COMBINED KNOWLEDGE AND PHYSICS BASED APPROACHES TOWARDS PREDICTING PROTEIN STRUCTURES

NURUL BAHIYAH AHMAD KHAIRUDIN

UNIVERSITI SAINS MALAYSIA
2009

COMBINED KNOWLEDGE AND PHYSICS BASED APPROACHES TOWARDS PREDICTING PROTEIN STRUCTURES

by

NURUL BAHIYAH AHMAD KHAIRUDIN

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

December 2009

PENDEKATAN TERGABUNG BERDASARKAN PENGETAHUAN DAN FIZIK KE ARAH PERAMALAN STRUKTUR PROTEIN

oleh

NURUL BAHIYAH AHMAD KHAIRUDIN

Tesis yang diserahkan untuk
Memenuhi keperluan bagi
Ijazah Doktor Falsafah

Disember 2009

… If this doesn't make you free
It doesn't mean you're tied

If this doesn't take you down
It doesn't mean you're high

If this doesn't make you smile
You don't have to cry

If this isn't making sense
It doesn't make it lies …

To

My precious afiq

ACKNOWLEDGEMENTS

I have always thought that this day would never come, and now that it finally has, I would like to thank those who have helped me either directly or indirectly along the way. First and foremost, I would like to express all the praises and greatest gratitude to Allah SWT, the Most Gracious, Most Merciful, and Holder of all knowledge for giving me the chance and strength to complete this study and for all the opportunities He has given to me until now. Alhamdulillah.

Not a single accomplishment would be possible without my family, who started it all. Their love and support forms the heart for everything I do. I thank my parents for raising me, filled my life with love and educating me to be independent and for letting me choose my own paths. I cannot be more indebted to my mother for all the many sacrifices she made and for her unconditional love and endless support she has showered me each and every single day of my life, without which this thesis would never have come into being. Thanks mak for persevering with me every now and then. To my brothers and sisters nana, abang, che pi and kak ina, thank you for always being there and constantly praying for my success, especially abang and kak ina for willingly being my guarantor.

I was so blessed to have the opportunity to spend my five years of graduate study in Universiti Sains Malaysia, Penang. It was a great honour for me to have Assoc. Prof. Dr. Habibah A Wahab as my main supervisor, a humble lady with a brilliant mind and a strong scientific drive. She has shaped my personality greatly in the past five years and has faithfully guided me along the rugged path. Thank you Dr Habibah for your kindness, generosity, optimism and support whenever I needed them. I could never forget your patience in reading through my manuscripts and improving my English. It was an honor and pleasure to work with you. A special thanks also goes to my other two co-supervisors, Prof. Dr. Nazalan Najimudin and Assoc. Prof. Dr. Mohd. Razip Samian for providing a great mentorship, for being there whenever I needed them and for sharing their insightful wisdom for which I deeply respect them. It has truly been a learning experience.

Thanks also to folks in the simulation lab (most of them have graduated), Rashid, Imtiaz, Fitriah, Ezatul, Nani, En. Achsan, Wai Keat, Sek Peng, Sohail, Belal, Sy Bing, En Razip, Yee Siew, Hamdah, En. Ismail and most importantly, I thank Su for putting up a lot of time and effort into making HAWK smoothly available to the group. I thoroughly enjoyed all the discussions, laughter, chit chatting, delicious lunch hours, birthday celebrations, bowlings, aquarium cleaning and all the great times. It’s been a really roller coaster for all of us.

I would also like to express my gratitude to Universiti Teknologi Malaysia, for funding my study. My thanks also goes to MIMOS Berhad for providing high-speed computational facilities to do my simulations.

Last but not least, my deep appreciation also goes to my beloved husband, *sharul* whom I met a few months after I have completed my work and writing, and our eight-year-old baby boy *afiq* who was born during my thesis correction. To *sharul*, thank you for kept on pushing me and gave me strength to finish up my corrections. To *afiq*, thank you for the absolute joy you have brought for mama and ayah. I sincerely thank Allah for bringing both of you into my world, which has added extraordinary beauty to my life.

Finally I would like to thank all those who were involved, suggested, taught and participated: You have enriched my life tremendously. It was a great privilege. May Allah SWT grant you the best reward for all of your kindness and help. Thank you.

TABLE OF CONTENTS

Page

Acknowledgements ii

Table Of Contents iv

List of Tables ix

List of Figures xi

List of Symbols xvii

List of Abbreviations xix

List of Publications and Seminar xxi

Abstrak xxii

Abstract xxiii

Preamble xxiv

CHAPTER ONE - INTRODUCTION 1

1.1 The Biochemistry of Proteins 1

1.1.1 The Peptide Bonds 2

1.1.2 Limitations on Conformations 2

1.1.3 Structural Hierarchy 3

1.1.4 Driving Forces 4

1.2 The Protein Folding Problem 5

1.2.1 Comparative Modeling 6

1.2.1.1 Template Selection 7

1.2.1.2 Sequence Alignment 7

1.2.1.3 Model Building 8

1.2.2 Fold Recognition 9

1.2.3 *ab initio* Method 10

1.3 Fundamentals of Molecular Dynamics 1

1.3.1 The Force Field 14

1.3.2 The Equation of Motion 16

1.3.3 Numerical Integration of the Equation of Motion 16

1.3.4 Periodic Boundary Conditions 17

1.3.5 Computation of the Non-bonded Interactions 18

1.3.6 MD at Constant Temperature and Pressure 19

1.4 Trajectory Analyses 20

1.4.1 RMSD Calculation 20

1.4.2 Native Contacts 21

1.4.3 Radius of Gyration 21

1.4.4 Secondary Structure Assignment 21

1.4.5 Clustering Analysis 22

1.4.6 Solvent Accessible Surface Area (SASA) 22

1.4.7 Energetic Analysis 22

1.4.8 Hydrogen Bond Analysis 23

CHAPTER TWO – SEQUENCE ANALYSIS AND STRUCTURE PREDICTION OF TYPE II *PSEUDOMONAS* SP. USM 4-55 PHA SYNTHASE AND AN INSIGHT INTO ITS CATALYTIC MECHANISM 24

Abstract 24

2.1 Introduction 25

2.1.1 Polyhydroxyalkanoic acids 25

2.1.2 PHA Synthase 26

2.2 Methods 29

2.2.1 Data Mining and Sequence Analysis 29

2.2.2 Model Development and Evaluation 29

2.3 Results 29

2.3.1 Data Mining and Sequence Analysis 29

2.3.2 Secondary Structure Prediction 32

2.3.3 Template Selection 33

2.3.4 3D Model Building and Evaluation 36

2.4 Discussion 41

2.5 Conclusion 45

CHAPTER THREE - STRUCTURE PREDICTION OF LARGE PROTEIN USING THE COMBINATION OF THE COMPUTATIONAL KNOWLEDGE-BASED AND THE PHYSICS-BASED APPROACH: METHOD VALIDATION ON CHOLESTEROL ESTERASE. 46

Abstract 46

3.1 Introduction 47

3.2 Methods and System Preparation 49

3.2.1 Cholesterol esterase 49

3.2.2 Development of the Core Region of CE 52

3.2.2.1 Core region at % id cutoff > 70% (CE_{CRL}) 54

3.2.2.2 Core region at 30% < cutoff id < 60% (CE_{THG}) 54

3.2.2.3 Core region at % id cutoff < 20% (CE_{JKM}) 56

3.2.3 Development of the Complete 3D Model 57

3.2.4 Long MD Simulation 59

3.3 Results 59

3.3.1 Development of the 3D Models 59

3.3.2 Long MD Simulation 62

3.4 Discussion 69

3.5 Conclusion 73

CHAPTER FOUR - ON THE ROLE OF THE KNOWLEDGE-BASED CORE REGION AS A POTENTIAL 'NUCLEATION CENTER' PROMOTING DOWNHILL FOLDING OF PROTEINS 74

Abstract 74

4.1 Introduction 75

4.2 Materials and Methods 76

4.2.1 MD Simulation on Native NMR 76

4.2.2 Development of the Core Region 77

4.2.3 Development of the Complete Protein 77

4.3 Results 77

4.3.1 Structure Refinement on the Native HP-36 77

4.3.2 Development of the Core Region 78

4.3.3 Development of the Complete Protein 79

4.3.3.1 Structural Properties 79

4.3.3.2 Clustering and Energetic Analysis 85

4.4 Discussion 91

4.5 Conclusion 92

CHAPTER FIVE - THE EFFECTS OF GAS PHASE ON THE PROTEIN CONFORMATION: A STRUCTURE PREDICTION STUDY ON EOTAXIN-3 CYTOKINE 93

Abstract 93

5.1 Introduction 94

5.2 Materials and Methods 96

5.2.1 CC Chemokine Eotaxin-3 96

5.2.2 Simulation Details 97

5.2.2.1 Native NMR in Gas Phase 98

5.2.2.2 Native NMR in Explicit Water 98

5.2.2.3 Development of the Core region 98

5.2.2.4 Development of the Complete 3D Model 98

5.3 Results 99

5.3.1 MD on NMR Structure of 1G2S 99

5.3.2 Structure Prediction of 1G2S 110

5.3.2.1 Development of the Core Region 110

5.3.2.2 Development of the Complete 3D Structure 111

5.3.2.2 MD Refinement in Gas Phase and Solvated Phase 112

5.4 Discussion 123

5.5 Conclusion 126

CHAPTER SIX – GENERAL DISCUSSION, CONCLUSION AND FUTURE DIRECTIONS 127

6.1 General Discussion 127

6.1.1 The Knowledge-based and the physics-based structure prediction approaches 129

6.1.2 The proposed combined knowledge-based and physics-based approach for protein structure prediction 130

6.2 Proposed Future Studies 134

6.3 Conclusion 137

LIST OF TABLES

		Page
Table 1.1	Geometric specifications for the identification of hydrogen bonds in Hbplus	23
Table 2.1	Proposed template structures obtained from three threading methods along with the calculated scores	34
Table 2.2	Comparison of the Cα distances for the catalytic residues between the template structure 1HGL and the predicted structure of PhaC1$_{\text{P.sp USM 4-55}}$	38
Table 3.1	Summary of the BLAST results with scores and % global identity	52
Table 3.2	Proposed templates obtained from the three threading methods along with the % id and the calculated scores	53
Table 3.3	The chosen templates for the 3D core region structure prediction of CE at various cutoff of sequence identity	53
Table 3.4	The summary of properties of the models prior to MD simulation	59
Table 3.5	The RMSD$_{\text{back}}$ for the complete protein and the core region	61
Table 3.6	Native secondary structures with their respective residues for CE$_{\text{CRL-87}}$, CE$_{\text{THG-45}}$ and CE$_{\text{JKM-14}}$	62
Table 3.7	Average values of the structural and energetic properties of the predicted 3D structures during the last 10ns of MD refinement	64
Table 3.8	Total number of residues involved in the formation of the secondary structures for the native crystal, CE$_{\text{MD-avg}}$ and the CE models	68
Table 4.1	Comparison of structural properties between each conformational clusters	85

Table 4.2 Summary of various energy terms calculated for MD trajectories of HP-36 89

Table 5.1 The $\text{RMSD}_{\text{back}}$ for the developed 3D structure calculated after each residue additions 111

Table 5.2 Comparison of protein-protein hydrogen bonds 117

Table 5.3 Average SASA (Å^2) per residues for G2S_{vac} and G2S_{wat} with their maximum and minimum values 120

Table 5.4 Averaged values of the various energetic components for the gas phase and solvated phase simulations of the 1G2S native structures and the predicted 3D structures 124

LIST OF FIGURES

Page

Figure 1.1 A Peptide bond in a protein primary structure. The bond lengths and angles between the atoms are also illustrated. 2

Figure 1.2 The torsion/dihedral angles of ϕ and ψ that determine the rotational limits for the peptide group. 3

Figure 1.3 Periodic boundary condition. Atom from the central box is free to move into the other box, as they will be replaced by their image, which move into the central box. 18

Figure 2.1 General structure of Polyhydroxyalkanoate (PHA). n represents the number of repeating monomer units. C represents Carbon. 26

Figure 2.2 Pairwise sequence alignment between PhaC1$_{\text{P.sp USM 4-55}}$ (Gln249 – Ala492) and the consensus sequence of the α/β hydrolase fold family obtained from conserved domain search using BLAST. The alignment showed 32% similarity and 18% sequence identity. 30

Figure 2.3 CLUSTALW multiple sequence alignment for Type II PHA synthases with the highly conserved catalytic residues shown in colored blocks; green (Cys296); purple (Ser297); dark blue (Asp451) and red (His479). The consensus pentapeptide motif for PhaC1$_{\text{P.sp USM 4-55}}$, which comprises of GACSG is highlighted in the blue box. 31

Figure 2.4 The consensus results of 5 different secondary structure prediction methods for all 559 residues. 33

Figure 2.5 Sequence-structure alignment between PhaC1$_{\text{P.sp USM 4-55}}$ (PhaC1_Seq) and 1HLG (1HLGA_seq) generated by 3DPSSM. 35

Figure 2.6 Overall topology of the predicted 3D model of PhaC1$_{\text{P.sp USM 4-55}}$. 37

Figure 2.7 Secondary structure assignment of the predicted three-dimensional model of PhaC1$_{\text{P.sp USM 4-55}}$ from residue 267- 484 using DSSP. 37

Figure 2.8 Comparison of the 3D structure between the template 1HLG and the model of PhaC1$_{\text{P.sp USM 4-55}}$ (residue 267- 484). The blue ribbon represents β strands, the red ribbon represents α helices and the green region represents turns. 39

Figure 2.9 Surface presentation of the active site containing the proposed catalytic residues shown as the stick models (red = Asp451, light purple = His479, orange = Cys296, dark purple = Ser297). 39

Figure 2.10 Ramachandran Plot of the 3D model of PhaC1$_{\text{P.sp USM 4-55.}}$ Red region represents the most favored region, yellow=allowed region, light yellow=generously allowed region, white= disallowed region. 41

Figure 2.11 Proposed catalytic mechanism for PhaC1$_{\text{P.sp USM 4-55}}$ involving Cys296, Asp451 and His479 with the protein dimer forming the loading and elongation site. 44

Figure 3.1 Ribbon representation of the 3D tertiary structure of CE. 50

Figure 3.2 Structure topology for CE showing the arrangements of the secondary structures with the core region containing seven parallel β-strands forming parallel β-sheet with α-helices in between. 50

Figure 3.3 Complete protocols for the knowledge-based and *de novo* structure prediction approach. 51

Figure 3.4 Pairwise sequence alignment between the template 1CRL and the target CE. 55

Figure 3.5 Ribbon representation of the 3D model of core region CE based on the template structure 1CRL. 55

Figure 3.6 Pairwise sequence alignment between the template 1THG and the target CE. 56

Figure 3.7 Ribbon representation of the 3D model of core region CE based on the template structure 1THG. 56

Figure 3.8 Pairwise sequence alignment between the template 1JKM and the model CE. 57

Figure 3.9 Ribbon representation of the 3D model of core region CE based on the template structure 1JKM. 57

Figure 3.10 Protocol for the development of the 3D complete protein. 58

Figure 3.11 3D structures of the developed CE models after the complete addition of the amino acid residues. 61

Figure 3.12 Ribbon representations of the models before and after the 20 ns of MD refinement. The knowledge-based core region (yellow); segments occupied by the added residues (blue). The green and red CPK models represent Ala1 and Val534 which are the first and the last residues respectively. a,b and c represent, respectively, CE_{CRL-87}, CE_{THG-45} and CE_{JKM-14}. 63

Figure 3.13 Time evolution (ps) of $RMSD_{back}$ (Å) for CE_{CRL-87}, CE_{THG-45} and CE_{JKM-14} 64

Figure 3.14 Time evolution (ps) of the R_{gyr} (Å) for CE_{CRL-87}, CE_{THG-45}, CE_{JKM-14} and $CE_{MD-avg.}$ 65

Figure 3.15 Time evolution (ps) of Total SASA ($Å^2$) for CE_{CRL-87}, CE_{THG-45}, CE_{JKM-14} and $CE_{MD-avg.}$ 66

Figure 3.16 Time evolution (ps) of Nonpolar SASA ($Å^2$) for CE_{CRL-87}, CE_{THG-45}, CE_{JKM-14} and $CE_{MD-avg.}$ 67

Figure 3.17 Time evolution (ps) of Polar SASA ($Å^2$) for CE_{CRL-87}, CE_{THG-45}, CE_{JKM-14} and $CE_{MD-avg.}$ 67

Figure 3.18 3D structures of the developed CE models after 20 ns of MD refinement. 69

Figure 4.1 Ribbon representations of (a) HP-36 NMR_{native}. The N-terminal residue is Met1 and the C-terminal residue is Phe36. Helix 1 (Asp4- Lys8), Helix 2 (Arg15- Phe18) and Helix 3 (Leu23- Lys30) are shown as ribbons. (b) NMR-MD_{avg} after 50 ns of MD simulation. Helix 1(Asp4-Phe11), Helix 2 (Arg15-Asn20) and Helix 3 (Leu23-Lys33). 78

Figure 4.2 Ribbon representations of; (a) The predicted 3D model of Core Region of the HP-36 Model (Ala9-Asn28); (b) The predicted 3D model of HP-36 with the added residues shown as the extended regions colored in yellow (Met1-Lys8; Leu29-Phe36). 79

Figure 4.3 Time evolution of backbone RMSD; ALL represent the backbone RMSD for all residues while CORE represent the RMSD for the core region (residue Ala9- Asn28). 80

Figure 4.4 Time evolution of radius of gyration for model and the NMR-$MD_{avg.}$ 80

Figure 4.5 Time evolution of fraction of native contacts for the model. 81

Figure 4.6 Time evolution of total SASA ($Å^2$) with the contributions from polar and nonpolar residues. 81

Figure 4.7 Time evolution (ps) of secondary structure formations for the model. 83

Figure 4.8 Level of occupancies for hydrogen bonds present in the three helices at different time interval. α-helix 1 (Leu2@O – Asp6@N), 3/10-helix 1 (Leu2@O – Glu5@N), α-helix 2 (Thr14@O - Phe18@N; Arg15@O - Phe19@N; Ser16@O-Asn20@N), α-helix 3 (Leu23@O - Gln27@N; Trp24@O - Asn28@N). The level of occupancies corresponds to the percentage of occurrence at a given time interval, 1(0-5%); 2(5-20%); 3(20-40%); 4(40-60%); 5(60-80%). 84

Figure 4.9 Atomic interactions in the developed knowledge-based core region of HP-36. (a) The starting model prior to MD simulation; (b) Model after 100 ps of MD simulation. The collapse of Phe11 and Val10 towards the hydrophobic cluster core region was shown. 84

Figure 4.10 Distribution of six clusters with each containing different sets of MD trajectories. RMSD= RMSD of the trajectories with respect to the centroid structure. 86

Figure 4.11 Superimpose between the NMR-MD_{avg} (green ribbon) and the best model from cluster 2 (blue ribbon) with backbone RMSD 3.53 Å 86

Figure 4.12 Deviation from the NMR-MD_{avg} SASA presented by the total SASA, nonpolar and polar contributions for each residue of the best model. 87

Figure 4.13 Time evolution of the solvation free energy (filled circle) and the internal electrostatic energy (x). 90

Figure 5.1 NMR-solution structure of CC Chemokine Eotaxin-3 (1G2S) 97

Figure 5.2 Time evolution (ps) of $RMSD_{back}$ (Å) for $G2S_{NMR\text{-}vac}$ and $G2S_{NMR\text{-}wat.}$ 100

Figure 5.3 Root-mean square fluctuations (RMSF) (Å) shown for each residue. 100

Figure 5.4 Superimpose between g2s$_{nmr-vac}$ (red ribbon) and g2s$_{nmr-wat}$ (blue ribbon) and g2s$_{nmr}$ (yellow ribbon). N-term (n-terminus), c-term (c-terminus). The circled and blurred regions correspond to the highly mobile c-termini region and n-loop region, respectively. 101

Figure 5.5 Time evolution (ps) of secondary structure formation for G2S$_{NMR-wat.}$ 101

Figure 5.6 Time evolution (ps) of secondary structure formation for G2S$_{NMR-vac.}$ 101

Figure 5.7 Time evolution (ps) of total number of hydrogen bonds for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 103

Figure 5.8 Time evolution (ps) of atomic distances between the sulfur atoms of Cys10 and Cys34 for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 103

Figure 5.9 Time development (ps) of various dihedral angles for Cys10 (dark blue), Cys11 (pink), Cys34 (yellow), Cys50 (light blue). (a) phi in vacuum (ϕ_{vac}); (b) phi in water (ϕ_{wat}); (c) psi in vacuum (ψ_{vac}); (d) psi in water (ψ_{wat}); (e) chi in vacuum (χ_{vac}); (f) chi in water (χ_{wat}). 106

Figure 5.10 Time evolution (ps) of atomic distance between the sulfur atoms of Cys11 and Cys50 for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 107

Figure 5.11 Time evolution (ps) of SASA (Å^2) for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 108

Figure 5.12 Time evolution (ps) of R$_{gyr}$ (Å) for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 108

Figure 5.13 Time evolution (ps) of SASA (Å^2) for Gln13 shown for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 109

Figure 5.14 Time evolution (ps) of SASA (Å^2) for Ser15 for G2S$_{NMR-vac}$ and G2S$_{NMR-wat.}$ 109

Figure 5.15 Pairwise sequence alignment between the template 1JE4 and the target 1G2S. The core region was shown as the filled box. 111

Figure 5.16 Ribbon representation of the developed core region of 1G2S (Trp21 – Thr51) consisting of three antiparallel β-strands. 111

Figure 5.17 Time development (ps) of $\text{rmsd}_{\text{back}}$ (å) for g2s_{vac} and g2s_{wat} calculated for all atom and core region. Vac-all and vac-core represents, respectively, $\text{rmsd}_{\text{back}}$ on whole protein and $\text{rmsd}_{\text{back}}$ on core region for g2s_{vac}; wat-all and wat-core represents, respectively, $\text{rmsd}_{\text{back}}$ on whole protein and $\text{rmsd}_{\text{back}}$ on core region for g2s_{wat}. 112

Figure 5.18 Root mean square fluctuations (RMSF) (Å) averaged over 150 ns for G2S_{vac} and G2S_{wat}. 113

Figure 5.19 Time development (ps) of fraction of native contacts for G2S_{vac} and $\text{G2S}_{\text{wat.}}$ 113

Figure 5.20 Time development (ps) of the distance between Cys10 and Cys34 in (Å). 114

Figure 5.21 Time development (ps) of the distance between Cys11 and Cys50 in (Å). 115

Figure 5.22 Time development (ps) of secondary structural elements for $\text{G2S}_{\text{vac.}}$ 116

Figure 5.23 Time development (ps) of secondary structural element for $\text{G2S}_{\text{wat.}}$ 116

Figure 5.24 Time evolution (ps) of total number of hydrogen bonds for G2S_{vac} and $\text{G2S}_{\text{wat.}}$ 118

Figure 5.25 Time development (ps) of total SASA for both G2S_{vac} and $\text{G2S}_{\text{wat.}}$ 119

Figure 5.26 Time development (ps) of R_{gyr} for both G2S_{vac} and $\text{G2S}_{\text{wat.}}$ 119

Figure 5.27 Snapshots of G2S_{vac} (green ribbons) and G2S_{wat} (purple ribbons) taken at time intervals 10 ns (a & b), 50 ns (c & d), 100 ns (e & f) and 150 ns (g & h). The white and black beads represent, respectively, the N-terminus and the C-terminus. 122

LIST OF SYMBOLS

Cα	The central carbon alpha atom in amino acid
ϕ	Dihedral angle in protein structure (phi angle)
ω	Dihedral angle in protein structure (omega angle)
ψ	Dihedral angle in protein structure (psi angle)
α-helix	Secondary structure of proteins (alpha helix)
π-helix	Secondary structure of proteins (pi helix)
β-sheet	Secondary structure of proteins (beta sheet)
Å	Distance unit in Armstrong
μs	Microsecond
l_{eq}	Equilibrium bond length
k_r	Force constant
k_θ	Angular force constant
V_n	Heights of rotational barriers
$\varepsilon_{\mathrm{ij}}$	Well depth for Lennard Jones Potential
σ_{ij}	Collision diameter for Lennard Jones Potential
r_{ij}	Distance between two atoms

k_B	Boltzmann Constant
κ	Isothermal compressibility
τ_p	Coupling constant
ΔE_{MM}	Molecular mechanical energies
ΔG_{solv}	Solvation free energies
ΔG_{PB}	Electrostatic solvation
ΔG_{np}	Nonpolar solvation energy
$\phi(r)$	Electrostatic potential
$\varepsilon(r)$	Position dependent dielectric function
$\rho(r)$	Charge density

LIST OF ABBREVIATIONS

3D	Three dimensional
BLAST	Basic local alignment search tool
CASP	Critical assessment of techniques for protein structure prediction
3DPSSM	3D position specific scoring matrix
MD	Molecular dynamics
CPU	Central processing unit
AMBER	Assisted model building with energy refinement
CHARMM	Chemistry at Harvard molecular mechanics
RMSD	Root mean square deviation
NMR	Nuclear magnetic resonance
MM-PBSA	Molecular modeling poison Boltzmann surface area
PBC	Periodic boundary condition
PME	Particle mesh ewald
FFT	Fast fourier transform
MMTSB	Multiscale modeling tools for structural biology

SASA	Solvent accessible surface area
PHA	Polyhydroxyalkanoate
HMM	Hidden markov model
3HB	3-hydroxybutyric acid
CoA	Coenzyme A
HA	Hydroxyalkanoic acid

LIST OF PUBLICATIONS & SEMINARS

1. Nurul Bahiyah Ahmad Khairudin, Habibah A Wahab, Mohd Razip Samian, Nazalan Najimudin (2008). “ An Approach Towards the Prediction of Protein Tertiary Structures: Molecular Modeling Perspectives”. In: Ida Idayu Muhamad, Chew-Tin Lee (eds). Special Topics in Bioprocess Engineering, Vol 3. Johor: Universiti Teknologi Malaysia Press. 1-34.

2. NB Ahmad Khairudin, HA Wahab, MR Samian, N Najimudin (2008). “The effects of gas phase on the protein conformation: A molecular dynamics study on eotaxin-3 cytokine”. JNCRE Modelling Innovation, FKKKSA, Special Edition:18-28.

3. NB Ahmad Khairudin, HA Wahab, MR Samian, N Najimudin (2008). “Structure prediction on large protein using the combination of knowledge-based and physics-based approaches: Method validation on cholesterol esterase”. ”. JNCRE Modelling Innovation, FKKKSA, Special Edition: 141-154.

4. Habibah A Wahab, Nurul Bahiyah Ahmad Khairudin, Mohd Razip Samian, Nazalan Najimudin (2006). “ Sequence analysis and structure prediction of type II Pseudomonas sp. USM 4–55 PHA synthase and an insight into its catalytic mechanism”. BMC Structural Biology; 6: 23.

5. Ahmad Khairudin NB, Samian MR, Najimudin N, A Wahab H. (2005). Structural assignment of a type II PHA synthase and an insight into its catalytic mechanism using human gastric lipase as the modeling template. 2005 International Joint conference of International Conference on Bioinformatics (InCoB), Association of Asian Societies for Bioinformatics (AASBi) and Korean Society for Bioinformatics (KSBI). Busan Korea. September 22-24.

6. Ahmad Khairudin NB, A Wahab H, Samian MR, Najimudin N. (2006). Proposed catalytic mechanism for type II Pseudomonas sp USM 4-55 PHA synthase based on the predicted 3D structure. 3rd Life Sciences Postgraduate International Conference, Penang. May 24-27.

7. A Wahab H, Ahmad Khairudin NB, Adnan R, Najimudin N, Samian R, Kumar S. (2003). The structure and function of PHA synthase: Bioinformatics perspectives. 2nd International Conference on Bioinformatics 2003. Penang. 2003.

8. Ahmad Khairudin NB, A Wahab H, Najimudin N, Samian MR. (2004). Structure and Function of PHA synthase. Structural Biology Colloquium 2004. Penang. April 25-28.

Pendekatan Tergabung Berdasarkan Pengetahuan dan Fizik ke arah Peramalan Struktur Protein

ABSTRAK

Protein merupakan satu bahan binaan kehidupan. Pengetahuan mengenai struktur tiga dimensi protein ialah asas dalam memahami fungsi-fungsi protein. Dalam kajian ini, satu metodologi baru telah dicadangkan untuk meramal struktur-struktur protein ini dengan menggabungkan teknik bersandarkan maklumat iaitu peragaan homologi dan teknik bersandarkan pengetahuan fizik iaitu simulasi dinamik molekul. Bahagian teras protein dibina menggunakan kaedah pemodelan homologi berdasarkan persamaan jujukan asid amino dengan struktur templat protein. Bahagian-bahagian protein yang lain yang tidak dijana iaitu di bahagian terminal protein dibiarkan untuk berinteraksi dengan bahagian teras protein bagi meneruskan proses pembentukan struktur protein dengan menggunakan kaedah simulasi dinamik molekul. Metodologi ini boleh dikategorikan pada tiga bahagian; pemodelan bahagian teras struktur protein, pemodelan struktur protein secara keseluruhan dan penghalusan model menggunakan simulasi dinamik molekul. Teknik baru yang dicadangkan ini telah diuji pada tiga jenis protein yang berlainan saiz, kolesterol esterase (534 asid amino), CC Chemokine Eotaxin-3 (71 asid amino) dan subdomain “villin headpiece” ayam (36 asid amino). Model-model yang terhasil telah diuji dan dibanding dengan struktur-struktur asal masing-masing. Daripada keputusan analisis, teknik yang dicadangkan ini telah berjaya meramalkan struktur protein yang bersaiz kecil tetapi memberikan keputusan yang negatif untuk protein yang bersaiz besar. Satu perkara penting yang dapat dilihat dalam kajian ini membabitkan penemuan fenomena yang bergelar ‘pusat nukleasi’ yang membantu proses pembentukan struktur protein.

Combined Knowledge and Physics Based Approaches Towards Predicting Protein Structures

ABSTRACT

Proteins are the building blocks of life. Knowledge of the three dimensional structure of proteins is of fundamental importance in understanding protein function. In this study, a new computational method was developed to predict these structures by combining the knowledge-based homology modeling and the physics-based molecular dynamics simulation methods. The core region of the protein was initially constructed via homology modeling based on sequence similarity with other solved protein templates. Then, the remaining end-terminal regions were allowed to fold towards the core region via MD simulation a few residues at a time. The method was categorized into three parts; the development of the core region of the protein, the development of the complete protein structure and the MD refinement simulation. This proposed techniques was tested on three proteins with different sizes, cholesterol esterase (534 residues), CC Chemokine Eotaxin-3 (71 residues) and chicken villin headpiece subdomain (36 residues). The developed models were analysed and compared with their respective native structures. From the results, it was found that this method could successfully predict the structure of a small protein but showed a negative result for larger size proteins. Another important highlight of the current work is the identification of a 'nucleation center' which facilitated the folding process of a small protein.

Preamble

Impressive advances in genomic sequencing technologies are flooding us with the complete genetic blueprints of human, rat, mouse, chimpanzee and various microorganisms at an extremely rapid pace. As these DNA sequences continue to accumulate, the challenge to determine the function of each gene and to establish the corresponding protein structures is paramount. The two most mature and conventional experimental techniques to solve the structure of a protein are the X-ray crystallography and Nuclear Magnetic Resonance (NMR). John Kendrew and Max Perutz shared the Nobel prize in 1962 for their pioneering achievement in solving the atomic level structure of the protein myoglobin and hemoglobin, respectively, using X-ray diffraction. Since then, many protein structures were solved and various roles of proteins in living cells were known. Despite the accuracy and the advances of these experimental techniques, such methods are very costly and it may take months to years for solving one structure.

The current number of 3D protein structures is very small compared to the number of protein sequences, creating a huge gap between them. As this gap is expected to keep on growing with the ongoing genome projects, the experimental techniques certainly cannot be expected to keep pace with the rapid flow of the sequences. This limitation has fueled up an awareness on the importance of computational work that relies heavily on theoretical studies that could be employed to predict the structures of proteins in order to bridge the gap between the number of protein sequences and the solved structures.

The emerging of the computational protein structure prediction methods formed the basis of the current work. The initial motivation of this study was to predict the structure and function of the enzyme polyhydroxyalkanoate synthase (PhaC1$_{\mathrm{P.sp\ USM\ 4\text{-}55}}$), of which solving the 3D structure using X-ray crystallography posed a great challenge. Since structure determines function, it is highly crucial to elucidate the 3D structure of this enzyme that will facilitate in understanding the enzyme's catalytic mechanism. Unfortunately, there were many regions in the protein that could not be modeled using the current available methods. There was obviously a strong demand for a new approach that could predict the complete 3D structure of a protein when other structure prediction methods fall short. This thus provided a great basis to propose a new procedure that could predict the complete 3D structure of proteins.

In general, protein structures can be predicted using two different approaches, the knowledge-based and the physics-based methods. The former method employs knowledge or information from structural databases such as Homology Modeling while the latter relies heavily on the theoretical and physical laws such as molecular dynamics simulation. These two approaches furnish the underlying motivation for this work. Thus, the main objective of the current study is to develop a new structure prediction method that combined both knowledge-based and physics-based methods. The proposed method was tested on proteins of different sizes ranging from 534 residues to as small as 36 residues. It is anticipated that this work could contribute to the knowledge of structure prediction of large-, medium- and small-sized proteins whose structures cannot be solved using the conventional knowledge-based approach alone. In addition to this, the effect of gas phase simulation on protein conformation was also investigated as reported in Chapter 5. A quick summary of every chapter in this thesis is given as follows:

Chapter 1 presented a brief introduction to the biochemistry of proteins. Different structures of proteins and the forces that govern them were discussed, while various protein structure prediction techniques relevant to this work were introduced. The underlying principles of the molecular dynamics simulation which was the main ingredient for the current work were also presented. In addition to this, a number of analysis methods were described in the final part of the chapter.

Chapter 2 reported on the findings of the prediction of the 3D structure of Type II *Pseudomonas* sp. USM 4-55 PHA synthase 1 (PhaC$_{\text{P.sp USM 4-55}}$). The formations of two tetrahedral intermediates and oxyanion holes were proposed to play a role in the catalytic mechanism of this enzyme.

Chapter 3 introduced the proposed structure prediction method which combined both homology modeling and molecular dynamics simulation. This method was tested on three different starting models of a large protein, cholesterol esterase (1CLE) with 559 amino acid residues.

Chapter 4 presented the structure prediction study of a small protein, 36-residue chicken villin headpiece subdomain (1VII) using the previous proposed structure prediction method. The well-developed core region of the protein model was proposed to serve as a 'nucleation center' or a template for subsequent rapid folding of other residues.

Chapter 5 reported on the effects of gas phase on the conformation of protein with the application on the small 71-residue enzyme CC chemokine eotaxin-3 (1G2S) using the proposed structure prediction method. The results obtained from both simulations in vacuo and in explicit water were compared and cross-checked.

Chapter 6 gave general discussion on the findings for every chapter and how the studies were related to each other. The final summary and conclusion was also presented with some thoughts on the potential future directions of protein structure prediction.

CHAPTER ONE

Introduction

Protein is the most abundant and diverse group of macromolecules highly essential to biological processes. It plays a central role in all biochemical reactions in all forms of life. It transports oxygen, mediates cell signals, regulate gene expression and many more. These functions are determined by their unique three-dimensional (3D) structures as a result of protein folding. Protein folding can be defined as the process in which proteins spontaneously arrange their linear sequence of amino acids into native 3D structures that will allow them to function properly. Thus, elucidation of the 3D structure of a protein is vital in understanding its function. However, it is not known how the newly synthesized polypeptide chains fold into a protein with specific function. It was not until 1973 that Anfinsen (Anfinsen, 1973) postulated that all the information needed for a protein to correctly fold into its native structure is encoded solely in its amino acid sequence and that the native state of the protein is the conformation with the lowest energy. Consequently, this important principle has brought immense interests among the experimentalists and theoreticians to investigate how proteins fold into their native structures. Despite the countless efforts and dedicated hard work to surmount this intangible problem over the last four decades, the protein folding enigma still remains the most outstanding unsolved challenge in structural biology today.

1.1 The Biochemistry of Proteins

Proteins are biological macromolecules that consist of more than tens and up to hundreds of subunits of amino acids linked together like beads in a linear chain. Amino acids are organic compounds, formed from carbon, hydrogen, nitrogen, oxygen and sulfur. These subunits are also addressed as residues or monomers since they are the building blocks of the protein polymeric chain. These residues are characterized by a central carbon alpha atom (Cα). In general, it consists of four constituents connected to a tetrahedral Cα; two functional groups that consists of carboxylic acid group (-COOH) and amino group (-NH); a proton –H and a side chain (R-group) that varies among all the 20 naturally occurring amino acids (Voet *et al*., 2006). It is the side chains that give each protein its unique identity and ultimately determines its function.

1.1.1 The Peptide Bonds

Amino acids are joined together by amide linkages known as peptide bonds as shown in Figure 1.1. Two amino acids that are joined together are called a peptide. If there are more than 50 amino acids, they are called polypeptides or protein. Each peptide chain has two free ends; the amino terminus and the carboxy terminus. A unique characteristic of a peptide bond that contributes to the overall structure of a protein is its rigidity which is due to the resonance of electrons that imparts 40% of double bond character. The electrons in the bond are delocalised across the C=O, C-N and N-H bonds making the rotation in the peptide backbone restricted (Voet *et al*., 2006). Almost all peptide bonds are found in *trans* conformation, since it is more energetically and sterically favorable compared to *cis* conformation.

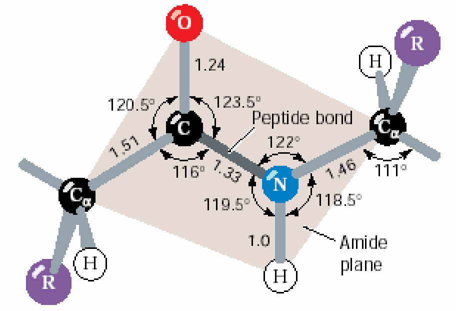

Figure 1.1 A peptide bond in a protein primary structure. The bond lengths and angles between the atoms are also illustrated (Voet *et al*., 2006).

1.1.2 Limitations on Conformations.

The peptide linkages along with the Cα atoms form the protein backbone. The conformation of this backbone can be described by the torsion angles or dihedral angles as shown in

Figure 1.2. The three important dihedral or torsion angles ϕ (phi), ω (omega) and ψ (psi) define the spatial conformation of the protein chain. The side chain attached to Cα of each amino acid limits the rotation about ϕ and ψ to avoid steric clashes between the side chain and the adjacent main chain atoms. The ϕ angle occurs due to the rotation of the bond joining N to Cα , the ω angle is the angle that joins together C and N of the chain while ψ is an angle about the Cα-C bond (Voet *et al*., 2006).

1.1.3 Structural Hierarchy

A protein can be described at four levels of structural organization; primary, secondary, tertiary and quaternary structures. The primary structure of a protein is the linear chain of amino acids connected to each other by a peptide bond. The secondary structure can be defined as regular, repeating shapes that form local spatial arrangements involving short segments of the primary structure (5 to 20 residues). Secondary structures are not formed from strong covalent bonding but rather via weak hydrogen bonds mainly within the peptide backbone (Pauling and Corey, 1951a, b).

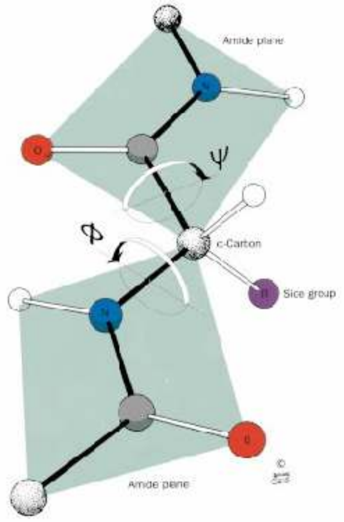

Figure 1.2 The torsion/dihedral angles of ϕ and ψ that determine the rotational limits for the peptide group (Voet *et al*., 2006).

The most common arrangements of secondary structure are helices and sheets. These two secondary structure elements satisfy a strong hydrogen bond network within the geometric constraints of the bond angles ϕ and ψ. Repeating values of ϕ and ψ form regular secondary structures. For instance, repeating values of $\phi \sim -57^{o}$ and $\psi \sim -47^{o}$ give an α-helix structure while repetitive angles in the region of ϕ = -110 to -140 and ψ = +110 to +135 give rise to parallel β-sheet fold (Voet *et al*., 2006). If the bond angles are not consistent in which the angles occupy both the regions for helices and sheets, the region will form random coil or loop. This segment is flexible and usually located on the protein surface. They are important in connecting elements of secondary structure together and also provide flexible hinges to support movements of the protein. Helices can be found as 3 types, α-helix, π helix and 3_{10} helix, depending on how tight the packing of the helices are (Voet *et al.*, 2006). The backbone atoms in π helix are so tightly packed whereas the atoms in 3_{10} helix are loosely packed. These two states are unfavorable. Only the backbone atoms in the α-helix are properly packed thus providing a stable structure. β-strands are usually made up of 5 to 10 residues in length. β-sheets are formed by hydrogen bonding network between two or more adjacent β-strands. They can be found in two patterns, parallel and anti-parallel depending on the orientation of the adjacent strands. In parallel sheets, the strands are oriented in the same N-terminal to C-terminal direction. In anti-parallel sheets, adjacent strands are oriented in the opposite way of each other (Voet *et al*., 2006).

1.1.4 Driving Forces

At physiological temperature, the stable state of a protein is governed by a number of weak, interdependent forces known as van der Waals interactions or dispersion forces between adjacent atoms. Although these non-bonded forces are weak, they play a vital role in the stabilization of proteins. These forces usually occur in the range of 3 to 4 Å between two atoms. Electron repulsion will prevent atoms from getting closer than 3 Å and the attraction force becomes weak beyond 5 Å.

The most important force that shapes protein tertiary structure is the hydrophobic effect (Kauzmann, 1959). This force forms the key feature of the folding of a protein in which the polar or charged residues tend to be at the surface of the protein exposing themselves to the solvents. This causes the hydrophobic residues to pack within the interior core of the protein protected from the solvent. The presence of hydrophobic side chains in aqueous solution results in the formation of water cages or structured water. This leads to an unfavorable reduction in entropy for the water molecules. In order to compensate for the loss of this entropy, the side chains of hydrophobic groups tend to pack with each other avoiding contacts with water molecules leading to the disruption of the water cages. Another important force in stabilizing the structures of proteins is hydrogen bond. Hydrogen bonds occur when a pair of nucleophilic atoms such as oxygen and nitrogen shares hydrogen between them. The hydrogen may be covalently attached to either nucleophilic atom (the H-bond donor) and shared with the other atom (the H-bond receptor). H-bonds are directional and their strength deteriorates dramatically as the angle changes (Voet *et al*., 2006).

1.2 The Protein Folding Problem

The protein folding problem has been discussed from two perspectives. The first one is the prediction of a protein structure from its amino acid sequence while the second concerns the elucidation of the protein folding pathway. While the former's objective is to solve the 3D structure of a protein, the latter aims at elucidating the protein folding kinetics and mechanisms involved towards achieving its native state with less interests in getting the functional structure. The protein structure prediction problem is commonly tackled using knowledge-based methods which rely heavily on the evolutionary relationship and information from structural databases. Meanwhile, the folding pathway problem relies strictly on the principles of physics employing the physics-based method such as molecular dynamics simulation. However, this approach is commonly hindered by the massive amount of computational power required.

Traditionally, methods of protein structure prediction can be categorized into three distinct levels, depending on the extent to which knowledge of structural databases is utilized (Rost, 1998). It ranges from the simplest and more accurate comparative modeling that relies on the structure of the homologous protein to the most difficult *ab initio* protein structure prediction. The latter attempts to predict the structure of a protein from the information of the amino acid sequence alone. Between these two extreme methods lies the fold recognition approach in which the model is built using a template protein that has very little or no similarity to the target protein.

1.2.1 Comparative Modeling

First reported by Browne and co-workers (Browne *et al*., 1969), comparative modeling (also known as homology modeling) is by far the most accurate and most successful method to predict the 3D structure of a protein. It has been agreed that conservation of the amino acid sequence will result in the conservation of the 3D structure (Chothia and Lesk, 1986). This approach exploits the idea that proteins with high similarity and identity in their amino acid sequence are evolutionarily related and tend to fold into similar structures (Westhead and Thornton 1998, Chothia and Lesk, 1986). The principal concept is to model the structure of a query protein (target) based on the backbone coordinates of one or more homologous proteins with known structure (template) (Blundell *et al*., 1987). This method produces an all-atom model of the protein derived from the sequence alignment between the target and the template proteins. The quality of the model is tightly linked to the closeness of the evolutionary relationship between the target sequence and the template structure. High sequence identity will result in a strong structural similarity and *vice versa*. In general, there are three steps involved; template selection, target-template sequence alignment and coordinates assignment through model building.

1.2.1.1 Template Selection

Template searching is usually carried out using search engine such as BLAST (Basic Local Alignment Search Tool) (Altschul *et al*., 1990). It compares the query sequence with each sequence from the structural database in a pair-wise mode employing BLOSUM substitution matrix (Henikoff and Henikoff, 1992). A substitution matrix estimates the numerical score associated with the cost or reward of sequence mutations or conservations. In other words, it penalizes the cost of aligning an amino acid with a different amino acid. For instance, aligning histidine with isoleucine will cost a negative score due to unlikely mutation. However, aligning isoleucine with leucine will cost a positive score as regard to acceptable mutation. The first thing to be considered in selecting the most appropriate template is by looking at the degree of sequence identity and similarity between the protein sequences. Although the percentage identity of the pair-wise sequence alignment determines how well the sequences are related to each other, the significance of the alignment can also be measured using the score and the E value (Expectation value). The higher the score, the more similar the two sequences are whereas the higher the E value, the less similar those sequences are (Claverie and Notredame, 2003).

1.2.1.2 Sequence Alignment

After identifying the appropriate templates from the structure database, the next step is aligning the protein sequences to identify regions that are conserved. Obtaining good alignments appears to be the ultimate key in accurate prediction. The accuracy of the alignment depends on the percentage of sequence identity and similarity between the two proteins. The alignment deteriorates once the sequence identity drops below 30% (Westhead and Thornton 1998, Venclovas *et al.*, 1999). If the sequence identity of the alignment exceeds 60%, the quality of the model can be regarded as comparable to that of the low resolution X-Ray structures (Marti-Renom *et al.*, 2000, Gerstein and Levitt, 1998a). There are a vast number of alignment methods available such as ClustalW (Thompson *et al*., 1994), T-coffee (Notredame *et al*., 2000) and DIALIGN 2 (Morgenstern, 1999) just to name a few .

1.2.1.3 Model Building

There are currently two major approaches of model building; fragment-based and restraint-based approach. However, only the restraint-based approach is discussed in this chapter. Restraint based approach or better known as modeling by satisfaction of spatial restraints is implemented in the program MODELLER (Sali and Blundell, 1993). In this approach, the model is constructed by optimizing and satisfying all the violations to the geometrical restraints derived from the sequence alignment between the model and the template. Such restraints include the backbone and side chain dihedral angles and limits on distances between pairs of Cα atoms (Sali and Blundell, 1993, John and Sali, 2003). These restraints were obtained empirically from a database of protein structure alignments which are derived from homologous structures, NMR experiments, rules of secondary structure packing, analyses of hydrophobicity and many more. The CHARMM22 (Brooks *et al*., 1983) forcefield is then combined with the spatial restraints to produce an objective function. Finally, the 3D model is obtained by optimizing the objective function in the Cartesian space by the use of the variable target function employing methods of conjugate gradients and molecular dynamics with simulated annealing. The resulting 3D protein model of the target sequence contains all main chain and side chain non-hydrogen atoms. Due to the fact that this method exploits many information on the target sequence, it can be considered as the most reliable method for structure building technique.

1.2.2 Fold Recognition

When the sequence identity between a target and a template drops below 30%, the comparative modeling method is no longer appropriate to be used to predict the protein structure. The quest for searching the correct fold thus becomes more difficult and the accuracy of the developed model deteriorates (Moult, 2005). Although it is already agreed that similar sequence adopts similar structures, the opposite is not necessarily correct. In contrast, two proteins can share similar structure topology even if there is no obvious sequence similarity (Swindells, 1992). Previous studies have shown that proteins sharing less than 10% of sequence similarities tend to have similar structures (Brenner *et al.*, 1998, Gerstein and Levitt, 1998a). Inspired by the notion that structure is evolutionary more conserved than sequence, fold recognition method aims to identify relationships between remotely related proteins for which comparative modeling methods are unable to detect any potential template (Sippl and Flockner, 1996, Jones and Thornton, 1993).

One of the important and widely used methods in fold recognition is the threading method. It attempts to fit a query sequence to a library of known folds to identify pairs of proteins that share similar structures without any evolutionary links. This method assumes that two different proteins might achieve global energy minimum in the same area on the potential energy surface (Godzik *et al.*, 1992, Jones *et al.*, 1992). The earliest threading work was introduced by Bowie and Luthy using the '3D profiles' as the scoring method (Bowie *et al.*, 1991, Luthy *et al.*, 1992) in which model that fits to the wrong fold will score poorly. The threading methods of Jones (Jones *et al.*, 1992) as well as Godzik (Godzik and Skolnick, 1992) are based on the residue pairwise interaction energy method. The degree of compatibility of the proteins is evaluated using a set of empirical potentials derived from known protein structures (Sippl, 1990). The alignment score is calculated by adding up all the pairwise interaction energies between each residues of the target and the template. It is believed that the native state of a protein corresponds to the global minimum of the free energy. Consequently, the correct fold for a target protein can be inferred by calculating the energy for each threading process. Jones (Jones *et al*., 1992) found that threading a sequence into its own native structure exhibited the lowest energy fold. Due to this, they developed a threading algorithm known as Genthreader (Jones, 1999) by assuming that the structure which produces the lowest energy fold is said to be the best threading template for the target sequence.

Another technique that contributes a huge improvement in fold recognition approach is the sequence comparison with the incorporation of specific structural information such as the secondary structure elements and the surrounding environments (such as solvent, pH and ligands) of which the residues preferred to be. It was proven that structures developed using this method were modeled with high accuracy (Jaroszewski *et al*., 1998) and that it performed well in the CASP meetings for the fold recognition category (Olszewski *et al.*, 2000, Sippl *et al.*, 2001). The program 3D Position Specific Scoring Matrix (3D-PSSM) evaluates the match of the query sequence to the sequence of the template, the secondary structure and the solvent accessibility pattern of the template (Kelley *et al*., 2000). Another successful method that should be highlighted is FUGUE (Shi *et al*., 2001) which utilizes the environment-specific substitution tables and structure-dependent gap penalties. The scores are evaluated based on the local environment of each amino acid residue in a known protein structure.

1.2.3 *ab initio* Method

The absence of a potential template poses immense difficulties in predicting the 3D structure of the protein. The only method that can be applied is the *ab initio* or *de novo* prediction method. In its purest form, this approach attempts to fold a protein into its novel structure using only the amino acid sequence information alone without the knowledge of any similar folds. The CASP6 and CASP7 meetings held in December 2004 and 2006, respectively, showed some success in *ab initio* category (Moult *et al.*, 2005) indicating that the technique has matured since the first CASP meeting in 1994 (Moult *et al*., 1995). However, despite the impressive development, almost all of the *ab initio* benchmark structure prediction attempts were focused on small to medium size proteins because they require less computational power compared to the large proteins (Simons *et al.*, 2001, Liwo *et al.*, 1999).

In general, the a*b initio* protein structure prediction can be further classified into two types based on two underlying principles. The first kind involves the utilization of the information from structural databases or fragment libraries (Kolodny *et al.*, 2002, Huang *et al.*, 1999, Yue and Dill, 2000). The incorporation of the knowledge from protein structures has somewhat blurred the distinction between the first principle *ab initio* approach and the fold recognition approach. An example of this is the Rosetta method developed by David Baker's group in 1997 where short segments obtained from library of known fragments independently sample distinct region of local conformations (Simons *et al.*, 1997, Rohl *et al.*, 2004). These short segments are assembled in a Monte Carlo search strategy until low free energy interactions are achieved, the hydrophobic residues are buried, beta-strands are paired and other non-local interactions are favorable. Successes of protein structure prediction from Rosetta can be seen from the third, fourth and fifth CASP meetings with predicted structure having Cα RMSD ranging from 3-7 Å compared to the native fold for small proteins (Bonneau *et al.*, 2001, Bradley *et al.*, 2003, Simons *et al.*, 1999).

Inspired by the fact that physical forces dictate the folding of a protein, the second type of *ab initio* method relies heavily on the physico-chemical approach without utilizing any information from structural databases. Computer simulation such as Molecular Dynamics (MD) is commonly employed to fold proteins from non-native to native states. This all-atom protein folding simulation is capable of providing a microscopic view of the entire important molecular folding events in atomic details. Thus it is commonly used to study the folding pathways of proteins. MD simulation was first initiated to study the interactions of hard spheres (Alder and Wainwright, 1959). Subsequently, Rahman and Stillinger reported the first MD on liquid water (Rahman and Stillinger, 1971). It was not until seven years later that MD was first applied to 58-residue protein BPTI and this has caused a dramatic change in the perspective of protein dynamics (McCammon *et al*., 1977). However, the subsequent progress in this area was very slow and it was doubtful that MD simulation can be applied to study the folding of proteins in the foreseeable future (Shakhnovich, 1997). Among reasons given include the timescale of folding which was far beyond what could be achieved and the available force fields were not robust and accurate enough for such a complex process.

Surprisingly, the remarkable evolution of this field especially in the last 10 years has prevailed over skepticisms on this method. This progress is mainly spurred by the development of sophisticated technologies especially the increase of computational power and speed. Current scenario has allowed computer simulation to be carried out utilizing large distributed clusters comprising hundreds to thousands of CPUs (Gnanakaran *et al.*, 2003, Fersht and Daggett, 2002, Daggett, 2000, Pande *et al.*, 2003, Duan and Kollman, 1998). Another important factor that contributes towards the success of this method is the much improved molecular modeling potential functions or force fields. Among those widely used are AMBER (Cornell *et al*., 1995), CHARMM (MacKerell *et al*., 1998) and GROMOS (Daura *et al*., 1998).

The current scenario has perceived MD folding simulation to be carried out in all-atom representation with the presence of solvent molecules. The study of 1 μs MD simulation of villin headpiece subdomain containing 36 residues in explicit water marked the breakthrough for protein folding simulation in leaping from nanosecond to microsecond time scale (Duan and Kollman, 1998). This study was facilitated by the advent of parallelization of the simulation code for supercomputers. The early stage of folding was dominated by hydrophobic collapse and helix formation followed by intermediate stable state before reaching the native-like structure. This finding was supported by another study on the same protein by another group (Zagrovic *et al*.,

2002) that was reported to exceed 300 μs of total dedicated simulation time using worldwide-distributed computing technique utilizing thousands of CPUs. Known as one of the fastest folding proteins (villin headpiece subdomain) with estimated folding time of 5 μs (Kubelka *et al.*, 2003), this work has achieved the native structure with average RMSD around 1.7 Å and 1.9 Å that is comparable to low resolution X-ray or NMR solved structures. Thus, suffice to say that MD by no doubt is the most powerful tool in solving the structure of proteins given an ample amount of time and a boost in computational power. Nevertheless, the method is still restricted to very small proteins and peptides with the time regime limited to mostly hundreds of nanoseconds (Fersht and Daggett, 2002).

Apart from the complete folding simulation, various studies have also applied MD for structure refinement in the endgame of protein folding. The endgame of protein folding refers to the ultimate phase in the folding process (Lee *et al.*, 2001c). It is thought that at this stage the overall fold has already been achieved and that the orientation of the amino acid side chains are the elements that need prior focus. The method developed by Wang and colleagues (Wang *et al.*, 1995a, Wang *et al.*, 1995b) used solvation parameters (Eisenberg and McLachlan, 1986) obtained from a training algorithm that maximized the solvation energy to differentiate between the native and nonnative structures generated by MD simulation. In another related study, (Huang *et al.*, 1996) the ability of a hydrophobic contact function to identify the correct model using MD simulations was investigated at two different operating temperatures. The averaged RMSD to the native structure were found to be 1.5 Å and 4.1 Å for MD at 298 K and 498 K, respectively. The study managed to distinguish 330 false positives out of 10,000 models that were initially developed by knowledge-based prediction method with MD refinement.

In contrast to these studies, three related work carried out by the Kollman group (Lee *et al.*, 2001a, Lee *et al.*, 2001c, Lee *et al.*, 2000) used MD and molecular-mechanics-Poisson

Boltzmann/surface area (MM-PBSA) (Kollman *et al.*, 2000) to discriminate the predicted model from false positives. Models of two small proteins, 36-mer villin headpiece domain (HP-36) and 65-mer region of ribosomal protein (S15) obtained from Rosetta structure prediction program were subjected to MD for refinement and the average free energies were calculated using MM-PBSA for each structure ensemble. The ensemble that had the lowest average free energies were observed to have the lowest Cα RMSD in the core region with values of 2.1 Å for the former and 1.8 Å for the latter (Lee *et al.*, 2001a).

1.3 Fundamentals of Molecular Dynamics

1.3.1 The Force Field

A classical model of a system is usually represented by a potential function or force field. Together with related parameters, it expresses the internal potential energy exerted on an atom as a function of the positions of the other atoms. The function is represented as a sum of five terms that can be classified into bonded and non-bonded terms as shown in Equation 1.1 (Leach, 2001).

$$U(\mathbf{r}^N) = \underbrace{\sum_{bonds} \frac{k_r}{2}(l - l_{eq})^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1+\cos(n\phi - \gamma))}_{\text{Bonded Terms}}$$

$$\underbrace{+\sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}}}_{\text{Non - bonded terms}}$$

Equation 1.1

The first term describes the energy involved in the stretching of a bond between pairs of bonded atoms. The energy is treated as a simple harmonic potential that obeys Hooke's Law where *l* represents the observed bond lengths, l_{eq} is the corresponding equilibrium bond length and k_r is the force constant in Kcal/mol. The second term describes the angle bending energy between three covalently bound atoms. Similar to bond energy, angle bending energy can also be described using the harmonic function in which θ is the observed angle with θ_0 as the equilibrium angle while k_θ represents the angular force constant. This term controls the bond geometry in molecules.

The dihedral or the torsional energy is represented by the third term in the force field equation. It describes rotation around a chemical bond. This term adopts a cosine function with ϕ as the dihedral angle between two planes. V_n is the heights of rotational barriers and n is an integer that describes the number of minima in a 360° rotation. γ is the phase factor that determines the location of the minima. The fourth term describes the 12-6 Lennard Jones potential that represents the non-bonded van der Waals energy between two atoms. This term contains repulsion and attraction subterms. Repulsion forces occur when the distance between two atoms is less than the sum of their van der Waals radii. However, this force becomes weaker when the atoms are distantly separated ($r \rightarrow \infty$). On the other hand, the attraction forces also known as dispersion forces take place when the distance between two atoms is more than their van der Waals radii in which there is no overlap between the two electron clouds. This force dominates due to the formation of instantaneous dipoles in an atom which will induce the formation of new dipoles in other atoms. In the equation, ε_{ij} determines the well depth, σ_{ij}, the collision diameter determines the separation of which the energy is zero and r_{ij} is the distance between two atoms.

The last term in the potential function is the coulombic interaction that represents the non-bonded electrostatic energy between pairs of partial charges in two atoms. q_i and q_j are the partial charges of the involved atom pair i and j. ε_0 is the permittivity of a vacuum, ε_r is the relative permittivity of the medium and r_{ij} is the non-bonded distance. In the equation, it is shown that electrostatic energy decays as r^{-1}, due to this, electrostatic interactions are considered as long-ranged forces.

1.3.2 The Equation of Motion

MD calculates the time dependent behavior of a molecular system such as atomic positions and velocities. It integrates the Newton's second law, which is the equation of motion (Equation 1.2) for an assembly of atoms on a potential energy surface. From the equation, the force F_i is the force exerted on a particle, m_i is its mass and a_i is its acceleration. The force F can be calculated as the negative gradient of the potential energy (U) as shown in Equation 1.3 (Leach, 2001). Therefore the equation provides a link between the molecular motions and the potential energy function.

$F_i = m_i a_i$ Equation 1.2

$F_i = - \nabla_I U$ Equation 1.3

The integrations of the equations of motion will result in atomic trajectories. These trajectories contain the detailed information about the time evolution of microscopic states such as velocities and positions in phase space.

1.3.3 Numerical Integration of the Equation of Motion

The positions and velocities of the particles are propagated using numerical integrators that employ the Taylor series expansion (Leach, 2001). The most basic and common algorithm is the Verlet integrator (Verlet, 1967) which is based on a forward and a backward Taylor expansion as shown follows:

$r(t + \delta t) = r(t) + \delta t\, v(t) + \frac{1}{2} \delta t^2 a(t) + \ldots$ Equation 1.4

$r(t - \delta t) = r(t) - \delta t\, v(t) + \frac{1}{2} \delta t^2 a(t) - \ldots$ Equation 1.5

Adding these two equations will then yield:

$r(t + \delta t) = 2r(t) - r(t - \delta t) + \delta t^2 a(t)$ Equation 1.6

This algorithm uses positions (*r*) and accelerations $\left(\frac{f}{m}\right)$ at time *t* as well as the past positions, *r (t-Δt)* to derive new positions *r (t+Δt).* From the equation, it is obvious that Verlet integrator does not require explicit velocities. One of the variations of the Verlet integrator is the Leap-Frog Algorithm (Hockney, 1970). Designed to improve the evaluation of the velocity, this algorithm gets its name from the way in which positions and velocities are calculated in an alternating sequence. For this algorithm, the velocities leap over the positions, and then the positions leap over the velocities. The velocities are first calculated at time *t* + (1/2)*δt*, which then are used to calculate the positions r, at time *t* + *δt*. The following equations represent the algorithm to evaluate the position and velocity;

$$r(t+\delta t) = r(t) + \delta t\, v\left(t+\frac{1}{2}\delta t\right) \qquad \text{Equation 1.7}$$

$$v\left(t+\frac{1}{2}\delta t\right) = v\left(t-\frac{1}{2}\delta t\right) + \delta t\, a(t) \qquad \text{Equation 1.8}$$

The velocities at time t can be approximated by the relationship:

$$v(t) = \frac{1}{2}\left[\; v\left(t-\frac{1}{2}\delta t\right) \;\right] + v\left[\; t+\frac{1}{2}\delta t \;\right] \qquad \text{Equation 1.9}$$

1.3.4 Periodic Boundary Conditions

To prevent the particles in a solution from drifting apart in the simulation, the system is usually confined in a box or a container that has a finite size. However, the finite volume of the system does not represent the bulk fluid properties as in real condition due to the occurrence of surface effects. As the system is being surrounded by surfaces, particles around the boundary would have fewer neighbors than the interior particles. This will lead to inaccuracy in calculating the potential forces exert on the outer particles. Thus, there will be a significant difference of behavior between particles on the surface and particles in the bulk. To eliminate this surface effect, a method called periodic boundary condition (PBC) is widely employed (Leach, 2001). PBC enables a simulation in such a way that all the particles experience forces as though they were in a bulk solution. In PBC, the simulation box containing the system is infinitely replicated in all the three Cartesian (x,y,z) directions filling the space to form a lattice. This removes any influence of the surface walls on the system since PBC mimics the condition of a bulk system. When a particle in the original central box moves, its virtual images will move in the same direction. Thus, particles that leave one side of the box will have one of its images entering the box at the opposite side as shown in Figure 1.3. Consequently, the number of particles in the central box and all the periodic boxes are always conserved.

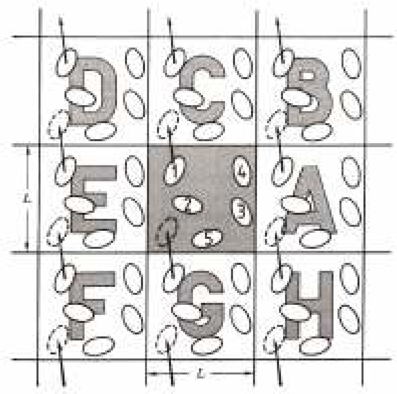

Figure 1.3 Periodic boundary condition. Atom from the central box is free to move into the other box, as they will be replaced by their image, which move into the central box (Allen and Tildesley, 1987).

1.3.5 Computation of the Non-bonded Interactions

The non-bonded energy interactions between every pair of atoms are evaluated at each step making it the most time consuming part in MD simulation. To speed up the computation, several methods such as the non-bonded cut-off and the Particle Mesh Ewald (PME) methods are employed. The non-bonded cut-off method is used to calculate the non-bonded forces within a cut-off distance making the interactions outside this distance negligible. In principle, when PBC is employed, the cut-off radius must not exceed half of the shortest box vector (length). This is known as the *minimum image convention*, which is to prevent an atom interacting with its own image or the same particle twice.

It is possible to truncate the short-ranged Lennard Jones potential using only a distance cut-off since the potential falls off very rapidly with distance. However, straight truncation of the long-ranged electrostatics by using cutoff method will result in substantial artifacts when simulating biomolecules such as proteins and DNA. Realizing this, Darden introduced the accurate and fast PME to properly treat the long-range electrostatic interactions (Darden *et al.*, 1993). This method splits the interactions into short-range and long range that can be evaluated in direct space using truncation method for the former and in the reciprocal space using Fast Fourier transform (FFT) for the latter. Each charge is screened by a Gaussian charge distribution cloud of equal magnitude and opposite sign to make it short-ranged. A cancelling Gaussian charge distribution of the same sign as the original charge is also added to compensate for the previous addition of charge distribution. FFT accelerates the solution of Poisson's equation in PBC by interpolating the charges onto a regular spaced grid.

1.3.6 MD at Constant Temperature and Pressure

Temperature is related to the kinetic energy of the system as given by equation 1.10, in which k_B is the Boltzmann constant, T is the temperature and N represents the number of particles in the system. Therefore, adjusting the velocities can control the temperature of the system. This is achieved by coupling the system to an external heat bath kept at the specified temperature (Berendsen *et al.*, 1984). The velocity at each step is scaled using Equation 1.11 that results in the change of temperature, which is proportional to the difference in temperature between the system and the heat bath. τ is the coupling constant that determines how tightly the system being coupled to the bath.

$$E_{kin} = \frac{3}{2} N k_B T \qquad \text{Equation 1.10}$$

$$\lambda_i = \sqrt{1 + \frac{\delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right)} \qquad \text{Equation 1.11}$$

The same control scheme applies for pressure regulation. Equation 1.12 shows that pressure is related to the isothermal compressibility κ and the volume of the system. Therefore, pressure is kept fixed by changing the volume of the simulation cell. This is achieved by coupling the system to a pressure bath (Berendsen *et al.*, 1984) that is kept at the specified pressure. The volume at each step is scaled using Equation 1.13 with coupling constant, τ_p.

$$\kappa = -\frac{1}{V}\left(\frac{\delta V}{\delta P}\right)_T \qquad \text{Equation 1.12}$$

$$\lambda_p = 1 - \kappa \frac{\delta t}{\tau_p}\left(P(t) - P_{BATH}\right) \qquad \text{Equation 1.13}$$

1.4 Trajectory Analyses

This section presents the various trajectory analyses used in the current work. All graphical molecular representations presented in this thesis are produced using VMD (Humphrey *et al.*, 1996), InsightII[1], DSViewer Pro[1] and Chimera 1.2184-linux (Pettersen *et al.*, 2004).

1.4.1 RMSD Calculation

A conventional way to compare two protein structures is to put one structure onto the other and calculate the RMSD or root mean square deviation. It can be calculated using the following equation:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2\right]} \qquad \text{Equation 1.14}$$

In which *x, y* and *z* are the atomic coordinates for the particle *i* and *j*. N is the number of atoms in the protein. The RMSD analyses were calculated using the PTRAJ module implemented in Amber8.

1.4.2 Native contacts

The native contacts was calculated using Multiscale Modeling Tools for Structural Biology (MMTSB) Tool set (Feig *et al.*, 2004). It is defined as the contacts between atoms for a given protein structure compared to the contacts in the native structure for the same atoms. This calculation is based on a minimum distance between heavy atoms of two residues of less than 4.2 Å. Only tertiary contacts are calculated in this method ignoring atoms i+1, i+2, i+3 and i+4.

1.4.3 Radius of Gyration

CARNAL module in Amber8 was used to calculate the radius of gyration of the predicted structure. It gives the information regarding the expansion or contraction of the structure by calculating the scalar length of each atom from its center-of-mass (COM). The radius of gyration can be described using the equation:

$$R_{gyr} = \sqrt{\frac{\sum_i^n m_i \left[x_i - x_{cm}\right]^2 + \left[y_i - y_{cm}\right]^2 + \left[z_i - z_{cm}\right]^2}{\sum_i^n m_i}} \qquad \text{Equation 1.15}$$

In which the coordinates and mass of each particle i are (x_i, y_i, z_i) and m_i, respectively. (x_{cm}, y_{cm}, z_{cm}) are the coordinates of the center of mass.

1.4.4 Secondary Sructure Assignment

DSSP (Kabsch and Sander, 1983) was used to assign secondary structure elements to the 3D models. The graphical presentations were obtained using the program PolyView (Peterson *et al.*, 2000) and the do_dssp program from GROMACS (Lindahl *et al.*, 2001).

[1] Accelrys Inc. San Diego

1.4.5 Clustering Analysis

The MD simulation on protein folding normally generates a lot of trajectories. This causes analysis procedure to be daunting. As a consequence, it is highly useful to classify these structures into classes that meet requirements such as clustering the structure into groups that are observed to have different intermediate states in the folding pathways. This was done using MMTSB (Feig *et al*., 2004) tool set.

1.4.6 Solvent Accessible Surface Area (SASA)

The solvent accessible surface area was calculated using the program Naccess (Hubbard and Thornton, 1993) based on the methods of Lee and Richards (Lee and Richards, 1971). The radius of the sphere representing the solvent molecule is assumed to be 1.4 Å.

1.4.7 Energetic Analysis

The conformational free energy of a protein was calculated using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) protocol (Kollman *et al.*, 2000, Srinivasan *et al.*, 1998).The average free energy (Equation 1.16) is calculated as a sum of the average gas phase molecular mechanical energies (ΔE_{MM}) and the average of the solvation free energies (ΔG_{solv}) minus the entropy contributions of several snapshots structures taken from the explicit MD trajectories.

$$\Delta G = \Delta E_{MM} + \Delta G_{solv} - T\Delta S \qquad \text{Equation 1.16}$$

The molecular mechanical energy (ΔE_{MM}) or the internal strain energy of the molecule is calculated by summing up the bonded and non-bonded energies:

$$\Delta E_{MM} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{vdw} + \Delta E_{elec} \qquad \text{Equation 1.17}$$

The electrostatic solvation free energy (ΔG_{solv}) shown by equation 1.18 is estimated as the sum of the electrostatic solvation (ΔG_{PB}), calculated by the numerical solution to the Poisson-Boltzmann (PB) (equation 1.19) and the nonpolar solvation energy (ΔG_{np}), calculated as SASA independent term. In the PB equation, $\phi(r)$ is the electrostatic potential, $\varepsilon(r)$ is the position dependent dielectric function, and ρ(r) is the charge density due to the solute. The nonpolar contribution is due to the cavity formation and the free energy of inserting the discharged solute into that cavity.

$\Delta G_{solv} = \Delta G_{PB} + \Delta G_{np}$ Equation 1.18

$\nabla\varepsilon(r)\nabla\phi(r) + 4\pi\rho(r) = 0$ Equation 1.19

The nonpolar solvation energy (ΔG_{np}) is estimated from Equation 1.25 with both $\gamma = 0.00542$ cal/mol.Å and $\beta = 0.92$ cal/mol (Sitkoff *et al.*, 1994).

$\Delta G_{np} = \gamma \text{ x SASA} + \beta$ Equation 1.20

The third term from Equation 1.16, TΔS, which represents the solute entropy can be estimated by normal mode analysis on a Newton-Raphson minimization method. Nevertheless, the calculation involved is the most time-intensive part of the MM-PBSA program and the contribution is much smaller compared to the other two terms in estimating the free energies as it varies negligibly from trajectory to trajectory (Vorobjev and Hermans, 1999). Furthermore, it has also been proven that this term could not distinguish between the protein natives and the protein decoys (Lee *et al.*, 2000). Thus, the calculation of the free energies in this study only involved the internal energy and the solvation energy as presented by Equation 1.21.

$\Delta G = \Delta E_{MM} + \Delta G_{solv}$ Equation 1.21

1.4.8 Hydrogen Bond Analysis

The program Hbplus (McDonald and Thornton 1994) was used in this study to calculate the total number of hydrogen bonds in the structures. The default geometric criteria for hydrogen bonds used in this program are shown in Table 1.1.

Table 1.1: Geometric specifications for the identification of hydrogen bonds in Hbplus

Criteria	Atoms involved	Limits
Maximum distances	D-A H-A	3.9 Å 2.5 Å
Minimum angles	D-H-A D-A-AA H-A-AA	90° 90° 90°

D = hydrogen bond donor, A = hydrogen bond acceptor, H = hydrogen

CHAPTER TWO

Sequence Analysis and Structure Prediction of Type II *Pseudomonas* sp. USM 4-55 PHA Synthase and an Insight into Its Catalytic Mechanism

Abstract

This study seeks to investigate the structural properties as well as the catalytic mechanism of Type II *Pseudomonas* sp. USM 4-55 PHA synthase 1 ($PhaC1_{P.sp\ USM\ 4\text{-}55}$) using computational approach. Sequence analysis demonstrated that $PhaC1_{P.sp\ USM\ 4\text{-}55}$ lacked similarity with all known structures in databases. PSI-BLAST and HMM Superfamily analyses demonstrated that this enzyme belongs to the alpha/beta hydrolase fold family. Threading approach revealed that the most suitable template to use was the human gastric lipase (PDB ID: 1HLG). The superimposition of the predicted $PhaC1_{P.sp\ USM\ 4\text{-}55}$ model with 1HLG covering 86.2% of the backbone atoms showed an RMSD of 1.15 Å. The catalytic residues comprising of Cys296, Asp451 and His479 were found to be conserved and located adjacent to each other. In addition to this, an extension to the catalytic mechanism was proposed whereby two tetrahedral intermediates were believed to form during PHA biosynthesis, an interesting feature which has never been highlighted before. These transition state intermediates were further postulated to be stabilized by the formation of oxyanion holes. Based on the sequence analysis and the deduced model, Ser297 was postulated to contribute to the formation of the oxyanion hole. The 3D model of the core region of $PhaC1_{P.sp\ USM\ 4\text{-}55}$ from residue 267 to residue 484 was developed using computational techniques and the locations of the catalytic residues were identified. Results from this study for the first time highlighted Ser297 potentially playing an important role in the catalytic mechanism of the enzyme.