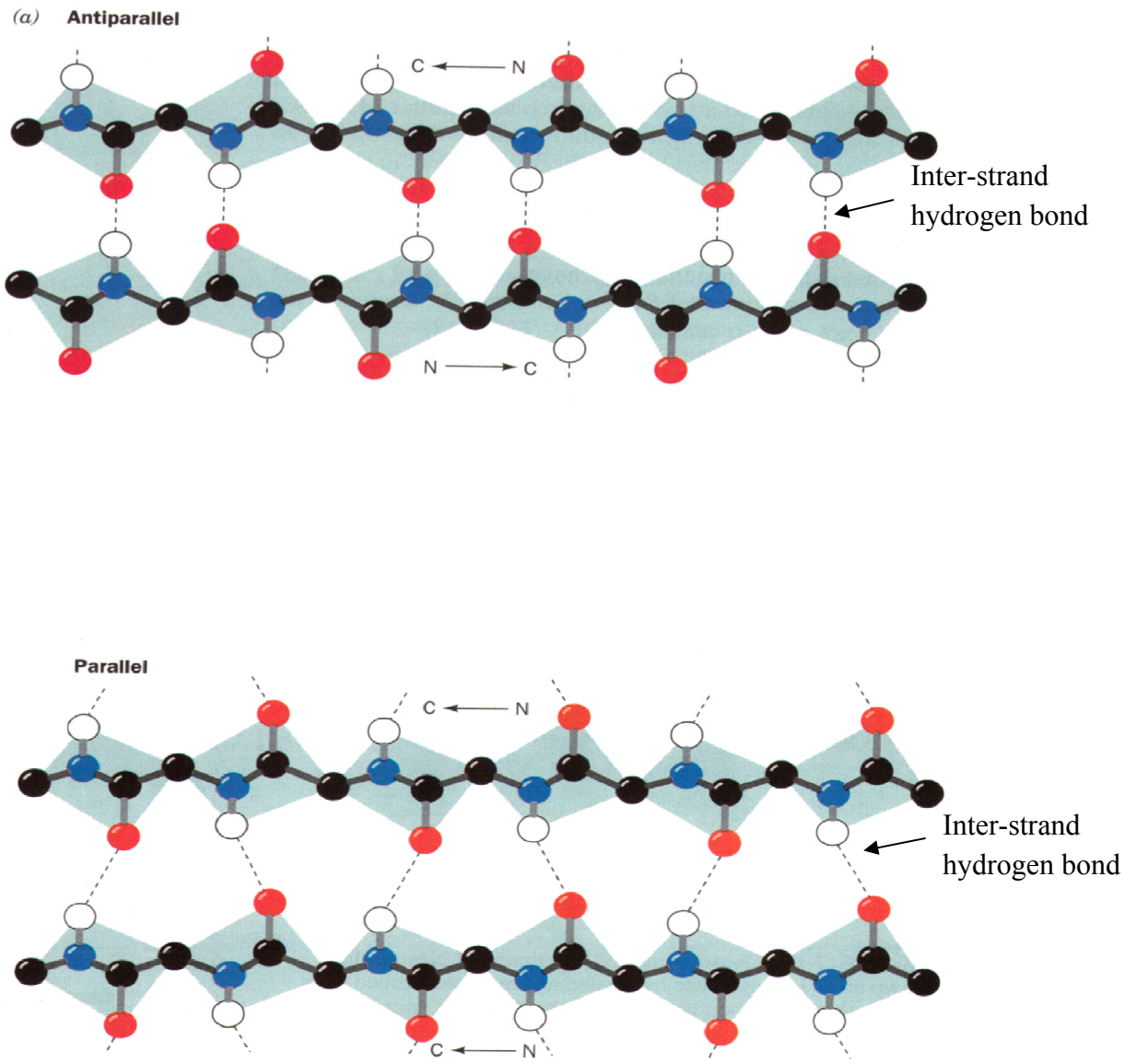

Figure 1.7. Two forms of β-sheets are commonly found in secondary structure, namely antiparallel and parallel β-sheets. The hydrogen bonding between the oxygen and hydrogen atoms at the carbonyl group and amino group are shown as dotted lines (Voet and Voet, 1995).

The right handed α-helix structure is the most identifiable unit of the secondary structure. The ideal value of the dihedral angles for α-helix structures is -57º and -47º of ϕ and ψ angles respectively.

In globular proteins, helices are stable in short structure. The average length is about 12 residues and the most probable helix length is less then 6 residues. Helix structure is stable in globular protein and remains 100% helical up to temperature near the denaturation point (Dill, 1990, Kabsch and Sander, 1983).

Beta strand is the second unit of secondary structures identified by Pauling and Corey (1951). It exists as an extended conformation. Two or more β strands that are located close to one another tend to form additional hydrogen bonds, which stabilizes the structure. A single β strand is not stable. Hydrogen bonds are formed between the oxygen of a carbonyl group in amino acid of one strand and a nitrogen in the backbone of a second amino acid of another strand (Figure 1.7).

Beta-sheets can be divided into two types, the parallel or anti-parallel. A parallel beta-sheet is formed when the amino acid terminal residue of each strand points in the same direction and the anti-parallel sheet is formed when the amino termini are pointing in opposite directions. The parallel β-strand has dihedral angles of -139º (ϕ) and +135º(ψ) respectively, while the anti-parallel β-strand has dihedral angles of -119º(ϕ) and +113º(ψ) respectively (Whitford, 2005).

Turns are one of the elements of secondary structure. Turn conformations play a very important role in protein structure as it enables the polypeptide to change directions in the protein structure. Residues with small side chains such as glycine, aspartate, asparagines, serine, cysteine and proline are dominant in this structure (Whitford, 2005). These helices and sheets are then assembled into a compact structure recognized as tertiary structure.

1.3.1.3 Tertiary structure of proteins

The tertiary protein structure is the final specific geometric shape of the protein which represents the folds of the polypeptide chain. Different fragments of secondary structure in the same chain become bonded together. Levitt and Chothia (1976) were classified globular proteins into four groups according to their secondary structure elements. The four protein classes are all α-helix, all β-sheet, α/β-fold and α+β fold protein. The all α and all β globular protein dominantly contain either α-helices and β-sheets in their secondary structures. The α/β proteins containing both α-helices and β-sheets which mostly show βαβ unit repeats which consist of two adjacent β strands connected by a single α helix (Russell, 2000). The α+β globular proteins exists as α-helix and β-strand secondary structure segments that do not mix. The α-helix and β-strand are segregated in the polypeptide chain, a mixture of all α and all β-sheets regioms within the same polypeptide chain (Michie *et al.*, 1996).

The driving forces responsible for protein folding comes from two types of energy. The first is the bonded interaction, which includes the sum of bond, bond angle, dihedral angle and torsional angle. The second is nonbonded interaction such

as van der waals and electrostatic interaction between all pairs of atom (Oostenbrink *et al.*, 2004, Jaenicke, 2000, Dill, 1990). The bonded and nonbonded energy are all considered when using force field algorithm for computational protein structure prediction.

In protein folding, the secondary structure element and the hydrophobic interaction, which is classified as nonbonded interaction in force field algorithm is hypothesized as the major driving force in the formation of tertiary structure. The secondary structure forms first, and the chain segment then pack together by hydrophobic force (Dill, 1990) to give the final tertiary structure (Levitt, 1976). The other force fields (bonded and nonbonded interaction) also contribute in tertiary structure stability.

1.3.1.4 Quartenary structure

The quarternary structure is formed when several protein molecules interact together. A multi-subunit protein maybe composed of two or more identical polypeptides or different polypeptides. The tertiary structure aggregates to form homo or hetero-multimers (Henrick and Thornton, 1998). Quartenary structures are stabilized by interactions between residues exposed on the polypeptide surfaces within a complex. These interaction include disulfide bridges, hydrophobic interactions, charge-pair interaction and hydrogen bonding (Whitford, 2005).

1.3.2 Protein structure determination

Currently, more than 55,000 sets of atomic coordinates of protein structure are deposited in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb) (September 2009) and the numbers are increasing daily. The protein structures were mostly

determined by X-ray protein crystallography and some of them were determined by nuclear magnetic resonance (NMR) spectroscopy.

Even though the number of protein structure deposited in the PDB keep increasing, many protein structures cannot be solved experimentaly using X-ray crystallography and NMR, due to the limitations of each method. In such cases, some protein structures can be predicted from amino acid sequences using computational methods. This domain is also known as Structural Bioinformatics.

1.3.3 Protein structure database

The current single worldwide depository for experimentally solved three dimensional structure of biological macromolecules is the Protein Data Bank (PDB) (Berman *et al*., 2000). The PDB was established in Brookhaven National Laboratories in 1971. As many as 55,285 protein structures were deposited in PDB as of September 2009.

Information on protein structures can also be found in the Universal Protein Resources (UniProt), The Structural Classification of Protein (SCOP) and a semi-automatic hierarchical domain classification of protein structures in PDB (CATH) (Orengo *et al.*, 1997).

The UniProt contains information on proteins sequences and their function, and is created by combining information from Swiss-Prot and Protein Information Resources (PIR) databases (Jain *et al.*, 2009, Barker *et al.*, 1999). SCOP shows the structure and evolutionary relationship between all protein structures by manual identification (Murzin *et al*., 1995). Protein structures with high structural similarity are classified in the same ‘fold’. The term ‘Superfamily’ is used for proteins which probably have common evolutionary origins, whereas the term ‘Family’ is used for proteins that have clear evolutionary relationships.

CATH is a database of manually curated classification of protein domain structures. Protein structures are classified using a combination of automated and manual procedures. The four major levels in this classification are class, architecture, topology and homologous superfamily (Cuff *et al.*, 2009, Orengo *et al.*, 1997).

1.3.4 Protein structure prediction

The fundamentals of structural bioinformatics involved the prediction of three-dimensional protein structure from its primary sequences. The main methods currently applied to predict protein structures are homology modeling, fold recognition and *ab initio* prediction (Jones, 2000).

Zhang (2008) divided protein structure prediction to two main methods. The first is a template-based structure prediction, which requires a protein structure that has at least weak homologies to the protein of interest. The second method is a free modeling method, which does not provide appropriate templates for the prediction. In this case, the protein model is built from scratch and is called as *ab initio* or *de novo* protein structure prediction.

The simililarity of the query sequence to an experimentally determined protein structure can be used as an indicator in making a decision to decide which method is to be used to predict the structure of an unknown protein. The sequence identity between the target and the template can be categorized into three classes. The first is the high homology category where the two protein sequences have 50% or above sequence identity. In the second category, the sequence identity is between 20 to 30%. This range is also known as the ‘twilight zone’ of sequence identity. The third is the ‘midnight zone’ for low sequence identity, covering the range of 8 to 10% (Rost, 1999, Rost, 1997, Chung and Subbiah, 1996).

The direct homology modeling can be applied if the target and its template amino acid sequence is 50% or greater because both structures normally share folds in the tertiary structure with a variation of 1Å in its backbone RMSD value (Gerstein and Levitt, 1998). If the amino acid sequence identity between the target and its template is below 30%, the fold recognition or threading-method can be used to assign the correct fold to the target sequence (Jones *et al.*, 1992). For proteins with a low amino acid sequence identity when compared to available determined structures, the folding cannot be assign by threading method. In these cases, the *ab initio* method can be applied in order to assign a new fold of protein (Hardin *et al.*, 2002, Bonneau and Baker, 2001, Bonneau *et al.*, 2002).

1.3.4.1 Homology modeling

Homology modeling is the prediction of an unknown protein structure and mapping it to a known structure, usually where both proteins have evolutionarily relationship and show high amino acid sequence similarity. Proteins with high sequence similarity share similar structures. This method is usually applied to proteins which have 50% or above amino acid sequence identity to proteins with solved three dimensional (3-D) structures in the PDB.

Once a suitable template for homology modeling is identified, Modeller is used to construct three-dimensional models (Sali and Blundell, 1993, Sali and Overington, 1994, Eswar *et al*., 2006, Eswar *et al.*, 2007, Marti-Renom *et al.*, 2000). Modeller performs comparative modeling by predicting the 3-D structure of a given protein sequence (target) based primarily on its alignment to one or more known structures of proteins (templates). The prediction processes involves fold assignment, target-template alignment, model building and model evaluation.

1.3.4.2 Fold Recognition

The fold-recognition approach in protein structure prediction involves the identification of experimently determined protein structures that can accommodate the target protein sequence. The structure based fold recognition is also known as the threading method. Threading works by computing a scoring function that performs the best fit of a sequence against an experimentally determined structure. Typically, a threading computer program comprises of four components. First is the template structure, which is usually a protein structure deposited in the Protein Data Bank database. Second is the evaluation of the compatibility between the target sequence and the template fold. This is followed by the algorithm to compute the optimal alignment between the target sequence and the template structure. Finally the statistical significance is estimated based on the ranking (Fischer *et al.*, 1996).

The target-template compatibility functions typically using two types of structural feature, the pattern of secondary structure element and local environment classes. The local environment classes are the combination of solvent accessibility, polarity of the side chain environment and local backbone conformation (Rost *et al.*, 1997).

Cymerman *et al*., (2004) defined the fold-recognition algorithms to two main types namely the sequence based fold recognition and structure based fold recognition method. The principle of the sequence based fold recognition method is to search structurally characterized proteins from the protein structure database that exhibited significant sequence similarity to the target protein. This step is done by using BLAST or PSI-BLAST in order to identify suitable template from the protein data bank.