Molecular modeling and statistical analysis on proteins and lipids

Ping, Zhi

2018

Ping, Z. (2018). Molecular modeling and statistical analysis on proteins and lipids. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73737>

<https://doi.org/10.32657/10356/73737>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**MOLECULAR MODELING AND STATISTICAL ANALYSIS
ON PROTEINS AND LIPIDS**

PING ZHI

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

2017

MOLECULAR MODELING AND STATISTICAL ANALYSIS ON PROTEINS AND LIPIDS

PING ZHI

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2017

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

24/06/2017

Date

.....
Ping Zhi

Abstract

In this work, the application of statistical coupling analysis in a kind of transmembrane protein, Aquaporin, is studied. The major objective of this work is to discover potential correlated residues in *E. coli* AqpZ involved in molecular selectivity and permeability. Mutating such residues would introduce glycerol permeability to the protein. To obtain this goal, SCA technique combined with network analysis and 3D structural analysis are employed for critical position identification. MD simulation is also used for verifying the accuracy of prediction. This work also applies SCA technique in AQP evolutionary study as well as butelase 1/ligase-related study. To further understand the details of AQP, MD simulation on characteristics of phospholipid bilayer and vesicle is conducted.

This research performs SCA analysis on a treated database and establishes corresponding networks according to SCA scores. A set of potential correlated positions are identified through combination analysis of network and molecule structure. Simulation is conducted to approve the correctness of this co-mutation sites prediction. SCA is also shown to be applicable in phylogenetic categorization and to help to find important positions related to the AQP evolutionary process between higher and lower animals. Meanwhile, SCA technique is also demonstrated to allocate critical residues in butelase 1 for ligation catalysis.

The findings of this research show the application of SCA technique in bioinformatic research and imply the significance of correlated residues in proteins besides highly-conserved ones. It will be promising to use SCA to allocate critical residues for diseases in AQP family and the research into many other protein families as well.

Acknowledgements

To my dearest mom and dad, and everyone in my life

My appreciation also goes to my supervisor, Associate Professor, Dr Su Haibin, for his dedicated guidance and enlightenment. His insight inspiration, advices and selfless support in both of my PhD study and life gave me strength and confidence even during the hardest moment. It is my great fortune and honor to study and work with him for the last five years.

I would also like to thank Dr Lin Xin, for introducing me into Prof Su's research team and his beneficial advices. Without his pilot work, it would be extremely rough for me to complete.

I also wish to thank Professor Marvin Weinstein, Professor James P Tam and Dr Hemu Xinya for their kind and helpful invaluable advices during collaboration.

My gratitude also goes to many of my groupmates, Dr Zhou Feng, Dr Ma Xiaohua, Dr Wang Jingbo, Dr Wang Qiang, Mr Zhu Kaichen, Dr Wang Jianfeng, Dr Xuan Fengyuan and Mr Zhang Chaodong, for their helpful discussion and sharing of knowledge. And I would thank all other groupmate for their kindness support along the time.

At last, I thank my wife, Dr Ling Han, for her selfless love and support in my life, especially during my recuperation from ulcerative colitis. Without the love she gave me, it is impossible for me to complete the work.

Table of Contents

Abstract.....	i
Acknowledgements	iii
Table of Contents	v
Table Captions	ix
Figure Captions.....	xi
Abbreviations	xv
 Chapter 1 Introduction.....	 1
1.1 Hypothesis.....	2
1.2 Objectives and Scope.....	4
1.3 Dissertation Overview	6
1.4 Findings and Outcomes.....	7
References.....	8
 Chapter 2 Literature Review	 9
2.1 Overview of Aquaporin Protein Family	10
2.2 Typical Structures of AQP and AQGP.....	12
2.3 <i>E.coli</i> AqpZ and GlpF.....	14

2.4 Mutations on AqpZ for Function Manipulation.....	15
2.5 Overview of Proteases and Ligases	17
2.5.1 Proteases.....	18
2.5.2 Ligases.....	18
2.6 Review of Cyclotides and Butelase 1	19
References.....	23
 Chapter 3 Experimental Methodology.....	27
3.1 Method selection.....	28
3.2 Multiple Sequence Alignment	30
3.3 Statistical Coupling Analysis.....	32
3.4 Molecular Dynamic Simulation Methods.....	34
3.4.1 Modelling and Force-Fields	35
3.4.2 All-atom molecular dynamics simulation and opl-s-aa force-field	37
3.4.3 Umbrella Sampling	38
3.4.4 Replica Exchange Molecular Dynamics (REMD) Simulation	39
3.4.5 Potential of Mean Force	40
3.4.6 MARTINI Coarse-Grained MD simulation	41
3.5 Network Theory	44
References.....	45
 Chapter 4 Structural Context of Selectivity Filter of AQP/AQGP and Residue Clusters Affecting Molecule Selectivity	49
4.1 Introduction.....	50
4.2 Principle Outcomes.....	51
4.2.1 Single site conservation analysis.....	51

4.2.2 General network analysis of Aquaporin results	51
4.2.3 Analysis of Top ranked correlation pairs SCA networks of AQPs	53
4.2.4 Analysis of Top ranked correlation pairs SCA networks of AQGPs	55
4.2.5 Comparison of results of DQC methods and SCA network analysis	57
4.2.6 Replica Exchange Molecular Dynamics (REMD) simulations of mutated Aqp Z and Glp F models	58
4.2.7 Potential of Mean Force Calculation	60
4.3 Discussion	61
4.3.1 Additional residues should be mutated based on the SF mutation.....	61
4.3.2 Suggested additional co-mutation sites close to SF region	61
4.3.3 Suggested mutation residues in C-loop and M6-M7 linker	64
4.4 Conclusion	66
References	68
 Chapter 5 Statistical Coupling Analysis in Aquaporin Evolution Study	71
5.1 Introduction.....	72
5.2 Results and Discussion.....	73
5.2.1 SCA application on phylogenetic aspect	73
5.2.2 Analysis of AQP1 and AQP4 in evolutionary aspect	75
5.3 Conclusion	79
References.....	80
 Chapter 6 Statistical Coupling Analysis Application in Butelase 1 Related Research	83
6.1 Introduction.....	84
6.2 Results and Discussion	85

6.2.1 Single site conservation of activation peptide (AP) region.....	85
6.2.2 Sequence alignment of a group of AP regions from typical AEPs	88
6.2.3 SCA application on AEP family proteins	89
6.3 Conclusion	94
References.....	97
 Chapter 7 Molecular Dynamic Simulation on Phospholipid Self-Assembly and Vesicle Fusion	99
7.1 Introduction.....	100
7.2 Molecular context of membrane and phospholipid bilayer	101
7.3 Simulations of lipid aggregates fusion process	104
7.3.1 Fusion of vesicle and vesicle: Spontaneous fusion at stalk position	104
7.3.2 Fusion of vesicle and bilayer	106
7.3.3 Self-assembly of Lipid Vesicle	107
7.4 The Trigger Event of Fusion – One Special Lipid Molecule	108
References.....	111
 Chapter 8 Summary and Future Works	113
8.1 Summary and Discussions	114
8.1.1 Features of AQP/AQGP selectivity filter and potential mutation sites	114
8.1.2 SCA method application in protease/ligase study	115
8.1.3 SCA application in aquaporin evolutionary study	115
8.2 Recommended Future Works.....	116
8.2.1 Investigation of the relationship between aquaporin/other proteins and diseases	116
8.2.2 Residue interaction network analysis of AQP family proteins	117

8.2.3 SCA application in other proteins function or structural research	118
8.2.4 Further investigation on butelase 1-facilitated ligation mechanism	119
References.....	120
Appendix	123

Table Captions

Table 2.1 Classification of aquaporin family proteins based on phylogenetic analyzes, their organism distribution and glycerol permeability.

Table 2.2 Proteoliposome permeability for water and glycerol.

Table 3.1 Mapping of subtypes and interactions between subtypes in MARTINI model.

Table 4.1 Cross reference table of positions and their equivalent residues in Glp F and Aqp Z.

Table 6.1 Equivalent residues of the positions from Butelase 1, Jack bean legumain and the general major conserved residues in plants.

Figure Captions

Figure 2.1 Schematic presentation of aquaporin transmembrane helices.

Figure 2.2 Structure of Aqp Z and Glp F.

Figure 2.3 Cross section of SF size comparison in wild type rat Aqp1 and its mutants.

Figure 2.4 Pore size and molecule permeability of AqpZ and GlpF.

Figure 2.5 MS characterization of peptide cyclase activity/Structure model of Butelase

Figure 3.1 Schematic presentation of MSA procedure.

Figure 3.2 Force field representation.

Figure 3.3 Simulation systems used in this work.

Figure 3.4 Two cases of potential landscape.

Figure 3.5 Schematic presentation of umbrella sampling across the potential energy landscape.

Figure 3.6 Schematic presentation of replica exchange steps in REMD.

Figure 3.7 Types of interaction centers in MARTINI CG model.

Figure 3.8 MARTINI CG model of DPPC and water.

Figure 3.9 A typical scale-free network and its node degree power-law distribution.

Figure 4.1 The multiple sequences alignment (MSA) matrix.

Figure 4.2 Degree distributions of SCA network.

Figure 4.3 Top ranked SCA network and critical residues identification in tetramer structure of AQPs.

Figure 4.4 Top ranked SCA network and critical residues identification of AQGs.

Figure 4.5 Channel radius of 3NKC, AqpZ, GlpF and mAqp Z along channel axis.

Figure 4.6 PMF along z-axis of Glp F and Aqp Z mutant of glycerol and water passage.

Figure 4.7 Comparison of equivalent residues in structures of Aqp Z and Glp F near SF.

Figure 4.8 Function of position 173.

Figure 4.9 Interactions of C-loop residues with Arginine gate of SF in Glp F.

Figure 4.10 Critical residues located on M6-M7 linker.

Figure 5.1 Phylogenetic tree of aquaporin protein family.

Figure 5.2 Top-ranked SCA network of AQP subfamilies.

Figure 5.3 AQP family proteins distribution in human organs.

Figure 5.4 Cys102 in Human Aqp1.

Figure 5.5 Met70, Cys106 and His151 in human Aqp 4.

- Figure 6.1** Single site conservation result generated by Weblogo tool.
- Figure 6.2** Schematic presentation of electrostatic status of different parts in legumain.
- Figure 6.3** MSA of a set of legumain.
- Figure 6.4** Demonstration of His-Trp π -interaction.
- Figure 6.5** Top-50 scored pairs of nodes in SCA analysis of plant legumains.
- Figure 6.6** Top-50 scored pairs of nodes in SCA analysis of plant legumains.
- Figure 6.7** “Cap” segment in the AP region.
- Figure 6.8** Position 332 (Tyr350 in Butelase 1) in helix-loop-helix part of AP region.
- Figure 6.9** Representation of reactive center in legumain.
- Figure 6.10** Schematic presentation of proposed mechanism of butelase 1 triggered cyclization.
- Figure 7.1** Typical structures of transmembrane proteins.
- Figure 7.2** Mass density profile of DPPC bilayer membrane.
- Figure 7.3** Mass density and electric charge density profile of POPC/POPE mixed bilayer.
- Figure 7.4** Electrostatic potential distribution along z axis of a POPC/POPE mixed lipid bilayer.

Figure 7.5 Simulation snapshots of vesicle-vesicle fusion.

Figure 7.6 Stalk formation of vesicles fusion.

Figure 7.7 Simulation snapshots of vesicle-bilayer fusion.

Figure 7.8 Simulation of vesicle self-assembly from mixture lipid bilayer.

Figure 7.9 Critical lipids in lipid aggregates fusion process.

Abbreviations

AEP	Asparaginyl Endopeptidase
AQP	Orthodox Aquaporin
AQGP	Aquaglyceroporin
BLAST	Basic Local Alignment Search Tool
CNS	Central Nervous System
DPPC	1,2-dipalmitoyl- <i>sn</i> -glycero-3-phosphocholine
DPPE	1,2-Bis(diphenylphosphino)ethane
MD	Molecular Dynamics
MSA	Multiple Sequence Alignment
PMF	Potential of Mean Force
POPC	1-Palmitoyl-2-oleoylphosphatidylcholine
POPE	1-palmitoyl-2-oleoyl- <i>sn</i> -glycero-3-phosphoethanolamine
REMD	Replica Exchange Molecular Dynamics
SCA	Statistical Coupling Analysis

Chapter 1

Introduction

The current bioinformatics research of proteins usually focuses on the highly conserved sites, but neglect the less conserved co-evolved sites. Statistical coupling analysis is a powerful biotechnological tool for discovering correlated residues in proteins. This chapter presents a brief overview of orthodox alignment strategy, statistical coupling strategy and network theory. It also calls for a better understanding of the molecule selectivity of aquaporin family protein in the evolutionary perspective. The rationale of choosing aquaporin as the objective is provided. An overview of the structure of the whole dissertation is given afterwards.

1.1 Hypothesis

Point-mutation and co-mutation are conventional techniques which are widely employed in proteomic/genetic investigation and protein design. However, to ascertain the mutation site is always a challenging process. Without lots of works in bioinformatics research, identifying the mutation site for the purpose of protein dysfunction or introduce new function requires a massive effort in trial and error tests.

Nowadays, one of the most popular bioinformatic techniques in protein design is multiple sequence alignment, which normally aligns a large number of protein sequences to identify conserved residues and motifs highly related to protein functions and structures that give evolutionary relationships for protein family/superfamily. Many researchers have developed plenty of algorithms and tools to confine the alignments and dig information from the them as much as possible.

The technique of multiple sequence alignment (referred as MSA herein after) is no doubt successful and has broad application (1). Nevertheless, it has its own natural deficiency. The alignment alone would only give information about single reserved residues or motifs, but lack of any information as to how the other residues evolve with the variation of one specific site – coupled mutation. In many cases of manipulating protein functions by point mutations, after identifying the critical residues by MSA of related protein families, mutation of only these residues could not achieve the expected result. Some other residues, may affect the physical or chemical properties of these critical residues indirectly, although not located close to them in secondary or tertiary structure. The protein itself could be considered as a complex system where the residues are the basic elements. The mutual relationships between residues or residue clusters, including physical, chemical and evolutionary connections, play significant roles in protein function and structure. Hence it is important to treat the protein as a whole when searching for suitable mutation sites as no element in the system can be altered without affecting the others.

Statistical coupling analysis (referred as SCA herein after) is a popular technique developed in 1999 and is getting more and more mature in the last decades. (2) This technique is employed to measure the covariations between the selected pairs of residues

in the protein MSA, which means it can quantify how the residue changes in the particular position with the variation of the residue in another particular position (3,4). With this technique, it is more convenient to identify the co-evolved pairs of residues in the protein family/superfamily.

Network theory has long been a hot topic in almost all fields of research. Although it was first developed for computer sciences, it can be applied to many other fields as long as they have object individuals with complicated relationships, or the object could be considered as a complex system. Biological systems contain many elements with physical, chemical, electrical and many other connections, which are perfect for application of network theory. Furthermore, it has been proved that network analysis could be a powerful tool in research of biology, especially in bioinformatic studies, including protein interaction network (5), human disease network (6), phylogenetic network (7) etc.

In this work, aquaporin/aquaglyceroporin superfamily was chosen as the main objective system with its extensive existence in almost all kinds of organisms all over the environment. In this superfamily, the orthodox aquaporin (referred as AQP herein after) can select water molecules and transport them efficiently, while the aquaglyceroporin (referred as AQGP herein after) is able to transport water as well as glycerol (8). Plenty of genomes of proteins from this superfamily had been sequenced and the corresponding protein sequences could be acquired expediently from databases and some of the crystalized protein 3D structures were solved (9,10). The manipulation of orthodox aquaporin and aquaglyceroporin functions has been a meaningful topic since the aquaporin family proteins were first discovered and shown their remarkable ability in filtering specific molecules and ions, including water and glycerol. This specialty of these proteins makes them perfect subject for research into sea water desalination and fabricating biological filters with specific selectivity. Moreover, many experiments have been done for altering the selectivity of these proteins. Some of the point mutation experiments did not achieve the expected results owing to the effects of some less conserved residues which did not showing their significance in MSA. Therefore, the aquaporin superfamily is suitable for bioinformatic research on residue coevolution and mutation sites prediction.

Besides identification of critical highly correlated residues, it is also very interesting and important to study the evolutionary process. Many strategies and algorithms have been developed for phylogenetic analysis of DNA and protein sequences for the purpose of evolutionary studies. Trying to discover the possibility of applying SCA technique in protein evolutionary research is also necessary. In this work, aquaporin family proteins are again employed for the study in order to investigate the evolution process of different subfamilies from lower animals to higher animals.

Similar procedure for analyzing aquaporin protein family could also be employed into research of other protein families. Butelase 1, as an efficient protein ligase, was a remarkable proteomics discovery in the past few years. Although butelase 1 belongs to legumain, a protease family, it shows great ability in catalyzing the reverse reaction of hydrolysis, which protease normally does. In this work, an SCA application is conducted on legumain family proteins for investigations of possible molecular mechanisms of butelase 1 triggered ligation.

To further understand the function and behaviors of aquaporin or other transmembrane proteins in cells, it is also necessary to study the surrounding environment of aquaporin – cell membrane, or in simplified model as phospholipid bilayers. The basic form of phospholipid bilayer in cells is vesicle. To get a detailed understanding of vesicle self-assembly as well as fusion process, molecular dynamics simulation technique is employed in this work.

1.2 Objectives and Scope

Herein, the major objective of this work is to apply SCA methodology as well as network analysis concept in discovering the pairs of coevolved residues that contribute to the molecular selectivity of AQP/AQGP and the function of protease/ligase, propose suitable mutation sites that could assist the manipulation of protein function, and dig up the hidden evolutionary connections among residues of aquaporin superfamily.

In order to achieve the above goal, first, a dataset of selected aquaporin superfamily protein

sequences is established. Meanwhile, to get a more accurate outcome from SCA and to avoid noise from data set, the number of homolog sequences should be large enough and share less sequence identity. The aquaporin dataset is employed for further study of mutation sites discovery. With the dataset successfully established, the sequences are aligned using MSA algorithm. After that, sequences with too many gap insertions during MSA would be removed as incomplete sequences.

To discover the evolutionary hidden connections between residues and identify the critical pairs of residues which affect aquaporins' molecular selectivity, the sequences are going to be separated into two groups as orthodox aquaporin and aquaglyceroporins. The SCA technique is applied to calculate the covariation of these two separate groups respectively. Networks made up of highly correlated positions based on SCA scores could be established and compared. Based on the comparison, particular network pattern affects the protein selectivity would be able to be discovered. Meanwhile, 3D structures of prototype AQP/AQGP, which are *E.coli* Aqp Z and *E.coli* Glp F, would be employed as model system for structural study. Combining the SCA network results and analyzing their influence on channel-forming residues, with the further assist of dynamic quantum clustering, a number of potential mutation sites are identified.

To prove the accuracy of the predicted mutation sites and further investigate how these residues affect the channel selectivity, computer modelling technique is employed. The radius of both natural and mutated channel of AqpZ are to be measured and compared. Furthermore, replica exchange molecular dynamics modeling would be applied to demonstrate how the water and glycerol molecule pass through the natural and mutated protein channels (11). In-depth structural analysis on these residues could be performed and the molecular details of selectivity filter configuration would be revealed. Later on, the analysis into evolutionary aspects of different aquaporin subfamilies could also be conducted to discover the hidden connections among residues over evolution.

With the experience of analyzing the AQP/AQGP superfamily by SCA network analysis, it is promising to further utilize this technique to study the possible mutual transformation of protease and ligase, i.e. legumain family protein and butelase 1 in this work.

1.3 Dissertation Overview

This dissertation addresses the applications of SCA in protein bioinformatic research, including aquaporin superfamily and protease/ligase transformation, and contains eight chapters.

Chapter 1 gives an overview of this work and introduces the rationale, the goal and an outline of this work.

Chapter 2 reviews literatures concerning the AQP/AQGP superfamily, the different molecule transportation preference of AQP and AQGP, as well as the mutation attempts to manipulate the selectivity. A brief introduction of AQP/AQGP superfamily will be given first. After that, previous techniques, research and experiments on critical residue identification and manipulation of protein function, especially for AQP/AQGP function transition, will be reviewed. The literatures concerning butelase 1 as well as legumain protein family will also be given afterwards. Based on the literature reviews, problems needing to be solved will be proposed.

Chapter 3 introduces the theoretical bases of techniques which are employed in this work, including multiple sequence alignment, statistical coupling analysis, network analysis, molecular modeling (including MARTINI coarse-grained force field), replica exchange molecular modeling and dynamic quantum clustering.

Chapter 4 shows the results of SCA analysis application in investigating the AQP/AQGP substrate selective permeation in molecular basis. A series of important residues are given based on the analysis. Mechanisms that affects the molecule selectivity are also proposed in detail. Modelling is conducted and a series REMD simulations are performed to prove the accuracy.

Chapter 5 describes an evolutionary application of SCA in AQP subfamily research and reveals that SCA method could also be employed to categorize proteins. Furthermore, some SCA-critical residues are selected for analysis to discover their roles and importance during

evolution process.

Chapter 6 discusses the SCA method application in protease/ligase investigation based on the discovery of butelase 1, an efficient protein/peptide ligase with similar structure and sequence identity with AEP protease. Some featured residues are predicted to be responsible for the ligation process. Ligation mechanism detail is also presumed according to the results.

Chapter 7 shows the MD simulation of a set of phospholipid bilayer/vesicle self-assembly as well as vesicle fusion process. The curvature force is proved to be the driving force of bilayer-vesicle self-assembly. A trigger event of appearance of a lipid molecule with specific conformation is revealed.

Chapter 8 summarizes the implications and the highlights of this dissertation, and proposes the future recommended works based on the results.

1.4 Findings and Outcomes

This work gives several novel findings as follows:

- a. Giving a few potential mutation sites, which could suggest guidance to experiments aiming AQP/AQGP function switching manipulation.
- b. Establishing a protocol of identify critical pairs of correlated residues by SCA network analysis in protein research and design.
- c. Showing the possibility of application of SCA network analysis in protein evolutionary trend research of big data.

References:

1. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS computational biology. 2007;3(8):e123.
2. Lockless, S.W. and R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families. Science, 1999. 286(5438): p. 295-9.
3. Tesileanu T, Colwell LJ, Leibler S. Protein sectors: statistical coupling analysis versus conservation. PLoS computational biology. 2015;11(2):e1004091.
4. Watanabe L, de Moura PR, Bleicher L, Nascimento AS, Zamorano LS, Calvete JJ, et al. Crystal structure and statistical coupling analysis of highly glycosylated peroxidase from royal palm tree (*Roystonea regia*). Journal of structural biology. 2010;169(2):226-42.
5. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell. 2006;125(4):801-14.
6. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(21):8685-90.
7. Finnila, S., M.S. Lehtonen, and K. Majamaa, Phylogenetic network for European mtDNA. Am J Hum Genet, 2001. **68**(6): p. 1475-84.
8. Agre, P., The aquaporin water channels. Proc Am Thorac Soc, 2006. **3**(1): p. 5-13.
9. Verkman AS, Mitra AK. Structure and function of aquaporin water channels. American journal of physiology Renal physiology. 2000;278(1):F13-28.
10. Verkman AS. Aquaporins at a glance. Journal of Cell Science. 2011;124(13):2107-12.
11. Zhou, R., Replica exchange molecular dynamics method for protein folding simulation. Methods Mol Biol, 2007. **350**: p. 205-23.

Chapter 2

Literature Review

In this chapter, literature in the related fields, including aquaporin protein family as well as protease and ligase proteins, is reviewed. Aquaporin, as one of the most important protein families, has stimulated much research interest. Introduction to previous mutation attempts on AqpZ to manipulate its molecule permeability is presented. The simple mutation of key residues is proved to obtain an undesired outcome. One of the reasons is that the effects of residues correlated to these key residues but not directly involved in the function were overlooked. The better understanding of the proteins and previous experimental attempts gives the rationale of selecting methods and analysis for understanding the in-depth detailed mechanism of protein functions.

2.1 Overview of Aquaporin Protein Family

Aquaporin is a general name of a sub-family of transmembrane channel proteins which belongs to the family of Major Intrinsic Proteins (MIP)(1). They exist widely in most of the organisms including prokaryotes and eukaryotes (2). The main function of aquaporin is associated with selective permeation of water and other small solutes across the biological membrane. Aquaporin could be classified into two categories according to its specificity of water and glycerol permeability (2, 3). Orthodox aquaporins (referred as AQP herein after) are exclusively permeable to water; while the aquaglyceroporins (referred as AQGP herein after) or glycerol facilitator proteins (GLPs) are permeable to both water and glycerol. Since the crystal structure of the Aquaporin 1 was first determined in 2000 (2), the structures of several proteins belong to the aquaporin family have been discovered in the last two decades (4-7).

During the latest few decades, aquaporin attracted the interest of researchers not only for its application in water transportation in metabolic pathways, but also in clinical studies and therapeutic development. It is known that different aquaporin homologues are essential membrane proteins distributed in different organ tissues in the mammalian bodies with different transportation ability and thus different functions. Aquaporin deficiency and mutations in particular sites will cause various diseases.

For example, AQP3 and AQP4 appear in basolateral plasma membranes in kidney, the knockout of these two aquaporins in mice will cause mild urine-concentrating defect and severe polyuria respectively (8). The AQP0 in lens will maintain the transparency, two typical common mutations of AQP0, Glu134Gly and Thr138Arg, could lead to cataracts (9). Aquaporin 2 was also reported to have relevance with nephrogenic diabetes insipidus (NDI), 39 mutations of Aquaporin 2 were discovered to cause NDI in 10% of NDI patients, including Arg254Leu, Glu258Lys, which will cause autosomal dominant NDI (10).

In cancer-related pathological research, aquaporin has also attracted a lot of interests. Some aquaporin proteins are found to enhance cell proliferation, migration and survival in several kinds of cancer diseases. Recently, AQP3 was also reported to induce E-cadherin repression and promote epithelial-mesenchymal transition in gastric cancer via

PI3K/AKT/Snail signaling pathway, which provides a novel drug target for cancer therapy. (11) Some publications also reported that AQP1 and AQP5 expression levels are found to be high in various kinds of tumor cells, and believed to trigger rapid water exchange in and out of cells and lead to rapid changes of cell shape and volume. (12-14)

Table 2.1 Classification of aquaporin family proteins based on phylogenetic analyzes, their organism distribution and glycerol permeability

Table 1

Classification of aquaporin sequences from phylogenetic analyses

Name	Organism	Glycerol Permeability
AQP subfamily (Orthodox aquaporins)		
AQP0	Mammals	No
AQP1	Mammals	No
AQP2	Mammals	No
AQP4	Mammals	No
AQP5	Mammals	No
AQP6	Mammals	No
AQP8	Mammals	No
AQPZ	E.coli	No
AQY1	Yeast	No
AQY2	Yeast	No
Plasma-membrane intrinsic proteins (PIPs)	Plant	Depends
Tonoplast intrinsic proteins (TIPs)	Plant	Depends
Nodulin-like intrinsic proteins (NIPs or NLMs)	Plant	Depends
Small basic intrinsic proteins (SIPs)	Plant	Not known
GLP subfamily (AQGP)		
AQP3	Mammals	Yes
AQP7	Mammals	Yes
AQP9	Mammals	Yes
AQP10	Mammals	Yes
GlpF	E.coli	Yes
GLPY1	Yeast	Yes
GLPY2	Yeast	Yes

Furthermore, aquaporin is also a candidate target of therapy involved in Alzheimer's

disease (AD) and Parkinson's disease (PD). The expression level of AQP4 is significantly reduced in diagnosed AD brain with cerebral amyloid angiopathy (CAA) which related to the accumulation of amyloid- β , the major cause of AD (15, 16). AQP4 is also involved in dopamine regulation on proliferation of striatal astrocytes in vitro, which is considered to be related to PD (17).

Besides pathological research, aquaporin family proteins have shown great significance in many other fields because of their great importance in different metabolic pathways. For example, AQP3, which is classified as aquaglyceroporin, show great importance in epidermal and stratum corneum hydration by facilitating glycerol transport. Experiments in mice also suggested its value in improving skin elasticity and wound healing. Therefore, to increase AQP3 expression could be a target of cosmetics that improve skin moisturization.

2.2 Typical Structures of AQP and AQGP

Aquaporins usually consist of around 320 amino acid residues and have masses of approximately 28 kDa. Both AQP and AQGP have homologous structures because of their high protein sequence identity. Generally speaking, an aquaporin family protein is a pore forming protein consisting of six transmembrane helices and two half helices. The aquaporin proteins usually form tetramers embedded in the cell membrane. For each monomer, a channel is formed by the above-mentioned helices and act as the transport pathway for water or other small molecules like urea and ammonia.

One of the most featured characteristic of almost all the aquaporin proteins, either AQP or AQGP, is the two Asn-Pro-Ala (NPA) highly conserved motifs locates on the two half helices (M3 & M7) and in the center of the channel lumen. These two motifs are proven to play significant roles in water transportation by flip the water molecule's orientation during transportation by hydrogen bonding interactions provided by Asn.

Figure 2.1. Schematic presentation of aquaporin transmembrane helices. Two NPA motifs and C-loop are represented in orange color.

Another particular feature of aquaporin is the aromatic/arginine selectivity filter (ar/R SF) located near the extracellular entrance. As its name shows, the SF always determines the size of molecules that would pass through this channel. In orthodox aquaporin, the ar/R SF consists of four residues (typically Arg, Phe, His and the fourth one) and is so small that water or other similar sized molecules could pass through. Correspondingly, the ar/R SF in aquaglyceroporin normally established by three residues (typically Arg, Phe, and Trp/Thr), but is broader compare with that in orthodox aquaporin, which gives the protein more choice of molecule transportation such as glycerol. Therefore, the ar/R SF actually provides a distinguishing feature for identifying subfamilies of the aquaporin family.

2.3 *E.coli* AqpZ and GlpF

Aquaporins not only exist in animals and plants, but can also be found in bacteria. In *E.coli*, one of the most common model system organism, both AQP and AQGP are discovered as aquaporin Z (AqpZ) and glycerol facilitated protein (GlpF).

Figure 2.2. Structure of Aqp Z and Glp F. (A) Superposition of 3D structure of Aqp Z (grey) and Glp F (cyan), the two NPA motifs locate on M3 and M7 are represented in green color; (B) Selectivity filter region of Glp F, consists of W48, G191, F200, R206, represented by spheres, the area of pore size is $9.3\text{\AA}^2$; (C) Selectivity filter region of Aqp Z, consists of F43, H174, T183, R189, the area of pore size is $3.9\text{\AA}^2$.

They have been widely studied during the last few decades with 3D crystal structures determined in 2002 and 2003 by Stroud et al. (19, 20) (PDB ID: Aqp Z – 1RC2; Glp F – 1LDA). They could be considered as the typical structures for AQGPs and AQPs since aquaporin proteins have high sequence identity and structure similarity. In this study, the structures of *E.coli* Glp F and Aqp Z (refer as Glp F and Aqp Z) are used to demonstrate the structural analysis of bacteria aquaporin family proteins.

Like all the other AQPs, AqpZ has a typical ar/R selectivity filter with the cross-section

area of $3.9 \text{ \AA}^2$, which consists of four residues F43, H174, T183, R189. Correspondingly, GlpF has a much broader ar/R SF with cross-section area of $9.3 \text{ \AA}^2$. The equivalent residues of SF as that in AqpZ are W48, G191, F200, R206, but with G191 hiding away from the SF.

2.4 Mutations on AqpZ for Function Manipulation

Ever since subfamilies of aquaporin were discovered, scientists have not stopped trying to understand the detailed mechanism of selectivity and to exchange the molecule selectivity by manipulating some of the critical residues.

Figure 2.3. Cross section of SF size comparison in wild type rat Aqp1 and its mutants (21). The mutants of R195V(B), H180A(C), H180A/R195V(D), F56A/H180A(E), all have enlarged SF pore size compare to the wild type(A) (21).

In 2006, Beitz et al. reported that some multiple mutations of ar/R SF residues of rat AQP1 could change its selectivity (Fig 2.3). In their publication, mutation of single residue to alanine and valine among the SF residues including Arg and His expanded the SF but did not change the permeability of water molecules. Furthermore, two double mutants, AQP1-H180A/R195V and AQP1-F56A/H180A, further expanded the SF. For AQP1-H180A/R195V, the expansion of SF allows urea to pass through. Meanwhile, AQP1-F56A/H180A gives the maximal diameter enlargement of SF and shows the glycerol permeability of this mutant. However, in the mutants with Arg mutated, protons are allowed to pass through the channel because of the removal of positive charge. In this work,

the researchers proved that the ar/R SF, or constriction, plays the role as major checkpoint for molecule selectivity during transportation. (21)

Figure 2.4 (A) Schematic presentation of Arg of SF in wild type AqpZ and GlpF. (B) Pore size of wild type AqpZ and its mutants compared to GlpF. Water (C) and glycerol (D) raw curve an osmotic challenge assay along with time in wild type AqpZ and its mutants compared to GlpF, indicating the molecular permeability of these proteins. (18)

Since the significance to molecule transportation of SF was discovered, researchers are more and more interested in manipulation of the selectivity of aquaporin family proteins. Nevertheless, it is difficult to achieve the expected goal. In 2010, Stroud et. al reported their experiment on *E. coli* AqpZ to mutate the residues of SF except the arginine to the equivalent residues of *E. coli* GlpF. They anticipated that the mutations would expand the ar/R SF and introduce glycerol permeability to AqpZ. The wild type AqpZ, GlpF and AqpZ mutants were purified and reconstituted into liposomes. The permeability of water/glycerol was measured by challenging the proteoliposome with hyperosmolar buffer respectively. As shown in Fig 2.4C and Fig 2.4D, normalized light scattering curves were obtained from

measuring the shrinkage rate of the proteoliposome. Molecular permeability constant k of water and glycerol for each kind of protein was then calculated by curve fit to an exponential function. However, the SF did not expand, in contrast, the size of protein SF was even smaller comparing with the wild type, which was unexpected. In their publication, the researchers stated that the major determinant of selectivity for larger amphipathic molecules such as glycerol was the size of selectivity filter. Meanwhile, they compared the crystal structure between AqpZ and GlpF and figured Furthermore, they also revealed that some of the residues that do not directly interact with the substrates could be necessary for selectivity determination, such as some residues located in helices-connecting loops.(18)

Table 2.2. Proteoliposome permeability for water and glycerol. Wild type AqpZ and GlpF shows their good permeability to water and glycerol respectively. The mutants however, perform poor permeability to both water and glycerol (18).

	$k_{\text{wat}}, \text{s}^{-1}$	$k_{\text{gly}}, \text{s}^{-1}$
WT AqpZ	102.9 ± 1	0.102 ± 0.001
F43W	56.8 ± 3	0.105 ± 0.001
H174G/T183F	34.8 ± 0.2	0.100 ± 0.001
F43W/H174G/T183F	12.3 ± 0.1	0.107 ± 0.004
GlpF	20.2 ± 0.4	$4.82 \pm .06$
Empty liposomes	6.3 ± 0.3	0.094 ± 0.002

In previous research, scientists focused on the residues that forms SF and conducted mutation trials on them only. However, the influence of other “less-critical” surrounding residues were neglected. Therefore, the study of molecular selectivity of aquaporin family proteins was not comprehensive. In this work, SCA is applied to discover the correlated residues and further explore the detailed interactions between residues and the molecule-selective mechanism.

2.5 Overview of Proteases and Ligases

Proteases and ligases are two of the most important protein families for living organisms. Although they have opposite functionality, they both play significant role in metabolic pathways.

2.5.1 *Proteases*

A protease, sometimes also called a peptidase or a proteinase, is an enzyme which has the main function of executing proteolysis or catabolism by hydrolysis of peptide bonds.(22) In other words, proteases cleave peptide bonds of peptides or proteins for protein degradation, signaling or metabolic purposes. Proteases have been evolved multiple times and adapted to various conditions like different pH and reductive environment. Therefore, they are found to be existed in almost all kinds of organisms including virus. Meanwhile, proteases use different catalytic mechanisms to conduct hydrolysis which classified them into several categories including serine, cysteine and threonine proteases.(23) Proteases could cleave protein substrates either from terminal or from the middle, which gives three subfamilies as aminopeptidase, carboxypeptidase and endopeptidase.

Proteases regulate the location and activity of various proteins, coordinate protein-protein interactions activate zymogens, and contribute to cellular signaling.(24) Therefore, proteases can influence many cellular processes, including DNA replication and transcription, cell proliferation, unfolded/misfolded protein degradation, inflammation, programmed cell death etc.(25) They are so important for metabolic pathways that dysfunction of some proteases in certain organs could cause particular diseases. So far, dysfunction of proteases lead to disordered proteolysis, which are found to be responsible for multiple pathological conditions such as cancer, cardiovascular, neurodegenerative, inflammatory, and some other diseases.(26)

2.5.2 *Ligases*

Ligases, or in other names, synthetase or synthase, are enzymes that usually bind two chemicals or compounds by forming new covalent bonds or creating ring structure coupled with high energy bond breakdown like ATP, GTP. The ligases can be classified as DNA ligase, RNA ligase, ubiquitin ligase, ATP synthase, etc. based on their substrates, products, or mechanisms. They are all very important for metabolisms of almost all organism. DNA ligases are critical for DNA replication and repair(27), which is one of the most studied ligases at present. Ubiquitin ligase recognizes protein substrate and catalyzes the ubiquitin transfer to substrate and leads to ubiquitination. It regulates homeostasis, cell cycle and

DNA repair.(28) Thus, ubiquitin ligase is a significant target for cancer research.

Protein ligation could play a significant role in protein engineering research. By adding specific peptide to a recombinant protein or ligate two recombinant proteins would be very powerful in subsequent research. For example, by ligation, a biophysical probe could be attached to a protein for characterization; some unnatural amino acid or peptides could also be introduced to a protein for altering protein structure or function.

2.6 Review of Cyclotides and Butelase1

Cyclotides are a small group of disulfide-rich peptides which have a peptide bond linkage of their N- and C-terminal, leading to a circular backbone (29, 30). This particular structure provides cyclotides remarkable structural and proteolytic stability comparing with their linear counterparts (31). The number of amino acids which form a cyclotide could range from 6-11 residues (32) to over 70 residues (33, 34). Generally speaking, smaller cyclotides (<15 residues) are often found in microorganisms, and have no well-defined three-dimensional structure (31). In contrast, larger cyclotides (14-78 amino acids), distributed in bacteria, plants, as well as animals, with well-defined 3D structures. It is reported that because of their small size and outstanding stability, cyclotides could modulate protein-protein interactions in different ways, e.g. binding to surface of large proteins with high affinity and specificity (35, 36). Since cyclotides have various therapeutically valuable functions, including antimicrobial and anti-HIV activity, they could be a promising class of peptides in clinical applications (37). Recently, several cyclotides have been FDA-approved or in clinical trials in different clinical therapies and attracted more and more interest in therapeutic research field (31).

Plants are rich resources of cyclic peptides, mainly produced through non-ribosomal biosynthetic pathway (38). By contrast, cyclotides which are gene-coded arise from large precursors and the bioprocessing involves a ligase or cyclase. One of the most important steps is to join the N and C termini to form the circular backbone. However, the in-vitro synthesis of cyclotides is not easy. Only three peptide cyclases for ribosomal peptide synthesis have been purified so far and they perform at a very low efficiency, with Kinect

K_{cat} of 1h^{-1} or 1d^{-1} . The low efficiency determines that these enzymes are not suitable for mass-production of cyclotides in industrial applications.

Asparaginyl endopeptidase (AEP), or legumain, sometimes with the name of vacuolar processing enzyme, is a family of cysteine proteases which catalyze hydrolysis of proteins or peptides specifically at the site of asparaginyl bonds (39). They were first discovered in plants during the 1980s with the name of CPR (cysteine proteinase) (40). After that, more proteins like CPR were discovered in other plants until they were given the name of legumain with the appearance in legume seeds (41, 42). The crystal structure of prolegumain and active AEP was discovered in 2013 (43), which gave us better reference in studying the function of legumains. Generally speaking, the major role for legumains was to catalyze the hydrolysis and no ligase activity was shown. Only in particular situation, when an aspartic acid transits to aspartimide through deamination, the human legumain could ligate the cleaved human cystatin E (hCE) at the cleavage site (44). In contrast, some plant legumains show ligase activity to produce cyclic peptides but in low efficiency (45).

A recent discovery of a novel plant AEP, showed that this specific plant legumain from *C. ternatea*, with the name of butelase1, could act as an Asx-specific ligase which enables peptide macrocyclization of kalata B1 (kB1) with an extraordinary high efficiency as high as 95% yield within an hour (46). As it is known, AEP or legumain is usually used as a lysosomal endopeptidase that cleaves substrate at Asp/Asn specific site (47). However, with the exact same catalytic triad and 71% sequence identity with other legumain proteases, butelase 1 does not do the hydrolysis of normal legumain substrates as its homologs do. It was claimed that the cyclization is independent of disulfide bonds, and facilitated by a C-terminal sequence of NHV or DHV (46). Furthermore, other legumains, e.g. Jack bean legumain, could not perform ligase function. Meanwhile, minor cyclodimerization was also detected during the experiment, suggesting that butelase 1 also has the ability to catalyze intermolecular peptide ligation, which was also proved to be true.

Figure 2.5 (a,b,c) Mass spectrometry characterization of peptide cyclase activity. Butelase 1 shows great yield of kB1 cyclisation (b) compare to Jack bean legumain (c). (d) Modeling structure of zymogenic butelase 1 based on the structure of human legumain (46).

Another recombinantly-expressed AEP from a cyclotide-producing plant *Oldenlandia affinis*, OaAEP1b, was also identified to have the ability of efficient cyclization recently (48). It was reported that this enzyme would facilitate the cyclization of kB1 with C-terminal sequence of NGL or DSLP with very high yield within 5 hours without involvement of water. Akin to butelase 1, the removal of disulfide bond in kB1 analogues improved the catalytic efficiency remarkably as well, suggested that OaAEP1b facilitated cyclization is also independent of disulfide bond. Although the efficiency of cyclization is not as perfect as butelase 1, OaAEP1b shows promising prospect in application since it could be fabricated by recombinant expression.

Although butelase 1 and OaAEP1b seems to be the most promising catalyzes candidates for industrial cyclotides production, the mechanism of butelase 1/ OaAEP1b triggered ligation remains undiscovered. It is also a very interesting topic to study why a protease family protein mainly catalyze ligation. In this work, SCA is applied to discover the critical correlated pairs of residues that might play important roles in facilitating ligation reaction

and prevent hydrolysis.

References

1. Zardoya R. Phylogeny and evolution of the major intrinsic protein family. *Biol Cell*. 2005;97(6):397-414.
2. Borgnia M, Nielsen S, Engel A, Agre P. Cellular and molecular biology of the aquaporin water channels. *Annu Rev Biochem*. 1999;68:425-58.
3. Hansen M, Kun JFJ, Schultz JE, Beitz E. A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J Biol Chem*. 2002;277(7):4874-82.
4. Preston GM, Carroll TP, Guggino WB, Agre P. Appearance of Water Channels in *Xenopus* Oocytes Expressing Red-Cell CHIP28 Protein. *Science*. 1992;256(5055):385-7.
5. Gonen T, Sliz P, Kistler J, Cheng YF, Walz T. Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature*. 2004;429(6988):193-7.
6. Sui HX, Han BG, Lee JK, Walian P, Jap BK. Structural basis of water-specific transport through the AQP1 water channel. *Nature*. 2001;414(6866):872-8.
7. Lee JK, Kozono D, Remis J, Kitagawa Y, Agre P, Stroud RM. Structural basis for conductance by the archaeal aquaporin AqpM at 1.68 angstrom. *P Natl Acad Sci USA*. 2005;102(52):18932-7.
8. King LS, Kozono D, Agre P. From structure to disease: The evolving tale of aquaporin biology. *Nat Rev Mol Cell Bio*. 2004;5(9):687-98.
9. Shiels A, Bassnett S. Mutations in the founder of the MIP gene family underlie cataract development in the mouse. *Nat Genet*. 1996;12(2):212-5.
10. Loonen AJM, Knoers NVAM, van Os CH, Deen PMT. Aquaporin 2 mutations in nephrogenic diabetes insipidus. *Semin Nephrol*. 2008;28(3):252-65.
11. Chen J, Wang T, Zhou YC, Gao F, Zhang ZH, Xu H, et al. Aquaporin 3 promotes epithelial-mesenchymal transition in gastric cancer. *J Exp Clin Canc Res*. 2014;33.
12. Galan-Cobo A, Ramirez-Lorca R, Echevarria M. Role of aquaporins in cell proliferation: What else beyond water permeability? *Channels*. 2016;10(3):185-201.
13. Direito I, Madeira A, Brito MA, Soveral G. Aquaporin-5: from structure to function and dysfunction in cancer. *Cellular and molecular life sciences : CMLS*. 2016;73(8):1623-40.
14. Soveral G, Casini A. Aquaporin modulators: a patent review (2010-2015). *Expert opinion on therapeutic patents*. 2017;27(1):49-62.
15. Kalaria RN, Ballard C. Overlap between pathology of Alzheimer disease and vascular dementia. *Alz Dis Assoc Dis*. 1999;13:S115-S23.
16. Wilcock DM, Vitek MP, Colton CA. Vascular Amyloid Alters Astrocytic Water and

Potassium Channels in Mouse Models and Humans with Alzheimer's Disease. *Neuroscience*. 2009;159(3):1055-69.

17. Kuppers E, Gleiser C, Brito V, Wachter B, Pauly T, Hirt B, et al. AQP4 expression in striatal primary cultures is regulated by dopamine - implications for proliferation of astrocytes. *Eur J Neurosci*. 2008;28(11):2173-82.

18. Savage DF, O'Connell JD, Miercke LJW, Finer-Moore J, Stroud RM. Structural context shapes the aquaporin selectivity filter. *P Natl Acad Sci USA*. 2010;107(40):17164-9.

19. Tajkhorshid E, Nollert P, Jensen MO, Miercke LJW, O'Connell J, Stroud RM, et al. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science*. 2002;296(5567):525-30.

20. Savage DF, Egea PF, Robles-Colmenares Y, O'Connell JD, Stroud RM. Architecture and selectivity in aquaporins: 2.5 angstrom X-ray structure of aquaporin Z. *Plos Biology*. 2003;1(3):334-40.

21. Beitz E, Wu B, Holm LM, Schultz JE, Zeuthen T. Point mutations in the aromatic/arginine region in aquaporin 1 allow passage of urea, glycerol, ammonia, and protons. *Proc Natl Acad Sci U S A*. 2006;103(2):269-74.

22. Turk B. Targeting proteases: successes, failures and future prospects. *Nature reviews Drug discovery*. 2006;5(9):785-99.

23. Oda K. New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases. *Journal of biochemistry*. 2012;151(1):13-25.

24. Davie EW, Fujikawa K, Kisiel W. The coagulation cascade: initiation, maintenance, and regulation. *Biochemistry-Us*. 1991;30(43):10363-70.

25. Koblinski JE, Ahram M, Sloane BF. Unraveling the role of proteases in cancer. *Clinica chimica acta; international journal of clinical chemistry*. 2000;291(2):113-35.

26. Lopez-Otin C, Bond JS. Proteases: multifunctional enzymes in life and disease. *J Biol Chem*. 2008;283(45):30433-7.

27. Pascal JM, O'Brien PJ, Tomkinson AE, Ellenberger T. Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature*. 2004;432(7016):473-8.

28. Lipkowitz S, Weissman AM. RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. *Nature reviews Cancer*. 2011;11(9):629-43.

29. Craik DJ, Daly NL, Bond T, Waine C. Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J Mol Biol*. 1999;294(5):1327-36.

30. Gruber CW. Global Cyclotide Adventure: A Journey Dedicated to the Discovery of Circular

Peptides from Flowering Plants. *Biopolymers*. 2010;94(5):565-72.

31. Conlan BF, Gillon AD, Craik DJ, Anderson MA. Circular Proteins and Mechanisms of Cyclization. *Biopolymers*. 2010;94(5):573-83.

32. Sivonen K, Leikoski N, Fewer DP, Jokela J. Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Appl Microbiol Biot*. 2010;86(5):1213-25.

33. Craik DJ, Daly NL, Mulvenna J, Plan MR, Trabi M. Discovery, structure and biological activities of the cyclotides. *Curr Protein Pept Sc*. 2004;5(5):297-315.

34. Craik DJ, Daly NL, Saska I, Trabi M, Rosengren KJ. Structures of naturally occurring circular proteins from bacteria. *J Bacteriol*. 2003;185(14):4011-21.

35. Yu H, Lin YS. Toward structure prediction of cyclic peptides. *Physical chemistry chemical physics : PCCP*. 2015;17(6):4210-9.

36. Nevola L, Giralt E. Modulating protein-protein interactions: the potential of peptides. *Chem Commun (Camb)*. 2015.

37. Craik DJ, Clark RJ, Daly NL. Potential therapeutic applications of the cyclotides and related cystine knot mini-proteins. *Expert Opin Inv Drug*. 2007;16(5):595-604.

38. Dutton JL, Renda RF, Waine C, Clark RJ, Daly NL, Jennings CV, et al. Conserved structural and sequence elements implicated in the processing of gene-encoded circular proteins. *J Biol Chem*. 2004;279(45):46858-67.

39. Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, et al. Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase. *J Biol Chem*. 1997;272(12):8090-8.

40. Shutov AD, Lanh DN, Vaintraub IA. Purification and Partial Characterization of Protease-B from Germinating Vetch Seeds. *Biochemistry-Moscow+*. 1982;47(5):678-85.

41. Kembhavi AA, Buttle DJ, Knight CG, Barrett AJ. The 2 Cysteine Endopeptidases of Legume Seeds - Purification and Characterization by Use of Specific Fluorometric Assays. *Arch Biochem Biophys*. 1993;303(2):208-13.

42. Ishii S. Legumain - Asparaginyl Endopeptidase. *Method Enzymol*. 1994;244:604-15.

43. Dall E, Brandstetter H. Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. *P Natl Acad Sci USA*. 2013;110(27):10940-5.

44. Dall E, Fegg JC, Briza P, Brandstetter H. Structure and mechanism of an aspartimide-dependent Peptide ligase in human legumain. *Angew Chem Int Ed Engl*. 2015;54(10):2917-21.

45. Saska I, Gillon AD, Hatsugai N, Dietzgen RG, Hara-Nishimura I, Anderson MA, et al. An asparaginyl endopeptidase mediates in vivo protein backbone cyclization. *J Biol Chem*.

2007;282(40):29721-8.

46. Nguyen GKT, Wang SJ, Qiu YB, Hemu X, Lian YL, Tam JP. Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nat Chem Biol.* 2014;10(9):732-8.

47. Zhao LX, Hua T, Crowley C, Ru H, Ni XM, Shaw N, et al. Structural analysis of asparaginyl endopeptidase reveals the activation mechanism and a reversible intermediate maturation stage. *Cell Res.* 2014;24(3):344-58.

48. Harris KS, Durek T, Kaas Q, Poth AG, Gilding EK, Conlan BF, et al. Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nat Commun.* 2015;6.

Chapter 3

Experimental Methodology

Chapter 3 presents the necessary role of the experimental methodology employed in this thesis and validates the experiment design. This chapter introduces basic background information and the necessity in this work about multiple sequence alignment, statistical coupling analysis, replica exchange molecular dynamics, as well as the network theory. Multiple sequence alignment is the foundation of performing statistical coupling analysis and applied to identify highly conserved positions or motifs. Statistical coupling analysis has been proven to be a powerful tool in finding correlated residues in proteins, and is chosen to be the main technique to discover the correlated residues in this work. Combined with network analysis, the correlated pairs are further filtered. MD simulations are employed to approve the correctness and accuracy of potential mutation sites.

3.1 Methods Selection

A complex adaptive system would basically have three typical features. First of all, the system consists of multiple heterogeneous elements with independent behaviors. Second, the elements would mutually interact with one another. And third, these interactions would lead to emergence, which means in the large-scale view, new pattern or particular characteristics appears through the interactions of individual elements.(1, 2) More than 2000 years ago, the great ancient Greek philosopher Aristotle once stated in his *Metaphysics* as, “The whole is more than the sum of its parts.” This could be considered as the first time that the concept “complexity” was come up with.(1) With the development of modern science, numerous of research objectives could be considered as a complex system. Therefore, to study the elements of a system without considering the connections between elements would be insufficient. Complexity theory had been widely employed in multiple scientific disciplines including physics, chemistry, computing, biology and social sciences.(2)

Any biological system, including DNA/RNA, proteins, cells, organs, organisms or even species and eco-systems, could be considered as a complex adaptive system. Hence complexity theory is always involved in biological research in different scales.(3) In system biology, bioinformatics and computational modeling techniques are often used in mutual interactions among individuals in biological systems.(4) Dozens of techniques and hundreds of software had been developed and applied in research of genetic/proteomic databases, metabolism flow path, protein-protein interactions and drug development.

Generally speaking, several ordinary techniques have been widely employed, including MSA-based single site/motif conservation analysis, which are most often used methods in proteomic bioinformatics research. Although these techniques are very powerful for finding out the conserved residue/motifs in protein families in order to identify the key functional residues, the mutations based on the identifications sometimes could not achieve the desired effect. Some alternation of critical residues could become totally useless for the protein function or lethal to the protein led by the effect of residues that locates even not near to the functional sites. Apparently, it is crucial not to neglect the connections between residues including the crucial ones and the ones seems nonsignificant in MSA. Therefore, to treat the system as a whole is essential for

research in complex system including protein bioinformatics.

Statistical coupling analysis is one of the most powerful tools for investigating the covariations of elements in a complex system. This technique was originally developed in 1990s especially for bioinformatic studies to quantify the energetic interactions in proteins, which makes it extremely suitable for this work of protein system.

Molecular dynamics simulation is also another powerful tool in computational modelling of biology. Either chemical or physical reactions of molecules are usually completed in an extremely short time scale. With the limitation of current scientific techniques, it is very difficult for researchers to investigate the reaction details at the molecular or atomic scale, including atom translocation, molecular configuration alteration, transition state structure etc. Hence computational MD simulation is often used for investigating these reaction details of molecules or atoms in microscopic scale. Models of atoms and algorithms to interpret forces between atoms, also known as force fields, are employed to simulate the real molecules or even larger systems. Although theoretically speaking, as long as unlimited computing power is provided, MD simulation could be used for research of system with any size, the amount of computation grows exponentially as the system size increases. Therefore, researchers at present usually employ MD for smaller systems which also shows promising outcomes and gives theoretical support for macroscopic experiments.

Nevertheless, in some situations, to determine reaction path or stable structures, the simulation need to cross the local minima of energy landscapes between different conformations, which require ergodic sampling. A standard MD simulation sometimes does not perform quite well in sampling when the energy barrier on potential energy surface between two local minima is relatively high. In order to cross the energy barrier and get a wider sampling phase, replica exchange molecular dynamics (REMD) simulation was developed with the application of umbrella sampling, which realized the random walk in specific energy interval space. Hence REMD simulation was selected for determination of protein mutant conformation.

Network theory is also a very useful tool for research of complex systems. After the computer was invented, people view it as the basic theory for computer science and it is the foundation for establishing the internet. Nowadays, people have found that the

networks exist almost everywhere, which made network theory applicable to many disciplines. As the authority of complex network research, Barabasi, wrote in his book, by using network theory and study the hidden pattern, 93% of human behavior could be predicted. It also attracts many researcher's attention in bioinformatic topics. Biologists have built multiple kinds of biological networks including protein interaction network, metabolic network, neural network, signaling network etc. With network theory in this work, to discover the critical pairs and individuals of residues based on the SCA measurements would be more precise.

Dynamic quantum clustering method is an auxiliary technique provided by Professor Marvin Weinstein. It is a general method for clustering of big data, which is an important problem that could never be neglect in most scientific or technological fields of study. In bioinformatic research, it could also have its promising application for clustering numerous of protein or gene sequences. This method is employed to recheck the SCA measurements, and to avoid any possible significant omissions.

3.2 Multiple Sequence alignment (MSA)

Multiple sequences alignment is a common technique which is frequently employed in biological research. This technique is used to align more than three presumptive homologous biological sequences, such as protein sequences, DNA and RNA sequences. The sequences are arranged into a rectangular array in order to discover residues which are homologous in a particular column, or homologous motifs which are superimposable. These conserved residues or motifs are usually considered to play important role either for protein 3D structure or function. In proteomics research, it is considered as an essential tool for protein structures and functions. It is also employed for phylogenetic inference of evolutionary history(5) and some other research of sequence analysis.(6) Many algorithms have been developed for MSA, including CLUSTALW (7), MAFFT (8), MUSCLE (9), etc.

Figure 3.1. Schematic presentation of MSA procedure. (7)

Generally speaking, a mature multiple sequence alignment usually has three steps. Firstly, pairwise alignments are performed to generate a distance matrix which presents divergence of all pairs of sequences. Secondly, a branching neighbor-joining tree is calculated from the previous matrix. Lastly, based on the guidance of NJ tree, the sequences are aligned gradually according to the branching order.(7)

In CLUSTALW alignment algorithm, compared with previously developed MSA algorithms, CLUSTALW introduces some new concepts including sequence weights and initial gap penalty which give it better accuracy. Different algorithm could be chosen according to the user's requirement in computing speed, sequence amount limitation, sequence identity preference, as well as operating system capability.

CLUSTALW and MAFFT are two progressive alignment algorithms and MUSCLE is an iterative alignment method. In this work, the data set obtained from BLAST contains more than 300 sequences with length varied from 200 to ~500 amino acids. As homolog proteins, the sequence identity is basically larger than 60% in this case. Theoretically speaking, any MSA tool could be chosen for the alignment (10). Considering the computing time as well as the availability of algorithm in VMD software package for windows system(11), CLUSTALW was chosen as the alignment algorithm.

In this work, the raw data set is collected from BLAST searches against a non-redundant protein database (12). Typical protein sequences are selected from the database for running a PSI-BLAST. Identical sequences and sequences with very high identity (normally > 90%) are removed. These filtered sequences are used as input for CLUSTALW MSA. Furthermore, sequences with too many gaps filled in (normally >30%) are also removed for minimizing the noises.

3.3 Statistical Coupling Analysis (SCA)

MSA could give researchers plenty information about highly conserved residues as well as motifs. Since conservation has long been common criteria for discover which part of the protein most likely to be critical for functions, MSA has been widely used in bioinformatic research. Nevertheless, MSA still has the limitation that it could not give information about residue covariance during evolution history directly. Recently, with the easy accessibility and the constantly improved database with increasing number of biological sequences, the information of higher dimension statistics, such as pair correlations of residues, is becoming possible and necessary for in-depth protein research.

Statistical coupling analysis was first introduced by Ranganathan et. al in 1999(13) for the quantification of the energetic interactions in the protein sequence based on MSA. Many versions of SCA algorithm were developed during the last decades. In this work, SCA tool box v3.0 is employed for the study. The basic principle is illustrated below:

The background possibility q for an amino acid a could be calculated from the mean frequency of amino acid a in all protein sequences from a redundant database. Therefore, the background possibility for the twenty natural amino acids in alphabetical

order will be:

$$q^{(a)} = (0.073, 0.025, 0.050, 0.061, 0.042, 0.072, 0.023, 0.053, 0.064, 0.089, \\ 0.023, 0.043, 0.052, 0.040, 0.052, 0.073, 0.056, 0.063, 0.013, 0.033)$$

With a given dataset of M protein sequences after alignment, the conservation of amino acid a in one independent position i with an observing frequency $f_i^{(a)}$ could be measured by the so-called Kullback-Leibler divergence(14), or relative entropy, $D_i^{(a)}$. Firstly, the probability $P_M[f_i^{(a)}]$ of observing frequency $f_i^{(a)}$ would be:

$$P_M[f_i^{(a)}] = \frac{M!}{[Mf_i^{(a)}]![M(1-f_i^{(a)})]!} [q^{(a)}]^{Mf_i^{(a)}} [1 - q^{(a)}]^{M(1-f_i^{(a)})} \quad (3.1)$$

When M is large, an approximation would be:

$$P_M[f_i^{(a)}] \cong e^{-MD_i^{(a)}} \quad (3.2)$$

Hence,

$$D_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{q^{(a)}} + (1 - f_i^{(a)}) \ln \frac{1-f_i^{(a)}}{1-q^{(a)}} \quad (3.3)$$

The value of $D_i^{(a)}$ here indicates that with the given background possibility $q^{(a)}$, how unlikely the observed frequency of amino acid a at position i would be.

For a raw alignment $x_{i,s}^{(a)}$, one pair of position i and j with amino acid a in position i and amino acid b in position j , the correlated conservation could be defined as:

$$C_{ij}^{(ab)} = [x_{i,s}^{(a)} x_{j,s}^{(b)}]_s - [x_{i,s}^{(a)}]_s [x_{j,s}^{(b)}]_s = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)} \quad (3.4)$$

In the weighted alignment of SCA:

$$\tilde{x}_{i,s}^{(a)} = \phi_{i,s}^{(a)} x_{i,s}^{(a)} \quad (3.5)$$

Where $\phi_i^{(a)}$ is the function of positional conservation $D_i^{(a)}$. Hence the correlated conservation of weighted matrix in SCA would be:

$$\tilde{C}_{i,j}^{(ab)} = \Phi_i^{(a)} \Phi_j^{(b)} C_{ij}^{(ab)} = \left| \frac{\delta D_i^{(a)}}{\delta f_i^{(a)}} \right| \left| \frac{\delta D_j^{(b)}}{\delta f_j^{(b)}} \right| C_{ij}^{(ab)} \quad (3.6)$$

And it could be inferred that the statistical energy to represent the correlation between position i and j , which is the change in the conservation of amino acid b at position j that caused by the perturbation introduced at position i , $\Delta\Delta E$:

$$\Delta\Delta E_{j,i}^{stat,b,a_i} = -\frac{1}{M} [\ln(P_M[f_j^{(b)}]) - \ln(P_M[f_{j|i}^{(b)|a_i}])] \quad (3.7)$$

Since:

$$f_{j|i}^{(b)|a_i} = \frac{f_{ij}^{(a_i b)}}{f_i^{(a_i)}} = f_j^{(b)} + \frac{f_{ij}^{(a_i b)} - f_i^{(a_i)} f_j^{(b)}}{f_i^{(a_i)}} = f_j^{(b)} + \frac{C_{ij}^{(a_i b)}}{f_i^{(a_i)}} \quad (3.8)$$

Under the assumption of $f_{j|i}^{(b)|a_i} \approx f_j^{(b)}$, expand the term:

$$-\frac{1}{M} [\ln(P_M[f_j^{(b)}])] \cong D_j^{(b)} + \frac{C_{ij}^{(a_i b)}}{f_i^{(a_i)}} \frac{\delta D_j^{(b)}}{\delta f_j^{(b)}} \quad (3.9)$$

Hence, it is concluded the statistical energy of SCA as:

$$\Delta\Delta E_{j,i}^{stat,b,a_i} \cong -\frac{1}{f_i^{(a_i)}} \frac{\delta D_j^{(b)}}{\delta f_j^{(b)}} C_{ij}^{(a_i b)} \quad (3.10)$$

Based on the statistical energy calculated, users can generate a matrix of pairwise correlation between positions in a protein sequence MSA. Accordingly, information could be inferred from the matrix.(15)

3.4 Simulation Methods

Due to the difficulty of protein multiple sites mutation in wet lab and to further investigate the mechanism of molecule selectivity, computer modelling and choose molecular dynamics simulation methods are chosen to verify the water and glycerol permeability in microscopic scale. As technique develops, many ways of MD simulation and various force-field algorithms could be employed.

First of all, protein models should be established based on their real structures. In this

work, proteins are obtained from PDB data bank, which provides most of the 3D protein crystalized structures verified by experiments. Pymol 1.6 is used for molecule model viewing, calculation of distance between residues or atoms, as well as residue mutations. HOLE 2.0 is used for protein channel lumen verification as well as the lumen radius determination. GROMACS 4.6.5 is employed as the platform to perform the MD simulations.

Secondly, several MD simulation techniques are employed, including all-atom MD simulation, umbrella sampling and replica exchange MD simulation, which will be introduced in detail hereinafter.

3.4.1. Modelling and force-fields

Modelling is the very first foundation for MD simulation. Atoms are basic elements to form ordinary matters. In MD simulation, researchers use 3D spheres of different size and weight to represent various kinds of atoms or functional groups. Meanwhile, different charges of atoms or functional groups according to experimental data are applied on these spheres. These spheres are the basic units for simulation. And according to whether the spheres represent atoms or functional groups, the simulations could be divided to all-atom MD simulation and coarse-grained MD simulation.

Figure 3.2. Force field usually contains bonded energy, comes from bond, angles and torsions, and non-bonded energy like electrostatic potential and Vdw potential.

Many kinds of interactions exist widely between atoms, including electrostatic forces, van der Waals forces, as well as bonded forces. In order to make the models more alike the natural molecules, another fundamental element to MD simulation is the

representation of these interactions as a function of atomic coordinates. The interactions are shown in the form of potential energy of the protein which could be represented by the energy function supposed to be populated at thermal equilibrium in low energy regions. The mutual interactions, or forces of atoms are related to this function, which are commonly mentioned as force fields.(16)

The potential energy function which used for a typical force field usually have two parts, the bonded energy and non-bonded energy. Apparently, bonded energy is named after the energy comes from interactions between atoms connected by covalent bonds, including bonds themselves, angles, as well as torsions (dihedral bond). Non-bonded energy commonly includes electrostatic energy, van der Waals forces etc. Generally, the function could be written as:

$$E_{total}=E_{bonded}+E_{nonbonded}=[E_{bond}+E_{angle}+E_{torsion}]+[E_{electrostatic}+E_{van\ der\ Waals}] \quad (3.11)$$

The covalent bond energy E_{bonded} coming from the stretch of the atomic bond, which is a harmonic potential that could be represented as $k_b(b-b_0)^2$ for approximation. In this representation, k_b is the force constant, b is the real bond length and b_0 is the ideal bond length. Similar to the energy of bond, the approximate energy of angle and dihedral angle could also be represented with similar function as $k_\theta(\theta-\theta_0)^2$ and $k_\phi(1-\cos(n\Phi))$. Hence:

$$E_{bonded} = k_b(b - b_0)^2 + k_\theta(\theta - \theta_0)^2 + k_\phi(1 - \cos(n\Phi)) \quad (3.12)$$

The non-bonded energy terms could also be represented by such functions. According to Coulomb's law, the electrostatic potential between two atoms i and j , could be written as:

$$E_{electrostatic} = \frac{q_i q_j}{r_{ij}} \quad (3.13)$$

And the van der Waals interactions is often represented by Lennard-Jones 6-12 potential in modelling technique according to atom-dependent constant A and C . It is usually written as:

$$E_{van\ der\ Waals} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \quad (3.14)$$

Therefore, equation (3.11) could be expanded as:

$$E_{total} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi(1 - \cos(n\Phi)) + \sum_{non-bonded} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \right] \quad (3.15)$$

Recently, many force fields have been developed for different purpose of simulations, including OPLS (17), AMBER (18), CHARMM (19), Gaussian (20), etc. Furthermore, some associated techniques, such as umbrella sampling, are also popular choices for computational modeling research.

3.4.2 All-atom molecular dynamics simulation and OPLS-aa force-field

All-atom MD simulation is one of the most commonly simulation methods nowadays. Compared with coarse-grained simulation mentioned above, all-atom simulation has much more degree of freedom, which exponentially increase the computing time. However, the model represents all the atoms in one molecule, which makes the simulation result more accurate. Furthermore, the details of mechanism in atomic level could be studied much more intuitively. Therefore, in this work, GROMACS 4.6.5 (21) and OPLS-all-atom force field (17) is employed for the investigation of protein channel size and function in molecular detail.

Figure 3.3 Simulation systems used in this work. AqpZ or GlpF models are solved in a cubic

simulation box with water model (Force-field: OPLS-aa) (Cyan: protein; red&white: water).

The models of *E.coli* AqpZ and GlpF are downloaded from PDB database with PDB ID of 1RC2 (22) and 1LDA (23). These models are directly used as the original pdb structure in GROMACS. The simulations are performed in an $8.66 \times 8.66 \times 8.66 \text{ \AA}^3$ cubic simulation box with spc216 water solvent model (24). To measure the protein channel radius, HOLE 2.0 (25) is employed to measure the radius of lumen cross section along the channel. According to previous studies, the narrowest part of proteins is the selectivity filter.

3.4.3 Umbrella sampling

Umbrella sampling is a powerful technique used in physics and chemistry simulation. Comparing with ordinary Monte Carlo sampling, umbrella sampling method is often employed to improve the sampling of a system where its potential energy landscape form hinders its ergodicity by one or more potential barrier. This method has better ability to overcome the potential barrier(s) by applying a potential that could cancel the influence of them.

Figure 3.4 Two cases of potential landscape. (A) No obvious energy barrier, sampling has ergodicity. (B) Obvious energy barrier, sampling could only be limited in a small interval without ergodicity.

To overcome these energy barrier, Torrie and Valleau first proposed this umbrella sampling method in 1977(26) to apply a harmonic biased potential on the system along the reaction to drive the system from reactant to product. Normally in umbrella sampling, a series of windows along the reaction potential cover the intermediate steps while an MD simulation is performed at each of them. With the applied biased potential, the sampling could pass through the energy barrier and gain ergodicity, which makes the sampling more accurate. In the analysis part, the effect of biased potential would be

reduced by some other technique like Weighted Histogram Analysis Method (WHAM) (27).

Figure 3.5 Schematic presentation of umbrella sampling across the potential energy landscape.

3.4.4 Replica exchange molecular dynamics (REMD) simulation

Replica exchange MD simulation is a cogent technique for protein folding research and transition state identification in computational modeling, which attracts more and more interest of researchers in the present. Generally speaking, because of the large amount of local minimum states in the potential energy surface, it is difficult to achieve accurate canonical distribution in low temperature by ordinary MD or Monte Carlo simulations within complex systems such as protein systems. It means that a random walk on the system potential surface is not applicable because of the numerous energy barriers. REMD gives a solution to allow a random walk to go over these energy barriers and have a much wider sampling of configurations by performing a generalized-ensemble simulation based on non-Boltzmann probability weight factors. The REMD method couples molecular dynamics trajectories with a temperature exchange Monte Carlo process for efficient sampling of the conformational space. A series of non-interacting copies, or replicas are run in parallel at temperatures ranging from low to high temperature with desired temperature in the range, at which the replica can easily surmount the energy barriers. Frequently, the configurations of neighboring replicas are exchanged, and this exchange is accepted or rejected based on a Metropolis acceptance criterion that guarantees the detailed balance. Normally, the period of exchange attempt would be more than 1 ps but not too large. In this work, the exchange rate was set to be 0.2, which means the replicas would exchange every 2 ps gives an average time at a

temperature of 10 ps.

Figure 3.6 A schematic presentation of how replicas exchange. Each color represents a parallel simulation process, or replica. For example, for replica 1, after the first step, it exchange with replica 2 and get accepted. This means that for the second step of simulation, it will take the trajectory from replica 2 after the first step. However, the exchange after the third step is rejected, so replica one will continue to use the trajectory from the third step to perform the next step of simulation.

3.4.5 Potential of Mean Force

When doing simulation for research, it is always very useful to know the change of free energy according to the reaction coordinate to investigate the details of reaction pathways. Therefore, potential of mean force (PMF) is developed to characterize the substrate's free energy along the reaction pathway, indicating the reaction intermediate stage, speed of transportation or selectivity. In liquid solvent state, the PMF between two molecules (or particles) within a given separation r describes an average over all the conformations of the nearby solvent.

A simple representation of PMF function $A(r)$ could be:

$$A(r) = -k_B T \ln g(r) \quad (3.16)$$

In the above equation, k_B is the Boltzmann constant, T is the temperature, and $g(r)$ is the radial distribution function of the system.(28)

For protein transmembrane channel protein, PMF is a perfect measurement for the investigation of transportation speed and permeability of small molecules passing through the lumen. Therefore, to study the permeability of aquaporins and their mutants,

it is very useful to employ PMF calculation. Through comparing the PMF regarding to different small molecules in different channel of wild type, the molecule permeability in mutants could be confirmed intuitively.

3.4.6 MARTINI Coarse-Grained MD simulation

MARTINI force field is a coarse-grained (CG) force field developed for reducing the degree of freedom thus to enhance the computing speed and extend the simulation longer (29). The model had been revised and updated for a few times since it was developed. The central principle is “four to one mapping”. A four-to-one mapping is a way used to represent the molecules in a simplified model, in which, on average, four atoms are represented by one interaction center. Nevertheless, this rule is not restrictive, as sometimes, it is appropriate to map three, five, or more atoms into one interaction center. Hydrogen atoms are neglected because of their small size and mass. And for ring structure, a different mapping is introduced as for benzene ring in Figure. 3.6B. To keep the model simple, only four types of interactions sites are considered: polar (P), nonpolar (N), apolar (C), and charged (Q). Each particle type has several subtypes, which allow for a more accurate representation of the chemical nature of the underlying atomic structure. The total number of subtypes is 18. Within a main type (P, N, C, Q), subtypes are distinguished either by a letter denoting the hydrogen-bonding capabilities (d=donor; a=acceptor; da =both; 0=none), or by a number indicating the degree of polarity/apolarity (1-5, from the lowest to highest). The mapping of these subtypes is shown in Table 3.1. (29, 30)

Figure 3.7 Types of interaction centers in MARTINI CG model (L). Mapping of different chemical structures in CG model. (29)

The force field potential function is in general similar to that of all-atom force field but with change of some constant variables. Bonded interactions between chemically connected sites are described by a weak harmonic potential $V_{bond}(R)$ with an equilibrium distance $R_{bond} = \sigma = 0.47nm$ and a force constant $K_{bond} = 1250kJmol^{-1}nm^{-2}$:

$$V_{bond}(R) = \frac{1}{2}K_{bond}(R - R_{bond})^2 \quad (3.17)$$

Similarly, another weak harmonic potential $V_{angle}(\theta)$ of cosine type is used for the angles with $\theta_0 = 180^\circ$ and $K_{angle} = 25 kJmol^{-1}rad^{-2}$:

$$V_{angle}(\theta) = \frac{1}{2}K_{angle}[\cos(\theta) - \cos(\theta_0)]^2 \quad (3.18)$$

The non-bonded interactions between interaction sites i and j are also described by the shifted Lennard-Jones (L-J) 12-6 potential energy function:

$$U_{LJ}(r) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^6 \right] \quad (3.19)$$

with σ_{ij} representing the effective minimum distance of approach between two particles and ϵ_{ij} the strength of their interactions. The strength of the interaction, determined by the value of ϵ_{ij} , is summarized in Table one for all possible interaction pairs. Ten levels of interaction are defined as follows: O, $\epsilon = 5.6$ kJ/mol; I, $\epsilon = 5.0$ kJ/mol; II, $\epsilon = 4.5$ kJ/mol; III, $\epsilon = 4.0$ kJ/mol; IV, $\epsilon = 3.5$ kJ/mol; V, $\epsilon = 3.1$ kJ/mol; VI, $\epsilon = 2.7$ kJ/mol; VII, $\epsilon = 2.3$ kJ/mol; VIII, $\epsilon = 2.0$ kJ/mol; and IX, $\epsilon = 2.0$ kJ/mol. The LJ parameter $\sigma = 0.47$ nm for all interaction levels except level IX (repulsion between charged Q and apolar C1 & C2), for which $\sigma = 0.62$ nm.

In addition to the LJ interaction, charged groups (Q type) bear a full charge $q_{i,j}$ interacting via a shifted Coulombic potential energy function:

$$U_{el}(r) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} \quad (3.20)$$

with relative dielectric constant $\epsilon_r = 15$ for explicit screening.

Table 3.1 Mapping of subtypes and interactions between subtypes in MARTINI model. (29)

TABLE 1: Interaction Matrix^a

	sub	Q				P					N				C				
		da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	III	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	I	O	II	II	II	II	II	II	II	IV	IV	V	VI	VI	VII
N	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
C	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

The MARTINI simulations described in this work were performed with the GROMACS simulation package version 4.5.5 (21). This model is designed for use close to room or physiological temperatures. In the simulations the non-bonded interactions are cut off at a distance $r_{\text{cut}} = 1.2$ nm. To avoid generation of unwanted noise, the standard shift function of GROMACS is used in which both the energy and force vanish at the cut off distance. The LJ potential is shifted from $r_{\text{shift}} = 0.9$ nm to r_{cut} , while electrostatic potential is shifted from $r_{\text{shift}} = 0$ nm to r_{cut} . Shifting of the electrostatic potential in this manner mimics the effect of a distance-dependent screening. A smaller time step of 30-40 fs is employed for stability instead of 50 fs for most systems. The parameters and example mdp files used for MARTINI coarse-grained GROMACS simulation are available at <http://md.chem.rug.nl/marrink/coarsegrain.html>.

A total number of 12 interaction centers are utilized to represent a single DPPC molecule with 51 atoms excluding hydrogen atoms (Fig 3.8). According to the previous study (31), the area per lipid for DPPC at 323 K is 0.65 nm^2 . The solvent used in the simulation is water and is also represented by one CG interaction center.

Figure 3.8 MARTINI CG model of DPPC and water. (30)

Generally, the beginning state of a simulation would be that a number of DPPC CG molecules are randomly distributed in a simulation box with the solvent water. All the simulations took place in the NpPT ensemble with constant particle number N_p , pressure P and simulation temperature T . All the MARTINI simulations in this work were performed at $T = 323$ K and $P = 1$ Bar. At this state point, DPPC remains in the biological relevant fluid state. The time step is set as 30 fs, and the output frequency is 30 ps^{-1} . The graphical images presented in this work were prepared using VMD(32).

3.5 Network theory

As ancient philosophers said, everything in the world is connected. Therefore, using network theory in scientific research has become not only a popular tool but also the trend. Network theory has been widely employed in many fields since the internet was invented and popularized.

Every natural network has its own pattern, no matter it looks random or regular. To categorize these networks according to their patterns, several parameters that represent the characteristic of networks are defined, including degree, betweenness, closeness, clustering coefficient etc.

Among these categories of networks, scale-free network is one of the most popular networks in scientific studies. Scale-free network model was first proposed by Albert-László Barabási et al. in 1999.(33) It is defined as a kind of networks whose degree distribution follows a power law, at least asymptotically. That is to say, the fraction $P(k)$

of nodes in the network having k connections to other nodes have the relationship as:

$$P(k) = ck^{-\gamma} \quad (3.21)$$

where c is a constant and γ is a parameter whose value is typically in the range $2 < \gamma < 3$. Generally speaking, in scale-free network, most of the nodes have few links, while only a small number of the nodes have numerous links with the others, which are called “Hubs”. Scale-free networks are often seen in our daily lives, including both natural and human-made systems, such as airline networks, citation networks and some society networks. Scale-free network is also a representative of hierarchical network, in which the hubs are considered to possess higher hierarchy.

Figure 3.9 A typical scale-free network and its node degree power-law distribution.

In this work, networks are established based on the SCA matrix (mentioned as SCA network herein after) from analyzing AQP and AQGP protein families. In the SCA network, residues (nodes) are connected by the SCA scores (edges). And if a pair of residues have no or very little correlation, which means the SCA score is 0 or close to 0, they are considered to be not connected. The pairs of residues with highest correlation, which could be considered as ‘critical pairs of residues’, are selected and constitute new set of networks. Through analyzing the pattern of these networks, critical residues and their correlated ones could be identified. However, the information from SCA networks is not sufficient for understanding the selectivity mechanism as well as intramolecular interactions and might give unnecessary noise. Therefore, to analyze the intramolecular interactions and to narrow down the data to channel function related information, combining SCA information with analysis of protein 3D structure is necessary.

References

1. Campbell DJ. Task Complexity - a Review and Analysis. *Acad Manage Rev.* 1988;13(1):40-52.
2. Sullivan T. Embracing Complexity. *Harvard Business Review.* 2011(September 2011).
3. Mazzocchi F. Complexity in biology - Exceeding the limits of reductionism and de using complexity theory. *Embo Rep.* 2008;9(1):10-4.
4. Combs CD, Barham SR, Sloot PMA. *Systems Biology and Health Systems Complexity. The Digital Patient: John Wiley & Sons, Inc;* 2016. p. 73-84.
5. Wang L, Jiang T. On the complexity of multiple sequence alignment. *Journal of computational biology : a journal of computational molecular cell biology.* 1994;1(4):337-48.
6. Edgar RC, Batzoglou S. Multiple sequence alignment. *Current opinion in structural biology.* 2006;16(3):368-73.
7. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research.* 1994;22(22):4673-80.
8. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research.* 2002;30(14):3059-66.
9. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5:113.
10. Do CB, Katoh K. Protein multiple sequence alignment. *Methods Mol Biol.* 2008;484:379-413.
11. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph.* 1996;14(1):33-8, 27-8.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research.* 1997;25(17):3389-402.
13. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* 1999;286(5438):295-9.
14. Kullback S. *Information Theory and Statistics: John Wiley & Sons;* 1959.
15. Ranganathan R, Rivoire O. 2012.

16. Ponder JW, Case DA. Force fields for protein simulations. *Advances in protein chemistry*. 2003;66:27-85.
17. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*. 1996;118(45):11225-36.
18. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *Journal of computational chemistry*. 2004;25(9):1157-74.
19. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry*. 2010;31(4):671-90.
20. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. *Gaussian 16*. Wallingford, CT2016.
21. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*. 2008;4(3):435-47.
22. Savage DF, Egea PF, Robles-Colmenares Y, O'Connell JD, 3rd, Stroud RM. Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z. *PLoS biology*. 2003;1(3):E72.
23. Tajkhorshid E, Nollert P, Jensen MO, Miercke LJ, O'Connell J, Stroud RM, et al. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science*. 2002;296(5567):525-30.
24. Toukan K, Rahman A. Molecular-dynamics study of atomic motions in water. *Physical review B, Condensed matter*. 1985;31(5):2643-8.
25. Smart OS, Goodfellow JM, Wallace BA. The Pore Dimensions of Gramicidin-A. *Biophys J*. 1993;65(6):2455-60.
26. Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*. 1977;23(2):187-99.
27. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *Journal of computational chemistry*. 1992;13(8):1011-21.
28. Darve E, Pohorille A. Calculating free energies using average force. *The Journal*

of Chemical Physics. 2001;115(20):9169-83.

29. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: Coarse grained model for biomolecular simulations. *J Phys Chem B*. 2007;111(27):7812-24.

30. Marrink SJ, de Vries AH, Mark AE. Coarse grained model for semiquantitative lipid simulations. *J Phys Chem B*. 2004;108(2):750-60.

31. Marrink SJ, Mark AE. Molecular dynamics simulation of the formation, structure, and dynamics of small phospholipid vesicles. *J Am Chem Soc*. 2003;125(49):15233-42.

32. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph*. 1996;14(1):33-8.

33. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-12.

Chapter 4

Structural Context of Selectivity Filter of AQP/AQGP and Residue Clusters Affecting Molecule Selectivity

Earlier attempts to induce glycerol permeation to E.coli Aqp Z by different single site mutations did not obtain the expected result. It is unlikely that any single mutation can render the expected substrate selectivity change considering the chain interactions between the pore forming residues. This chapter presents the use of statistical coupling analysis (SCA) analysis to detect co-evolved pairs of residues is implemented to analyze the two sub groups of the aquaporin family. Combined with structural analysis, networks of highly coupled pairs are established for each group. The functions of these sites involved in selectivity are analyzed. Based on the comparison of residues in Aqp Z and Glp F in structural level, some possible co-mutation sites near selectivity filter (SF) region, C-loop as well as M6-M7 linker part are proposed.

4.1 Introduction

Generally, aquaporin is a family of transmembrane channel proteins that are involved in water transportation through the cell membrane(1). According to molecule permeability, aquaporin could be divided into two groups, AQP and AQGP, with the specificity of water and glycerol respectively(2, 3).

In current research on aquaporin, researchers have paid much attention to the difference between AQP and AQGP. They would like to discover the secret of the key that creates a difference between two groups of proteins which share a high percentage of sequence identity and structure. Furthermore, to manipulate the molecule selectivity of permeation mutually is also a hot topic in aquaporin research.

Mutation is the main procedure of protein engineering and protein design nowadays. In aquaporin protein engineering research, many studies have aimed at changing the functionality of aquaporins by introducing mutations to modify the local structure of the channel or the selectivity filter. However, many of the trials did not produce the expected results because of the undesired conformational change or other physical/chemical alternation. Therefore, it is very important to consider the protein as a whole complex system, in other words, to analyze the whole picture.

Statistical coupling analysis (SCA) (4) combined with network analysis could be one of the techniques for predicting possible mutation site for function conversion. The SCA method is originally designed for identifying co-evolving residues in a protein family and will give indication of residues interactions purely through sequence analysis. By analyzing these co-evolved sites, some information of hidden connections between specific sites could be obtained that could not be discovered through single site conservation analysis. This technique has been successfully applied in some research so far, including de-novo protein structure prediction (5) as well as identification of motifs that mediate protein allosteric communications (6).

Inspired by the previous mutation experiment on Aqp Z, it would be promising to investigate the evolutionary connections between residues of aquaporin family proteins to get further details of the water/glycerol selectivity mechanism at the molecular level, and to give some other possible co-mutation sites for Aqp Z to gain the glycerol permeability by using SCA of treated bacteria AQP/AQGP sequences as well as

network construction and analysis. In protein research, although the evolutionary hidden patterns between amino acids in protein families were sometimes overlooked, they may result in many severe consequences. It is hoped that through this study, dealing with these hidden connections would be much more easier and a general procedure to tackle mutation determined dysfunction type of bio-medical obstacle could be provided.

4.2 Principle Outcomes

In the previous coupling study of aquaporin family proteins conducted by Lin Xin et al. (7), aquaporin sequences from all species were analyzed. In this study, in order to avoid the noise due to the unbalanced nature of different species, we choose only bacteria aquaporin sequences for investigating the detail of molecule conductivity.

The data set of 305 bacteria aquaporin protein sequences (including 220 AQPs and 85 AQGPs) treated after multiple sequence alignment was obtained by Lin Xin et al.(7), each sequence has 192 columns (or positions). For convenience, the residues of these positions will be represented by their corresponding residues in Aqp Z and Glp F, the two typical structures of AQP and AQGP. Detailed cross reference table is stated in appendix.

4.2.1 Single site conservation analysis

In AQGPs, there are only 8 highly conserved sites, including two NPA motif and two other sites (Gly96 in M4 & Arg206 in M7), with threshold of conservation rate being 90%. Orthodox aquaporin appears to have much more highly conserved sites. With the same threshold, 46 highly conserved positions were identified in orthodox aquaporin, including all 8 conserved positions appeared in AQGPs with identical most frequent amino acid. The only conclusion we could get from single site conservation analysis is that AQPs have much more stable sequence than AQGPs have.

4.2.2 General network analysis of Aquaporin results

An SCA network of bacteria aquaporin proteins including both AQPs and AQGPs has been generated. An 192x192 SCA score matrix is acquired after the analysis. (Fig 4.1)

The matrix is transferred into a network with 192 nodes and 18528 non-redundant links (not shown).

Figure 4.1 The multiple sequences alignment (MSA) gives a set of sequences with 192 columns each. An 192x192 matrix was given by SCA with 18528 non-redundant couples, represented in Figure 4.1, the SCA score are represented by gradient color from high to low, the dot $i_{j,k}$ represents the SCA score of two positions j and k . The value increases evenly as right shows.

The degree distribution analysis of this network shows an exponential decrease of frequency count of pairs as the correlation strength increases, which elucidates that most of the positions are not correlated (low value of SCA score), and only a few pairs of positions are statistically coupled. Furthermore, even if a cut-off value of 1.0 is set, which excluded 16605 weakly correlated pairs, the degree distribution still suggests a power-law-like distribution, with the power law function of $y = 16.204x^{-0.804}$ and the correlation factor of 0.817. Therefore these results insinuates that the correlation network of bacteria aquaporin family proteins have a scale-free property (8). In scale free network, the most correlated nodes, or hub nodes are critical to the whole network, thus the equivalent amino acids would be important for either structure or function of proteins.

Figure 4.2 A) Distribution of correlation strength. Frequency of coupling connections are counted in the interval of 0.1. The relationship between frequency counts of coupling connections and correlation strength is shown. It reveals that in Aquaporin, only a few couples of positions have high correlation while most of the positions have very weak or almost no correlation. B) Degree distribution of SCA network with threshold of 1.0.

To get further information of critical positions as well as how they could contribute to the difference between AQPs and AQGP, the SCA networks of AQPs and AQGPs with only top ranked correlation pairs of sites are established and analyzed respectively. Cytoscape 2.8.2 (9, 10) is used for network construction and network properties calculations.

4.2.3 Analysis of networks formed by strongly coupled positions of AQPs

To identify the critical positions of the proteins, we chose only the correlation pairs with highest scores to establish a much smaller network (referred as “SCA network”). For the AQPs, we selected 50 correlation pairs of positions with the highest SCA scores (>1.84) with twenty-three positions in total (Fig 4.3). Two residues, Trp14 and Trp209, appeared to be critical to this network with the highest degree and betweenness centrality value of 0.61 and 0.22 respectively. For the other nodes, the value was less than 0.1. This smaller network formed by strongly correlated positions of AQPs also displays the scale-free hierarchical pattern (8).

Figure 4.3 Top ranked SCA network and critical residues identification in tetramer structure of AQPs. (A) Sequences alignment of AQPs of different species with critical residues highlighted in red color. (B) SCA network of 50 top scored pairs of positions in AQPs, represented by equivalent residues in Aqp Z, Trp14 and Trp209 are the two residues which have the most connections (in dash boxes); (C) Tetramer structure of Aqp Z (PDB ID: 1RC2), Trp14 and Trp209 represented by colored spheres, Trp209 locates in the peripheral part of the tetramer which could act as anchor to reduce the immobility of the structure, while Trp 14 locates in the region that close to the adjacent monomer, which makes it possible that it might be structurally important for tetramer formation and stabilization

For Trp14 in Aqp Z, in the structure of Aqp Z monomer, its function for water selectivity could not be identified because the residue locates in the peripheral part and is relatively distant from the selectivity filter as well as the transporting channel. However, when put in to the tetramer, it is discovered that Trp14 in M1 of each monomer is the only residue which is close to M5 and M6 of its adjacent monomer ($<5.5\text{\AA}$) (Fig. 4.3). Thus, it implies the contribution of this residue to the formation or stabilization of tetramer configuration. In Glp F, the corresponding residue is Leu20, which is a chained residue instead of a ring-structured residue. No important function could be estimated for this residue either in monomer or tetramer.

Trp209 in Aqp Z also locates in the peripheral part of protein. Unlike Trp14, Trp209 could be considered as one of the neighbors of the NPA motif since the distance between Trp209 and Pro187 (one of the residues in NPA motif) is only $\sim 3\text{\AA}$, which makes it possible to stabilize the NPA motif by forming a T-shaped π stacking (11). Furthermore, Trp209 is also close to the gate residue Arg189 of SF, and in contact with The191 of M7 to stabilize the structure of SF. According to some of previous studies, Tryptophan could also act as anchor residue to the lipid membrane (12). Hence it is also possible that Trp209 in Aqp Z would have contributions for protein immobilization. In Glp F, the equivalent residue is Leu237, followed by residue Pro236, creates certain flexibility after the tight turn. The structural analysis of Pro236 in Glp F will be mentioned in the later part as well.

4.2.4 Analysis of Top ranked correlation pairs SCA networks of AQGs

Similarly, for the AQGs' SCA results, 50 correlation pairs of positions with the highest SCA scores (>1.84) with twenty-nine positions in total are selected. This network is constituted by a main network and two separate correlation pairs. (Fig. 4.4) Among the 25 positions in the main network, all the four residues which form the selectivity filter are found to be appeared, which are equivalent to Trp48, Gly191, Phe200 and Arg206 in Glp F.

Meanwhile, some other important positions, which contribute to the selectivity also appears in this network, including Phe135, Thr137, Asp207, Lys211 and Pro236 in Glp F. Phe135 and Thr137 are considered as the highly conserved FST triad motif locates on C-loop which is near the SF (13), while the mutation of this motif might cause loss of water permeability (14). Asp207, Lys211 and Pro236 were identified to have different physio-chemical properties in AQP and AQGP respectively (15). A double mutation of equivalent residues of Pro236 and Leu237 in an insect aquaporin led to the gain of glycerol permeability but the loss of water permeability (16). These residues also locate near Arg206, which is suggested to be involved in gating mechanisms (17), and the interactions between them and nearby residues help to stabilize the Arginine side-chain's point-up orientation in order to keep the channel open for passage of molecule (18).

Figure 4.4 Top ranked SCA network and critical residues identification in structure of AQGs. (A) Sequences alignment of AQGs of different species with critical residues highlighted in red color. (B) Top 50 scored pairs of nodes represented by equivalent residues in Glp F. Three major clusters could be observed from this network (Pink, Blue and Dark yellow), which also reflected in the structure. SF residues are labeled in purple and residues close to SF region (distance < 10Å) are labeled with dark yellow. (C) Critical residues in Glp F 3D structure (PDB ID: 1LDA), represented in sticks of different colors with respect of different clusters. Channel lumen is represented by marine dots.

However, some of the other nodes, Lys83 in Glp F, locate in the connecting part of M3 and M4, have no physical contact with the selectivity region as it locates in the intracellular part of this protein in the cell. Although the real function of such residue is not clear, since it locates far from the SF, it could be believed that these nodes would not directly engage in water/glycerol selection. Likely, 14 nodes would be excluded as they locate at least 3 amino acids away (>10Å) from the SF. The remaining 15 nodes and their equivalent amino acids would be the main objects for this study.

4.2.5 Comparison of results of DQC methods and SCA network analysis

With the data of critical positions generated from Dynamic Quantum Clustering (DQC) method, provided by Prof. Marvin Weinstein, the results could be compared.

In the DQC report, position 38, 103, 137, 155, 161, and 173 are identified as the most important sites to differentiate AQPs and AQGPs (corresponding residues indicated in Table 4.1). For position 137, in AQPs, 80/85 of this site is Isoleucine, while in AQGPs, 73/220 of this site is Isoleucine and 115/220 of this site is Valine. Due to the similarity of Isoleucine and Valine in both chemical and structural aspects, it is suggested that there would be no difference between AQPs and AQGPs in position 137. For other positions, significant difference in amino acid mutation between AQPs and AQGPs is suggested. These sites are also suggested as first candidates of co-mutation sites in the SCA network analysis as well. Therefore, the DQC result helps us to double check the results from SCA and also narrows down to some of the important residues. Although not all the positions got from DQC and SCA are exactly identical, it is believed that the two methods were consistent in co-mutation sites prediction to some extent.

Combined the mentioned two procedures, 15 residues including the SF residues that could affect pore size determination and thus the substrate selectivity are retained. These residues locate in various regions in the protein secondary structure. The overall identifications of position numbers and critical residues that might affect the selectivity filter pore size are represented in Table 4.1.

Table 4.1. Cross reference table of positions and their equivalent residues in Glp F and Aqp Z, the locations of these residues in protein are presented in the last column. The proposed mechanisms of how they affect the SF pore size are discussed in the following parts.

Position number	Equivalent residue in Glp F	Equivalent residue in Aqp Z	Locations in protein
38	Trp48	Phe43	Selectivity filter residues
145	Gly191	His174	
154	Phe200	Thr183	
160	Arg206	Arg189	
155	Ala201	Ser184	Close to SF in protein sequence
161	Asp207	Ser190	
164	Pro210	Val193	
165	Lys211	Ala194	
168	Ala214	Gln197	C loop
101	Phe135	Ala117	
103	Thr137	Asn119	M6-M7 linker
149	Gly195	Ile178	
153	Gly199	Asn182	Starting region of M8
171	Leu234	Trp206	
173	Pro236	Phe208	

4.2.6 Replica Exchange Molecular Dynamics (REMD) simulations of mutated Aqp Z and Glp F models

The replica exchange molecular dynamics simulation is a powerful tool invented to sampling different protein configurations, normally used to study the protein folding pathways (19). It could also be used for sampling the stable structure since REMD could sample a larger energy phase by passing the energy barrier comparing with traditional simulation (20). To prove the validity of the hypothesis and verify the influence of these residues to the local molecular configuration of SF, a set of REMD simulations are conducted in order to get the stable structures of Aqp Z mutant (referred as mAqpZ) with the above-mentioned residues substituted by the corresponding ones and Aqp Z mutant with only SF residues mutated, as well as the normal simulation of original structures. The original source structures of these two proteins are crystal structures

from PDB (PDB ID: Aqp Z – 1RC2; Glp F – 1LDA). A series of short MD simulations (~1 ns) was conducted on Aqp Z, Glp F, mAqpZ and the Aqp Z mutant from previous research (PDB ID: 3NKC). The Aqp Z mutant with identified residues mutated were created from the original Aqp Z structure by manually mutation via Pymol 1.5 (21). These proteins were solvated in cubic water system and embedded in a lipid bilayer consists of 116 DPPC molecules. Both normal and replica exchange MD simulation was executed with simulation package GROMACS 4.6.5 (22) for 1ns under NVT with periodical boundary condition. The conventional simulations were performed with the temperature of 300 K. A total number of 16 replicas were used with temperature from 300-500 K, which are 300, 311, 322, 334, 346, 358, 371, 384, 397, 411, 425, 439, 454, 469, 484, 500, respectively.

Figure 4.5 Channel radius of 3NKC, AqpZ, GlpF and mAqp Z along channel axis. (A) Channel radii determined by HOLE2 for all proteins plotted as a function of channel axis (z direction) position. Protein 3NKC (marine) which mutated only the SF residues has the smallest pore size. The mAqp Z (pink) with all the above-mentioned residues mutated get a pore size comparable with the Glp F channel (blue) but larger than that of original Aqp Z (red). (B) SF of Glp F (green) and mAqp Z (cyan). The channel sizes of them are comparable (orange) while the configurations of the SF residues are similar as well, indicates that the co-mutations might have effect on the SF configuration and pore size determination. (C) SF of Aqp Z (green) and 3NKC (cyan). The SF pore size of 3NKC is slightly smaller than that of original Aqp Z, indicates that it is not sufficient to enlarge the SF pore by only mutating the SF residues.

HOLE2 (23) is utilized to determine the radius of channels of the proteins, including original crystal structures and the structure with SF residues mutated after MD simulation, and structures of mAqp Z. This program reads atoms from the pdb file and sets up a van der Waals radius for each and use a Monte Carlo simulated annealing method to determine the largest available radius in every cross section along the channel.

The SF pore sizes (represented by the narrowest channel radius along the channel) of the Aqp Z mutant with mutation of only SF residues (0.513 Å) is smaller compared with that of original Aqp Z (0.596 Å), which was shown in a previous study (18). However, it is found that the pore size of mAqpZ (0.975 Å) is much larger than that of original Aqp Z; in contrast, it is closer to the pore size of Glp F (1.230 Å) with RMS of 0.411 Å. Furthermore, the configurations of SF residues in mAqpZ and Glp F are similar (Fig 4.5). Meanwhile, no π -stacking could be observed between Trp43 and Phe183. These promising results of simulation show that the co-mutation sites identified in this study do contribute to the SF pore size determination by affecting the local molecular configuration.

4.2.7 Potential of Mean Force Calculation

To further verify the validity of predicted mutation sites, a set of PMF calculation is performed. The conditions are basically identical to REMD simulation performed previously. In this set of simulations, an individual water molecule or glycerol molecule was placed at the entrance region of the lumen in the start. Harmonic forces along the channel direction (pull_coord_k = 1500kJ/mol·nm²) was applied on the molecules and pulled them to pass through respective protein channel. Trajectories were obtained from these simulations, and 80 windows (~1 Å each) were selected for further usage. For each window, an umbrella sampling simulation was performed to characterize the free energy landscape of this window. After the set of simulation was done, WHAM method was utilized for generating a PMF curve.

As in Fig. 4.6 shows, the Aqp Z mutant also shows good permeability of both water and glycerol molecules. The peak potential required for glycerol permeation of Glp F and Aqp Z mutant is ~10 kcal/mol and ~5 kcal/mol, respectively. It reveals that glycerol is more favorable to pass through the Aqp Z mutant channel comparing with Glp F. For water passage, the Aqp Z mutant also requires lower potential (~3 kcal/mol) comparing

with that of Glp F (~ 8 kcal/mol). The PMF calculation suggests that with the predicted sites mutated, the Aqp Z mutant shows favorable permeation functionality to both water and glycerol molecules.

Figure 4.6 PMF along z-axis of Glp F (black) and Aqp Z mutant (blue) of glycerol and water passage. The positions of glycerol/water (red & white) in channel protein (cyan) along z-axis is shown below the graph. For glycerol transportation (A), the highest potential required to pass through the channel is ~ 10 kcal/mol for Glp F, but only ~ 5 kcal/mol for Aqp Z mutant. For water transportation, the highest potential required to pass through the channel is ~ 8 kcal/mol for Glp F and ~ 3 kcal/mol for Aqp Z mutant. The PMF calculation reveals that the Aqp Z mutant shows excellent permeability to both glycerol and water molecules.

4.3 Discussion

4.3.1 Additional residues should be mutated based on the SF mutation

In order to induce glycerol permeability to Aqp, the strategy is to make the pore size larger but keep the original constitution and hydrophobicity states (18). The mutations of only SF molecules are proved to be not successful. Therefore, besides the SF residues, some additional co-mutation sites are proposed based on the REMD result and discuss in detail the molecular mechanism of the SF enlargement.

4.3.2 Suggested additional co-mutation sites close to SF region

Among these residues, the function of Asp207 would be considered to be keeping channel in open state. The strong electrostatic attractive force from Asp207 side-chain drags C loop close to M7, thus keep the orientation of Arg206 side-chain point to the

extracellular part. The close contact of C loop and M7 as well as the orientation of Arg206 make the pore size larger comparing with the pore size of Aqp Z. In Aqp Z, the equivalent residue is Ser190, which provides much smaller dragging forces without electric charges. Meanwhile, the carboxyl group of Asp207 also provides electrostatic attractive force (4.3Å) of the amide group of Arg206 towards the M7, which further contribute to the expansion of SF in Glp F. (Fig 4.7A)

Ala201 in Glp F, which is also in most of the bacteria GLP proteins (168/220), located adjacent to Phe200, part of SF; while in Aqp Z, the equivalent residue is Ser184 as in most of the bacteria AQP proteins (82/85). In both Glp F and AQP Z, these two residues both have close contact with the same equivalent residue – a Glutamic acid locates in M5. (Fig 4.7B) The interaction drags the structure of M5 towards the lumen center. But the interaction of this residue with adjacent residue is difference between Glp F and Aqp Z. In Aqp Z, Ser184 is also in contact with Ser190, located in M7, forcing the helix move closer to the lumen and thus decrease the size of SF. Furthermore, as a hydrogen bond acceptor, Ser184 plays an important role in water transportation. Whereas the equivalent residue Ala201 in Glp F cannot form hydrogen bond with water because of its hydrophobicity, thus water could not pass through Glp F efficiently with the interrupted water permeation pathway facilitated by a series of hydrogen bonds.

In Glp F Lys211 and Ala214 locate near Arg206 of SF, equivalent to Ala 194 and Gln197 in Aqp Z and Glp F. Lys211 interacts with Asp207 via electronic attractive force, which causes the direction of carboxyl group to point to the extracellular side. As mentioned before, Asp207 attracts the amide group of Arg206 in Glp F. Therefore, Glp F Lys211 indirectly contributes to the tilting orientation of Arg206. On the other hand, by having Ala214 instead of the larger size Gln197 in the equivalent position in Aqp Z, Glp F maintains the bulkiness of the structure nearby, regardless of the large size of Lys211 (Fig 4.7C).

Another residue, Pro210 in Glp F, locates near Asp207 and Lys211. It makes a tight turn for the helix. The unique rigidity of proline makes the nearby structure less flexible and fixes the position of Asp207 and Lys 211 to some extent. Hence the “gate” of SF will keep opening due to steady electrostatic attraction as well as rigid structure. Instead of proline in Glp F, the equivalent residue in Aqp Z is Val193. Together with Ala194, these two residues don't have such influence on the permeability of SF.

Figure 4.7 Comparison of equivalent residues in structures of Aqp Z and Glp F near SF. (A) Asp207 side chain electrostatically attracts Arg206 of SF in Glp F, causes the side-chain of Arg206 always points to the exocellular (Left), while the equivalent residue Ser190 side chain locates relatively distinct from Arg189 in Aqp Z, the side-chain of Arg189 then reduce the pore size of SF (Right), the channel lumen was represented by light-blue-color dots. (B) Ala201 (spheres) has polar interaction with Glu 152 but locates distinct from Asp207 in Glp F (Left), while the equivalent residue Ser184 (spheres) has polar contact with Glu138 as well as Ser190 in Aqp Z, which could make the structure more compact (Right). (C) Superposition of structures in helix M7 in Glp F (Green) and Aqp Z (Magenta) with critical residues highlighted. In Glp F, K211 attracts D207 with the rigidity provided by Pro210, which further stabilize the orientation of Arg206 side-chain, while Ala201 help to maintain the bulkiness of the structure.

In the starting region of M8, Pro236 in Glp F would be another important residue influencing the selectivity. Proline often plays a role as disruptor of protein regular secondary structures due to its cyclic side-chain which locks one of the dihedral angles at -60° . In this case, the Pro236 in Glp F here disrupts the α -helix of M8 and makes the structure more flexible while the M8 helix in Aqp Z remains stable with the equivalent residue being Phe208. Hence it may also appear as an important contributor to the larger pore size of Glp F. A study showed co-mutated equivalent residues of insect aquaporin AQPcic (Y&W in that case) made the protein lose the water permeability but gain the ability to conduct glycerol (16). Combined with the importance of Phe209 mentioned earlier, it may imply that in the insect mutation study, mutation of Phenylalanine could be the reason why the protein lost water permeability, while the mutation of Tryptophan could be the reason why it gains the glycerol permeability.

Figure 4.8 Function of position 173: Pro236 breaks the secondary structure of M8 in Glp F (Left), Phe208 does not affect the secondary structure in Aqp Z (Right)

4.3.3 Suggested mutation residues in C-loop and M6-M7 linker

The importance of C-loop and M6-M7 linker in delivery of water/glycerol was also emphasized in a previous study (18). C-loop is always a complicated part for alignment because the variety of loop part is huge between proteins even in the same family of proteins. However, it is counted that in this data set, the average C-loop length in AQPs was 5 amino acids shorter than that in AQGPs. This suggests that in AQPs, the structure will be more compact due to the shortness of C-loop comparing with the structure in AQGPs. Hence the pore size of AQPs would be smaller than that of AQGPs.

Speculated from the top-ranked SCA network of Glp F, the Phenylalanine and Threonine of the FST triad, which is a highly conserved motif in the aquaporin family protein in C-loop (24), are neighbors connected to Phe48 and Asp207. Phe135 possibly contributes to the gating mechanism by stabilizing the orientation of the tail of Arg206 through interacting with electrostatic forces. Meanwhile, although Thr137 does not directly contact with Arg206, it has electrostatic interactions with Asp207 (2.6Å) to lead the orientation of its side chain point towards to the extracellular part of the protein, which indirectly affect the orientation of Arg206 side chain. (Fig 4.9A) Therefore some of the residues in C-loop would also contribute to the SF pore size enlargement in AQGPs to some extent. On the other hand, the aligned positions of Phe135 and Thr137 in Aqp Z, which are Ala117 and Asn119, both have electrostatic interactions with

Arg189. (Fig 4.9B) However, these interactions cannot pull the Arginine side chain upwards without the help of other residues.

Leu234 in Glp F, locates near C-loop region of the structure, reduces the space in this region and create more flexibility of residues locates in C-loop by reducing the π -stacking interactions between ring-structured amino acids around. Furthermore, in Aqp Z, the equivalent residue Trp206 with similar bulky residues nearby, which are Phe207, Phe208 and Trp209, further squeeze the space to make the structure more compact near C-loop as well as in SF, limits the size of permeable molecules. (Fig 4.9C&D)

Figure 4.9 Interactions of C-loop residues with Arginine gate of SF in Glp F. (A) In Glp F, Phe135 attract R206 to make the carboxyl group point up to the extracellular direction, while Thr137 could not interact with R206 but D207 (left corner), which contribute to the point-up orientation of R206; (B) In Aqp Z, Ala117 and Asn 119 both interact with R189, no point-up orientation of R189 side was observed; (C) (D) Leu234 in Glp F instead of Trp206 in Aqp Z reduces the bulkiness of the nearby region and creates more flexibility for C-loop (orange).

Last but not least, another two Glycine residues in Glp F, locate in the turns of M6-M7 linker, and are also neighbors connected to Asp207 in the SCA network of AQGP (Fig 4.10). These two residues, Gly195 and Gly199, are followed by F200 of SF, also being the starting residue of M7 in aquaporin family proteins. Nevertheless, no interactions

(H-bonds, electrostatic interactions) between these two residues and SF residues are observed. It was explained in the previous research (18) that these two residues in Glp F, instead of Isoleucine and Asparagine in Aqp Z, reduce the bulkiness of the structure of M6-M7 linker, which makes the structure much more flexible. The flexibility also breaks the π interactions between Phe200 and Trp48, which was considered to be another “hidden gate” of this channel in a recent study (25). Meanwhile, the bulky M6-M7 linker makes Thr183 obtain a highly-constrained structure, led to the strained structure of SF in Aqp Z as well. Furthermore, unlike Asn182 in Aqp Z, which acts as a hydrogen bond acceptor in the permeation pathway (25), Gly199 could not form hydrogen bond with water in Glp F. It further suggests the importance of M6-M7 linker in water selectivity.

Figure 4.10 Critical residues located on M6-M7 linker. G195 and G199 in Glp F (Left) reduce the bulkiness of the turning region and create more flexibility comparing with I178 and N182 in Aqp Z (Right).

In this study, it is suggested that the residues in C-loop as well as in M6-M7 linker in this section are clearly further contribute to the enlargement of SF size pore besides the residues mentioned in previous section. However, it reveals that these residues were possible but not the necessary co-mutation sites for changing the water/glycerol conductivity, because the effect is suggested to be indirect while the coupling was also indirect to the SF residues.

4.4 Conclusion

In summary, an SCA analysis is demonstrated to identify the critical residues from

highly correlated pairs of residues in bacteria orthodox aquaporin as well as in the aquaglyceroporin proteins, by analyzing MSA treated bacteria aquaporin protein family sequences. Possible co-mutations sites for Aqp Z to introduce the function of glycerol permeability besides the SF residues are also predicted after selecting close evolutionary contact positions with SF residues combining with structural analysis. Dynamic quantum clustering method is conducted to narrow down and exclude less-important residues. These co-mutation sites, locating near the SF, might contribute to size determination of channel lumen as well the SF pore. A molecular dynamic simulation proves that the group mutation of these residues in *E. coli* Aqp Z do enlarge the SF pore size by changing the local structure to Glp-like configuration. In vivo/vitro experiment should be done to verify the influence of these mutations to protein folding process for future research.

The combination of SCA analysis and network construction and analysis had been first used in protein research. The prospect of blind, higher dimensional sequence analysis for finding the functionally adapted mutants seems feasible. Previous research regarding protein mutagenesis were mostly based on multiple sequence alignment and single site conservation analysis. Thus, the evolutionary connections between residues might be underestimated although they might more or less help key residues either structurally or functionally. More generally, this study suggests a new point of view in protein studies, as well as protein design, which is that either structural or correlation network analysis could be further utilized in protein research. Some pilot structural network analysis on proteins, had already shown the promising prospect of this technique including identification of key residues in protein folding (26, 27) or function (28), and predicting enzyme active sites (29). It suggests that this novel 2-dimensional analyzing techniques/concept would be helpful for uncovering the deep information of proteins, such as discovering the hidden driver residues of proteins (30), or signal transportation pathways inside proteins.

References

1. Zardoya R. Phylogeny and evolution of the major intrinsic protein family. *Biol Cell*. 2005;97(6):397-414.
2. Hansen M, Kun JFJ, Schultz JE, Beitz E. A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J Biol Chem*. 2002;277(7):4874-82.
3. Borgnia M, Nielsen S, Engel A, Agre P. Cellular and molecular biology of the aquaporin water channels. *Annu Rev Biochem*. 1999;68:425-58.
4. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999;286(5438):295-9.
5. Bartlett GJ, Taylor WR. Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. *Proteins*. 2008;71(2):950-9.
6. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins (vol 10, pg 59, 2003). *Nat Struct Biol*. 2003;10(3):232-.
7. Lin X, Hong T, Mu YG, Torres J. Identification of residues involved in water versus glycerol selectivity in aquaporins by differential residue pair co-evolution. *Biochimica Et Biophysica Acta-Biomembranes*. 2012;1818(3):907-14.
8. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-12.
9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.
10. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 2011;27(3):431-2.
11. Biedermannova L, Riley KE, Berka K, Hobza P, Vondrasek J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Phys Chem Chem Phys*. 2008;10(42):6350-9.
12. de Planque MRR, Bonev BB, Demmers JAA, Greathouse DV, Koeppe RE, Separovic F, et al. Interfacial anchor properties of tryptophan residues in

transmembrane peptides can dominate over hydrophobic matching effects in peptide-lipid interactions. *Biochemistry-Us*. 2003;42(18):5341-8.

13. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, et al. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*. 2000;290(5491):481-6.
14. Beitz E, Pavlovic-Djuranovic S, Yasui M, Agre P, Schultz JE. Molecular dissection of water and glycerol permeability of the aquaglyceroporin from *Plasmodium falciparum* by mutational analysis. *Proc Natl Acad Sci U S A*. 2004;101(5):1153-8.
15. Froger A, Tallur B, Thomas D, Delamarche C. Prediction of functional residues in water channels and related proteins. *Protein Sci*. 1998;7(6):1458-68.
16. Lagree V, Froger A, Deschamps S, Hubert JF, Delamarche C, Bonnec G, et al. Switch from an aquaporin to a glycerol channel by two amino acids substitution. *J Biol Chem*. 1999;274(11):6817-9.
17. Jiang J, Daniels BV, Fu D. Crystal structure of AqpZ tetramer reveals two distinct Arg-189 conformations associated with water permeation through the narrowest constriction of the water-conducting channel. *J Biol Chem*. 2006;281(1):454-60.
18. Savage DF, O'Connell JD, Miercke LJW, Finer-Moore J, Stroud RM. Structural context shapes the aquaporin selectivity filter. *P Natl Acad Sci USA*. 2010;107(40):17164-9.
19. Jani V, Sonavane UB, Joshi R. REMD and umbrella sampling simulations to probe the energy barrier of the folding pathways of engrailed homeodomain. *J Mol Model*. 2014;20(6).
20. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*. 1999;314(1-2):141-51.
21. DeLano WL, Lam JW. PyMOL: A communications tool for computational models. *Abstr Pap Am Chem S*. 2005;230:U1371-U2.
22. Bekker H, Berendsen HJC, Dijkstra EJ, Achterop S, Vondrumen R, Vanderspoel D, et al. Gromacs - a Parallel Computer for Molecular-Dynamics Simulations. *Physics Computing '92*. 1993:252-6.
23. Smart OS, Goodfellow JM, Wallace BA. The Pore Dimensions of Gramicidin-A. *Biophys J*. 1993;65(6):2455-60.
24. Fu DX, Libson A, Miercke LJW, Weitzman C, Nollert P, Krucinski J, et al.

Structure of a glycerol-conducting channel and the basis for its selectivity. *FASEB J.* 2001;15(4):A28-A.

25. Xin L, Helix-Nielsen C, Su HB, Torres J, Tang CY, Wang R, et al. Population Shift between the Open and Closed States Changes the Water Permeability of an Aquaporin Z Mutant. *Biophys J.* 2012;103(2):212-8.
26. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E.* 2002;65(6).
27. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. *Nature.* 2001;409(6820):641-5.
28. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanel D, Venger I, et al. Network analysis of protein structures identifies functional residues. *J Mol Biol.* 2004;344(4):1135-46.
29. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol.* 2003;330(4):719-34.
30. Liu YY, Slotine JJ, Barabasi AL. Controllability of complex networks. *Nature.* 2011;473(7346):167-73.

Chapter 5

Statistical Coupling Analysis in Aquaporin Evolution Study

SCA methods cannot only be applied in protein function research and key residues identification, but can also be applied in evolutionary studies. In this chapter, the results of the work on orthodox aquaporin evolution by employing SCA method and network theory are discussed. The network patterns of aquaporin subfamilies which are adjacent in phylogenetic tree show high similarity. Analysis of key residues and their correlated ones of individual subfamilies are also presented in this chapter. The analysis reveals that AQP1 and AQP4 are found to be different in some of the key sites in higher animals and lower animals. The residues of these sites show different key function features either in water flux regulation or ligand binding. Some of the positions also show relationships with particular diseases.

5.1 Introduction

Water is the source of life and the link that binds all the living beings together on this planet(1). As its main function is intercellular water transportation, aquaporin is an essential protein widely distributed in various organs and in almost all living organisms. Aquaporin is so important for living organisms that it is associated with many diseases (2) – not to mention its potential application in many industrial field like desalination of sea water and sewage treatment. Therefore, to investigate the detailed structure as well as mechanism of aquaporin functionality is extremely important. Nevertheless, it is also crucial to understand the evolutionary relationship between different kinds of aquaporin individuals and identify the historical events that led to novel or constrained function within or cross taxonomic lineages (3).

Most of the evolutionary studies are based on the analysis of genetic information with the increasing number of sequenced genome and transcriptome. The very first evolutionary research on aquaporin diverse subfamily studied the aquaporins from many species and classified aquaporin into two phylogenetic categories – AQP and AQGP based on their molecule selectivity (4). Phylogenetic studies of aquaporin family indicated that both categories evolved from prokaryote as a consequence of early gene duplication event (5). And with research continues, a phylogenetic tree of aquaporin family could be generated. Nonetheless, researchers are not clear about how the numerous aquaporin subfamilies evolved and why they have so various functions and distributions.

Statistical coupling analysis is a powerful bioinformatic tool. Its application should not be only limited in identifying key residues, but could be applied in phylogenetic studies as well. With the increasing number of various aquaporin protein sequences are available, SCA network analysis method could be applied for further evolutionary study of aquaporin family proteins, not only in category recognition but also more in evolutionary process.

In this work, a pilot evolutionary study on only orthodox aquaporin from aquaporin family is conducted. SCA method is applied and several networks established by highly-correlated residues are generated. The results are consistent with previous phylogenetic research based on genomes that the adjacent subfamilies in phylogenetic tree shares similar network

pattern and vice versa.

Figure 5.1 Phylogenetic tree of aquaporin protein family. Among this subfamilies, AQP0,1,2,4,5,6,8 and AqpZ are orthodox aquaporins, AQP3,7,9,10 and GlpF are aquaglyceroporins. Meanwhile, AQP11 and AQP12 are unclassified aquaporins.

5.2 Results and Discussion

5.2.1 SCA application on phylogenetic aspect

The data sets of each AQP subfamily, including AQP0, AQP1, AQP2, AQP4, AQP5, AQP6, AQP8, are generated with similar procedure described in the previous chapters. SCA analysis is conducted for each subfamily data set after MSA respectively. Based on the coupling index value, several networks consist of top correlated residues are built respectively. As showed in Fig 6.2, the SCA network patterns of these subfamily are revealed. They could be divided into four categories. The network pattern of AQP 8 is a whole network with two major clusters. For AQP1 and AQP4, two main networks have been identified. However, AQP0, AQP2 and AQP5 have similar network pattern as one major network and one or two much smaller separated networks. And for AQP6, the SCA result contains six separated small networks.

Figure 5.2 Top-ranked SCA network of AQP subfamilies. Subfamilies with similar network pattern are categorized.

According to the network patterns, the more related subfamilies could be classified. Amazingly, the AQP phylogenetic tree generated based on SCA network pattern is quite consistent with the phylogenetic tree generated after long-time genetic analysis. This suggests that SCA network pattern could also be employed for phylogenetic studies if sufficient data is provided.

5.2.2 SCA network analysis helps to identify evolutionary key residues in AQP1 and AQP4

According to previous research, a lot of information about the distribution in various organs of different aquaporin subfamilies could be obtained. Like in Fig 5.3 (6) shows, although AQP1 and AQP4 have various distribution in animal bodies, they are the only two AQPs distributed in Central Nervous System (CNS). Therefore, a robust co-evolution network pattern should be employed to make the proteins less vulnerable to undesired mutations.

Nature Reviews | Drug Discovery

Figure 5.3 AQP family proteins distribution in human organs.

By analyzing the network pattern of AQP1 and AQP4 protein subfamilies, both network shows scale-free characteristics with very good power-law degree distribution. Through investigating the corresponding residues of the hub positions in the network, the molecular role of these residues could be analyzed, either structurally or functionally, and discover the evolutionary significance of these residues.

Evolutionary key residues of AQP1

Position 95 and 237 are the most connected hub nodes in the AQP1 SCA network. The corresponding residues of these two positions in human Aqp 1 are Cys102 and Trp245, respectively. By align all the selected animal AQP1 sequences, it is apparent that either for position 95 or position 237 has two major kinds of residues. In terricolous higher animals including mammals, birds and reptiles, the corresponding residues of these two positions are cysteine and tryptophan respectively. However, for aquatic fishes, the corresponding residues are methionine and leucine respectively.

Figure 5.4 Cys102 in Human Aqp1 (PDB ID: 4CSK). Cys102 in magenta and the two NPA motif in orange.

It is known that methionine and leucine are generally non-reactive among the amino acids and rarely involved in protein functions. In contrast, cysteine is very reactive which could often be found in protein active and binding sites and it has good binding affinity to metals;

extracellular cysteine could form disulfide bridge and contribute to protein conformation. Tryptophan side chain are also relatively more reactive among all the amino acid residues and often involved in many kinds of functions such as binding to other ligands and interact through stacking.

In higher animals, CNS is usually more complex and exhibits much more functionality compared with lower animals. As in Fig. 5.4 shows, in human Aqp1, Cys102 located in the transmembrane region of M4 helix, near the two NPA motifs. Although not discovered yet, it is possible that the cysteine here could interact with ligands or intramolecular residues and regulate the water transportation or intermolecular signaling pathway by affecting the conformational change of NPA motifs. (7) Therefore, it could be supposed that some CNS diseases might be related to the site-mutation of AQP1. However, in fishes, the equivalent residue in this position is methionine, which is hydrophobic and relatively non-reactive. Therefore, AQP1 in fishes is not likely to possess ligand binding or regulation function in this region. It was reported that in zebra fish, AQP1 facilitate transcellular water uptake and tight junction restrict water influx.(8)

In human Aqp1, Trp245 locates in the C-terminal of the protein. Although the C-terminal region is often folded as disordered loops in cytoplasm and is not determined in the crystal structures, it is still very significant in some functions. In the previous research, the C-terminal of Aqp1 are considered as putative phosphorylation site for protein kinase A, protein kinase C, etc. It was reported that the C-terminal tail region of Aqp1 in PASM (human pulmonary artery cells) is required for cell proliferation and migration. (9)

Evolutionary key residues of AQP4

Three “hub” positions, position 47, 83 and 128, could be identified through SCA network analysis. The corresponding residues of this position in terricolous higher animals are normally methionine, cysteine and histidine. Meanwhile, in lower animals, including fishes and *Ornithorhynchus anatinus* (Platypus), the corresponding residues are mainly leucine, valine and asparagine, respectively.

These three residues locate either in the extracellular site or cytoplasmic site of the protein.

In other words, these residues are all exposed outside the membrane, which gives them higher chance to be involved in ligand binding and signaling pathways. Comparing with leucine, valine and asparagine, the corresponding residues in higher animals, methionine, cysteine and histidine are all more reactive and considered to be involved more in molecular functions.

Figure 5.5 Met70, Cys106 and His151 in human Aqp 4 (PDB ID: 3GD8), NPA motifs represented with orange color.

As Fig. 5.5 shows, in human aquaporin 4, the equivalent residues are Met70, Cys106 and His151 respectively. It is reported that the mutation of H151A and D69H will affect the antibody or IgG binding related to neuromyelitis optica (NMO) (10, 11), which means that His151 and Met70 could be either directly or indirectly involved in NMO antibody binding (12). For Cys106, no research has been conducted to suggest its molecular function yet. Nevertheless, cysteine exposed in cytoplasm is usually found to be employed for ligand binding or act as reaction center. Hence, it is reasonable to assume that Cys106 could also have significant functionalities.

In contrast, AQP4 in fishes is only expressed in sensory organs, could be involved in cell osmosis regulation and maintain physiological conditions (13). It is clear that AQP4 is

expressed in sensory organs in both higher and lower animals. Mutations in AQP4 will cause severe impairment (14). Nevertheless, higher animals have more complex sensory organs, which requires AQP4 to evolve to obtain additional functions compared with lower animals.

5.3 Conclusion

Scientists have already revealed the significant correlation of AQPs with CNS diseases. AQP4 and AQP1 are selected for SCA analysis in this work to investigate their evolution process from lower animals to higher animals. In human body, AQP1 is expressed in choroid plexus and plays important role in facilitating cerebrospinal fluid secretion and tuning pain perception. Meanwhile, AQP4 is expressed in astrocytes and is involved in water transportation, cell migration and neuroexcitation in the CNS (15). In contrast, lower animals such as fishes do not have such complex nervous system. Moreover, although ambiguous before recently, it was believed that fishes do not have the feeling of pain, which means that the AQPs are not involved in pain perception in fishes. In previous studies, it was reported that both AQP1 and AQP4 are involved in pathogenesis of multiple nervous system diseases, including Alzheimer's disease, Amyotrophic lateral sclerosis, Neuromyelitis optica, Parkinson's disease and some other neurodegenerative diseases (16-18). Some of the identification of critical residues in this work is correlated to disease. It is suggested that clinical experiment should be conducted on the other identified residues to discover the relationship between them and the CNS diseases.

In conclusion, for AQP1 and AQP4, residues in several critical positions are substituted with more reactive ones because of the higher complexity of CNS and sensory organs in higher animals. These substitution lead to the further complement of evolution of their correlated residues. Meanwhile, these residues might be critical in pathogenesis of some nervous system diseases.

References

1. Ki-Moon B. Secretary-General's Message for 2010 United Nations Water Day 2010. Available from: http://www.un.org/en/events/waterday/2010/sg_message_2010.shtml.
2. King LS, Kozono D, Agre P. From structure to disease: the evolving tale of aquaporin biology. *Nature reviews Molecular cell biology*. 2004;5(9):687-98.
3. Finn RN, Cerda J. Evolution and functional diversity of aquaporins. *The Biological bulletin*. 2015;229(1):6-23.
4. Park JH, Saier MH, Jr. Phylogenetic characterization of the MIP family of transmembrane channel proteins. *The Journal of membrane biology*. 1996;153(3):171-80.
5. Zardoya R. Phylogeny and evolution of the major intrinsic protein family. *Biology of the cell*. 2005;97(6):397-414.
6. Verkman AS, Anderson MO, Papadopoulos MC. Aquaporins: important but elusive drug targets. *Nature reviews Drug discovery*. 2014;13(4):259-77.
7. Kuang K, Haller JF, Shi G, Kang F, Cheung M, Iserovich P, et al. Mercurial sensitivity of aquaporin 1 endofacial loop B residues. *Protein science : a publication of the Protein Society*. 2001;10(8):1627-34.
8. Kwong RW, Kumai Y, Perry SF. The role of aquaporin and tight junction proteins in the regulation of water movement in larval zebrafish (*Danio rerio*). *PloS one*. 2013;8(8):e70764.
9. Lai N, Lade J, Leggett K, Yun X, Baksh S, Chau E, et al. The aquaporin 1 C-terminal tail is required for migration and growth of pulmonary arterial myocytes. *American journal of respiratory cell and molecular biology*. 2014;50(6):1010-20.
10. Pisani F, Mola MG, Simone L, Rosito S, Alberga D, Mangiatordi GF, et al. Identification of a point mutation impairing the binding between aquaporin-4 and neuromyelitis optica autoantibodies. *The Journal of biological chemistry*. 2014;289(44):30578-89.
11. Petzold A, Pittock S, Lennon V, Maggiore C, Weinshenker BG, Plant GT. Neuromyelitis optica-IgG (aquaporin-4) autoantibodies in immune mediated optic neuritis. *Journal of neurology, neurosurgery, and psychiatry*. 2010;81(1):109-11.
12. Owens GP, Ritchie A, Rossi A, Schaller K, Wemlinger S, Schumann H, et al. Mutagenesis of the aquaporin 4 extracellular domains defines restricted binding patterns of pathogenic neuromyelitis optica IgG. *The Journal of biological chemistry*. 2015;290(19):12123-34.
13. Zichichi R, Magnoli D, Montalbano G, Laura R, Vega JA, Ciriaco E, et al. Aquaporin 4 in the sensory organs of adult zebrafish (*Danio rerio*). *Brain research*. 2011;1384:23-8.
14. Nicchia GP, Ficarella R, Rossi A, Giangreco I, Nicolotti O, Carotti A, et al. D184E mutation

- in aquaporin-4 gene impairs water permeability and links to deafness. *Neuroscience*. 2011;197:80-8.
15. Papadopoulos MC, Verkman AS. Aquaporin water channels in the nervous system. *Nature reviews Neuroscience*. 2013;14(4):265-77.
 16. Hubbard JA, Szu JJ, Binder DK. The role of aquaporin-4 in synaptic plasticity, memory and disease. *Brain research bulletin*. 2017.
 17. Badaut J, Fukuda AM, Jullienne A, Petry KG. Aquaporin and brain diseases. *Biochimica et biophysica acta*. 2014;1840(5):1554-65.
 18. Foglio E, Rodella LF. Aquaporins and neurodegenerative diseases. *Current neuropharmacology*. 2010;8(2):112-21.

Chapter 6

Statistical Coupling Analysis Application in Butelase 1 Related Research

The discovery of Butelase 1 was a landmark achievement in enzyme research. Although it belongs to the legumain protease family, Butelase 1 plays a role as cyclase/ligase and shows extremely high efficiency in catalyzing ligation of two peptides comparing with previous-found enzymes. The key region contributes to preventing hydrolysis is presumed to be activation peptide region. In this chapter, the legumain family proteins including butelase 1 and OaAEP_b are analyzed using MSA as well as SCA combined with network analysis. Some potential key residues which contributes to specific ligation process are given. Their functional features in ligation process are analyzed and a mechanism at the molecular level is proposed.

6.1 Introduction

Peptide ligases are very rarely discovered but are extremely valuable in biotechnology research. Labeling proteins segmentally allows accurate investigations of protein internally(1); ligating different protein segments may create novel proteins with unique functionality. These promising processes all require stable, efficient peptide ligases as catalysts, which makes peptide ligases an essential biotechnology tool in research.

Legumain, also known as asparaginyl endopeptidase (AEP), sometimes with name of vacuolar processing enzyme (VPE) is a family of cysteine proteases. The main function of legumain is to catalyze hydrolysis of proteins or peptides specifically at the site of asparaginyl bonds. In general, legumain contains a highly conserved catalytic triad as interaction site and cleaves the protein/peptide substrate at Asp/Asn site. Nevertheless, legumain could also reverse the cleavage in some conditions. In some other cases, plant legumain shows ligase activity to produce cyclotides but in extremely low efficiency.

Recently, a novel member of legumain family proteins, butelase 1, isolated from *Clitoria ternatea*, was discovered and showed magnificent efficiency in catalyzing ligation of protein segments/peptides. Butelase 1, although shares 71% sequence identity with other legumain family proteins, does not catalyze the hydrolysis like them. It was proven that butelase 1 facilitates ligation or cyclization for substrates with broad N-terminal amino acids and a specific C-terminal sequence of NHV or DHV. The efficiency is very high with K_{cat} value of $0.04-17s^{-1}$ and the yield is normally greater than 95%(2).

Before butelase 1 was discovered, several peptide ligases were applied in scientific research, such as sortase A, PCY1, PatG and inteins. Nevertheless, these enzymes show very low efficiency as turnover number of $1h^{-1}$ or even $1day^{-1}$. The discovery of butelase 1 was an epochal success as butelase 1 increase the ligation efficiency for three or four orders of magnitude, which made it a perfect catalyst for peptide/protein ligation.

With the great efficiency and yield, the discovery of butelase 1 provided a new tool for C-terminus-specific intermolecular protein/peptide ligations. It has already been employed as catalyst in research and proven to be a valuable biotechnological tool in many aspects

including site-specific N-terminal labeling(3) and protein thioester ligation(4).

OaAEP1_b, a member of legumain family protein, isolated from another species of plant *Oldenlandia affinis*, was also discovered to be able to facilitate kBI cyclization efficiently(5). The efficiency of this protein is not as high as that of butelase 1. Nevertheless, as it could be recombinantly expressed, OaAEP1_b has its own advantage in massive application.

The discovery of these two exceptional legumain family proteins with great ligation-catalyzing ability might not be a coincidence. Since cyclotides are commonly found in some species of plants (*Rubiaceae*, *Violaceae*, and *Cucurbitaceae*), researchers might discover more and more legumain with similar functionality. Although limited data of only two known ‘ligationable’ legumain is available, SCA tools could be applied in discovering the characteristic pattern of them, identifying crucial residues and try to investigate the specific mechanisms of ligation/cyclization.

6.2 Results and Discussion

6.2.1 Single site conservation of activation peptide (AP) region

Construction of homology model of butelase 1(2) based on zymogen of human legumain suggested that butelase 1 could also divided into 3 parts like orthodox legumains: the AEP domain, the activation peptide (AP) domain and the legumain stabilization and activity modulation (LSAM) domain. By comparing the sequence similarity of butelase 1 and human legumain, it is indicated that 49.8% sequence identity and good structural alignment. Meanwhile, less than 10% sequence identity showed up for AP region, which is cleaved during activation for human legumain. However, in butelase 1-facilitated cyclization, according to mass spectroscopy, part of the AP region remains connected. Thus, it is reasonable to believe that it is possible that the AP region has some relevance to the ligase activity. In previous research, the AP region was also suggested to be related with the ligase function. Thus, in this study, the firstly focus is on the function of residues located in AP region of legumain. Furthermore, for AP region, since the random coil part is flexible and

would be far away from the reaction center, the importance of most residues belong to this part will be ignored during analysis.

In this work, the data set was prepared with the similar procedure as described in the previous chapter. The plant legumain sequences for MSA were obtained by using BLAST searches against non-redundant protein database. The protein sequence of human legumain AEP region was used as the basic query sequence for PSI-BLAST ($e < 0.005$). More than 1000 sequences were generated from PSI-BLAST. To avoid noise created by other kind of proteins, only the sequences with description as either “asparaginyl endopeptidase”, “legumain” or “vacuolar processing enzyme” were selected. Meanwhile, sequences with very high identity ($>90\%$) were removed. As a result, a data set with 323 sequences of legumain family proteins was established. These proteins are from bacteria, fungi, animals and plants. Among them, only proteins from plant were selected, which resulted in 116 sequences. To make comparison, a data set of animals legumain sequences was also obtained with 178 sequences.

Plant legumain proteins share higher similarity compared with animal legumain proteins. Nevertheless, as cyclotides have only been discovered in plants, it can be inferred that the animal legumain proteins would not have the functionality of protein ligation/cyclization. The AP region in each data set of legumain sequences were selected for analysis. The aligned animal legumain AP region are three amino acids longer comparing with that of plant legumain, which gives 62 and 59 columns for analysis respectively. The single site conservation result is represented by the figure generated by Weblogo tool (6).

Figure 6.1 Single site conservation result generated by Weblogo tool from two data set of animal and plant legumain. The proportion of alphabet in each column represents the proportion of the

residue it stands for in the corresponding position of protein sequences in this data set.

From the comparison of sequences between animal and plant legumains, it could be easily figured out that the amino-acid difference between these two groups is obvious for AP region. For example, the highly conserved motif “PSPDVP” in animal legumains corresponds to the “NQRDAD” motif in plant legumains; the helices-connecting loop “TND” in animal legumains corresponds to “APE” in plants, etc. As one of the previous study showed, in legumain, the AEP region mainly contains negative-charged amino acids on the AP region-facing surface, while the AP region contains positive-charged amino acids that could interact with the AEP region and leads to electrostatic binding (Fig 6.2). Furthermore, plant legumains have more positive charged residues in the helix-loop-helix part, thus this part could bind with the AEP region below closer compared with the animal legumains.

Figure 6.2 Schematic presentation of electrostatic status of AP and LSAM region (top) and AEP region (bottom) at pH 7.0. Blue: positive charge; red: negative charge.

6.2.2 Sequence alignment of a group of AP regions from typical legumains

To identify the critical residues in AP region, the AP region sequences of two enzymes capable of efficient cyclization (OaAEP1_b and butelase 1) and three enzymes without this function (butelase 2, Jack bean legumain and human legumain) were aligned for comparison (Fig 6.3).

The sequences of butelase 1 and OaAEP1_b have the highest sequences identity of ~57%. However, other two plant legumain only have ~35% sequence identity with butelase 1, whereas human legumain only have <10% sequence identity in this region. The sequence identity comparison indicates that legumain proteins with capability of facilitating ligation/cyclization should have high similarity in the AP region other than those with preference of hydrolysis.

Butelase1:	DPANDNNTFT	DKN-SLGTTPR	KAVSQRDADL	IHYWEKYRRA	PEGSSRKAEA	KKQLREVMAH	RMHID
OaAEP1 _b :	NPANDNYTSL	DGN-ALTPSS	IVVNQRDADL	LHLWEKFRKA	PEGSARKEVA	QTQIFKAMSH	RVHID
JackBean:	DPATVN--FP	PHNGNLEAKM	EVVNQRDAEL	LFMWQMYQRS	NHQPEKKTHI	LEQITETVKH	RNHLD
Butelase2:	DPATVN--LP	PHN-GIEAKM	EVVHQRNAEL	LFMWQMYQRS	NHLLGKKTHI	LKQIAETVKH	RNHLD
Human:	KRKASS----	FVPLPPVTHL	DLTPSPDVPL	TIMKRKLMNT	N-DLEESRQL	TEEIQRHLDA	RHLIE

Figure 6.3 MSA of a set of legumain AP region. Butelase1 and OaAEP1_b are both from plants and have the function of facilitating ligation. Jackbean legumain is also from plant but only has the function of promoting hydrolysis. Butelase2, although also comes from *Clitoria ternatea* like butelase 1, has proven to be an endopeptidase. Human legumain does not have any function of triggering ligation.

The alignment also gives the electrostatic difference between ligase and hydrolase from position 30 to 40 in AP region. The ligases apparently have much more charged residues in this region with more positive charges than hydrolases. With the fact that the AEP region has overall negative charge in the AP-region-facing surface, it is inferred that in ligase, AP region might have more compact interactions with AEP region, which could facilitate ligation.

Furthermore, the difference between butelase 1 and OaAEP1_b sequences might also be the reason why the cyclization substrates possess diverse C-terminal residues. In butelase 1, as indicated in Fig 6.4, the Trp-His π -interaction (position 33 in AP region) might lead to the preference for histidine in P1' site of the substrate. While in OaAEP1_b, the equivalent residue leucine is smaller and hydrophobic, thus would favor glycine in P1' site of the

substrate.

Figure 6.4 Demonstration of His-Trp π -interaction

6.2.3 SCA application on legumain family proteins

SCA analysis of plant legumain complete sequences

The first SCA analysis was on the whole sequences of plant legumain. A 458x458 matrix was generated after the coupling energy calculation. To find out the key residues during revolution, the range to the top-scored co-related pairs need to be narrowed down. Therefore, a network consist of top 50 correlated pairs of nodes were established. (Fig. 6.5) Most of the nodes in this network appear to be in the AP and LSAM region of the proteins, indicated that the AEP region was relatively robust during the evolution, but not the other parts, which would cause different functions besides the hydrolysis for legumains of some species. It is also suggested from this figure that two positions, position 332 and position 422 are the most connected nodes, which means that many of the other nodes co-evolve with these two positions, suggested that they could be key residues for either structural or functional roles.

Figure 6.5 Top-50 scored pairs of nodes in SCA analysis of plant legumains (whole sequence)

SCA analysis of AP region in plant legumains

The AP region of plant legumains was selected, and a separate SCA analysis was performed on this region. Similar to the whole sequence analysis, top 50 scored pairs of nodes are also chosen to build a network (Fig 6.6). From this network, it is clear that besides position 332, some other positions, like position 358, 361, 345, 323 and 327 were also connected with multiple other nodes. The corresponding amino acids in butelase1, Jack bean legumain, and the general major corresponding residues in plants. were represented in table 6.1.

Figure 6.6 Top-50 scored pairs of nodes in SCA analysis of plant legumains (AP region)

Table 6.1 Equivalent residues of the positions from Butelase 1, Jack bean legumain and the general major conserved residues in plants.

Position number	Butelase 1	OaAEP1b	Jack bean legumain	Plants
332	Y	L	M	F/M/L
358	A	S	K	S/K/R
361	M	V	N	K/M/S
327	A	A	A	A/I
303	D	N	D	N/D
323	S	N	N	N
345	R	R	K	E/R/K

Biological relevance of key residues

The biological functions of these key residues were analyzed using the homology model based on zymogen of human legumain (PDB ID: 4FGU), since it is the only legumain protein with known crystal structure.

As described in the previous research on butelase 1, the mechanism could be deduced from the PatGmac ligase mechanism. PatGmac contains a subtilisin-like core domain and an insertion which create a helix-loop-helix motif (orange region in Fig. 6.7). (7) The insertion

will act as a “latch” or “cap” to shield the cleaved intermediate from premature hydrolysis, and thus lead to cyclization. Furthermore, the insertion part looks exactly like the AP region of legumains, which possibly suggested that the butelase 1 could also perform ligase function similarly.

Figure 6.7 “Cap” segment in the AP region, represented by khaki color.

Position 332

As the most connected position in both SCA networks of whole sequences and AP region sequences, position 332 makes great difference in butelase 1 and Jack bean legumain. In Jack bean legumain, the corresponding residue of this position is methionine, which has a chained structure and hydrophobic properties. In contrast, the corresponding amino acid of this position is the ring-structured tyrosine, which could create more space between the two helices of AP region. (Fig 6.8) The space between these two helices could be critical for shape-fitting of substrates, which could be the reason why different substrates for butelase 1 would have different efficiency and yield. It could also contribute to the substrate recognition through either forming H-bond or π -stacking. Meanwhile, the hydroxyl group of tyrosine could act as a hydrogen-bond donor to prevent possible further hydrolysis from the invading water.

In general, the corresponding amino acids of position 332 in plant legumains have three options, phenylalanine, leucine and methionine. It is hypothesized that for plant legumains which have leucine and methionine in position 332, the substrate shape fitting will fail and lead to protease function only, while for plant legumains which have phenylalanine in this position, some ligase function will be performed but less efficient comparing with butelase 1 because no hydrogen-bond donor could prevent the hydrolysis of substrate intermediates. It was observed that in plants which produce cyclotides, especially plants of Violaceae and Psychotria genus, the legumains/AEP have phenylalanine in this position. This fact could probably prove the hypothesis to some extent.

Figure 6.8 Position 332 (Tyr350 in Butelase 1) in helix-loop-helix part of AP region

Position 345

One of the most correlated positions with position 332 is position 345. The common residues among all the plant legumains of this position are arginine, lysine and glutamic acid. (Table 6.1) Arginine in this position will always lead to phenylalanine in position 332, while for other two options gives multiple choices for position 332. Comparing with lysine and glutamic acid, arginine has positive charge and the highest pKa among all the amino acids, which makes it easier to interact and bind with the surface of AEP region having negative charge. In butelase 1, the co-evolving residue of position 332 is tyrosine, which could also interact with arginine through hydrogen bonding to enclose the reaction site and

block outside water, thus prevents hydrolysis.

Position 323

Position 323 locates in the highly conserved motif NQRD, which corresponds to PSPD motif in animals. In human legumain, this motif performs critical function for preventing autolysis because the motif physically located adjacent to the S1 pocket. The two proline residues, which have extreme rigidity, hinder the P1 residue serine from diving into the pocket (Fig 6.9). However, in normal plant legumains, no such structure rigidity is shown in the surrounding, which makes the P1 residue, normally asparagine, easily enter the S1 pocket. The S1 pocket would recognize this amino acid, and lead to auto-cleavage. Nevertheless, in butelase 1, the corresponding residue is serine, which makes the failure of recognition, and thus the nearby region could survive during the enzyme activation.

Figure 6.9 Representation of reactive center in legumain. Khaki: reaction site. Cyan: S1 pocket

6.3 Conclusion

Butelase 1 is a novel member of legumain family with the function of high efficient intermolecular ligation and cyclization. However, the only one known crystal structure of the legumain family is the human legumain (PDB ID: 4FGU); furthermore, only a few of proteins of legumain family are determined to have or not have the function of ligase. With

limited information, in this study, only some of the possible key residues could be proposed which could contribute to the ligase function as well as the substrate recognition. It is suggested that with the mutation of these key residues in other plant legumains, e.g, Jack bean legumain, it is possible to get a reversed function conversion from protease to ligase. Further wet lab research should be done to prove this hypothesis.

Here the mechanism is proposed of how these legumain homologues facilitate cyclization. As Fig 6.10 indicates, after the S1 pocket recognizes Asn and S2, S3 sites further binds to the substrate, the 'cap', or part of the AP region, confine the binding by attach to the His and Val with S1' and S2' sites. The confinement of binding to the substrate stabilizes the intermediate and may further keep out the water from coming into the active site and leading to hydrolysis. The active site cleaves the substrate at Asn site by creating a thioester intermediate. After that, with higher affinity of His and Val, the 'cap' removes the cleaved C-terminal part of the substrate, while the other part of substrate, binds to the S1' and S2' sites. The thioester intermediate then ligates with the GLP motif and gives the end product.

The discovery of some critical residues in this work could be significant to further studies on butelase 1/OaAEP1b. Some mutagenesis research could be conducted to prove the roles of these residues in legumain-triggered ligation. Moreover, referred to the proposed mechanism, it is possible to introduce similar residue and structure patterns into legumain proteins which are not "ligationable" to reverse their function. Therefore, produce ligase for industrial use from orthodox legumain would become reality.

However, the data is not sufficient in this study, only limited results were obtained. Only two legumain family proteins, butelase 1 and OaAEP1b, are confirmed to have the ability to facilitate protein ligase/cyclization. Moreover, modelled 3D protein structure of butelase1, instead of real crystalized structure, also gives limited information in structural analysis. More promising information to uncover the secrets of legumain-facilitated protein ligase will be obtained when sufficient sequence data and structural information is available.

Figure 6.10 Schematic presentation of proposed mechanism of butelase 1 triggered cyclization.

References

1. Cotton GJ, Muir TW. Peptide ligation and its application to protein engineering. *Chemistry & biology*. 1999;6(9):R247-56.
2. Nguyen GK, Wang S, Qiu Y, Hemu X, Lian Y, Tam JP. Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nature chemical biology*. 2014;10(9):732-8.
3. Nguyen GK, Cao Y, Wang W, Liu CF, Tam JP. Site-Specific N-Terminal Labeling of Peptides and Proteins using Butelase 1 and Thiodepsipeptide. *Angewandte Chemie*. 2015;54(52):15694-8.
4. Cao Y, Nguyen GK, Tam JP, Liu CF. Butelase-mediated synthesis of protein thioesters and its application for tandem chemoenzymatic ligation. *Chemical communications*. 2015;51(97):17289-92.
5. Harris KS, Durek T, Kaas Q, Poth AG, Gilding EK, Conlan BF, et al. Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nature communications*. 2015;6:10199.
6. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome research*. 2004;14(6):1188-90.
7. Koehnke J, Bent A, Houssen WE, Zollman D, Morawitz F, Shirran S, et al. The mechanism of patellamide macrocyclization revealed by the characterization of the PatG macrocyclase domain. *Nature structural & molecular biology*. 2012;19(8):767-72.

Chapter 7

Molecular Dynamic Simulation on Phospholipid Self-Assembly and Vesicle Fusion

To further study transmembrane proteins like Aquaporin, it is also very important to investigate the lipid membranes that transmembrane proteins distributed in. Molecular Dynamics Simulation is a powerful technique to study molecules and reactions in microscopic scale. In this chapter, molecular details including density, charge and potential of phospholipid bilayer are discussed. MD simulation on phospholipid bilayer self-assembly and vesicle fusion are conducted. The process of lipid aggregates assembly and fusion are shown. A trigger event that two fusion parts are connected by one lipid molecule with special conformation is discovered.

7.1 Introduction

Aquaporin and many other channel proteins are all transmembrane proteins embedded in the cell membranes. Transmembrane proteins have various function besides molecule transportation, they are also involved in signaling transmission and catalyze metabolism reactions (Fig 7.1). These proteins locate in biological membranes that function as barriers of living cells or organelles in eukaryotic cells (1). Therefore, to investigate the environment that membrane proteins locate in is also very important in studying membrane protein folding, immobilization and functions.

Figure 7.1 Typical structures of transmembrane proteins. The α -helix transmembrane proteins often act as recognition proteins and receptors; β -barrel structured transmembrane proteins often play roles for substrate transportation; the helical bundle proteins often act like enzymes, transporters and receptors.

As is well-known, lipid is one of the major building blocks of life. The lipids are usually defined as amphiphilic molecules, or a kind of surfactants, as they always have a hydrophilic head and one or more hydrophobic tails. In water solutions, the lipids will aggregate and form micelle, bilayer and vesicles according to different concentrations.(2) The driving force for the aggregation is the hydrophobicity. One kind of lipids, the phospholipids, constitutes the major part of cell membranes, thus becomes a very attractive target in both biochemical, physiological and biomaterial research. For example, phospholipids could form vesicles which could be used as the drug transporters in order to get much more efficient curative effects and much less side effects. Furthermore, the microstructures of lipid self-assembly are also used in biomaterial biotechnology of

designing novel soft materials(3). Therefore, research into the mechanisms of behaviors of phospholipid in solutions, including self-assembly and fusion became a very popular topic.

Computer simulation is a powerful tool for lipid aggregates (e.g. bilayer, vesicle) properties research. Direct structural or spectroscopic experimental techniques could not detect the exact conformation of every single lipid but to do an average, because of the widely existed intrinsic disorder of lipid aggregates in situ and in vitro. Generally speaking, computer simulation technique could be employed for tracking behaviors of individual molecule or atoms with molecular details and give information that wet lab experiments could not provide. Therefore, MD simulation would be an excellent approach for doing molecular investigation of research related to membrane and other lipid aggregates.

7.2 Molecular context of membrane and phospholipid bilayer

The main component of cell membranes are often phospholipid bilayers consist of many different kinds of lipids, including phosphatidylethanolamine (PE), phosphatidylcholine (PC), etc. (4) As mentioned, it is very important to study the environment that membrane proteins locate in, which is the cell membrane, or phospholipid bilayer as a simplified model. According to previous simulation studies on lipid bilayers, some basic information of DPPC and DPPE bilayers could be obtained. The average headgroup surface area of DPPC and DPPE are 0.65nm^2 and 0.52nm^2 , respectively. Meanwhile, because of the more ordered lipid tails in DPPE bilayer, the thickness of DPPC is smaller ($3.43\text{nm} < 4.00\text{nm}$). (5) The hydration state and electrostatic properties of different types of phospholipids, as well as the interactions between lipid head groups/tail and protein residues, would be essential for the investigation. (6)

The amphiphilic characteristic of phospholipids makes it very critical to study the hydration state in lipids bilayers. The water distribution of a lipid bilayer system could well reflect the hydration state in lipid membranes and it could be measured through mass density profile along the axis. It is clear that in Fig7.2, in pure DPPC bilayer system, water only appears around the lipid head groups and barely exist in the region of hydrophilic/hydrophobic interface. In other pure lipid bilayers, the mass density profile is

similar.

Figure 7.2 A) DPPC bilayer in water system. Red and white dots represent the hydrogen and oxygen atoms in water. The cyan dots represent the carbon atoms in DPPC. B) Mass density profile of DPPC bilayer membrane. Red line represents the system mass density; green and blue lines represent the mass density of DPPC lipid and water respectively. DPPC lipid shows this mass density profile because the headgroup has larger mass density compare to the tail region. This figure revealed that water molecules are prohibited inside the lipid bilayer. However, water-lipid interface exists around the DPPC head group region, which could be a very good target for study.

However, the lipid distribution and charge distribution changes according to different kinds of lipid ratios in mixed lipid bilayer. For example, with a smaller headgroup, lipid tail of PE would be more ordered comparing with PC. Therefore, the PE monolayer would be more condense and have higher density (Fig 7.3B).(7) Because of the densely packing of PE, water permeation in to the membrane is largely prohibited in PE side. Meanwhile, as Fig 7.3C shows, due to the NH_3 group instead of three CH_3 in PE, the charge density of PE would be different from that of PC.

Figure 7.3 A) Mixed POPC/POPE bilayer. Left: POPC monolayer; Right: POPE monolayer. B) Mass density profile of mixed POPC/POPE bilayer. C) Charge density profile of mixed POPC/POPE bilayer.

The asymmetry of different types of phospholipid also exist in electrostatic potential distribution, led by the different partial charge density. As Fig 7.4 shows, the value of peak electrostatic potential of POPC and POPE monolayer is 220mv and 300mv, respectively. In mixed PC/PE lipid bilayers, the increasing PE level will reduce the P-N vector tilt angle.(5) Therefore, the observed potential difference comes from the different dipole moment of the two lipid monolayers.

Figure 7.4 Electrostatic potential distribution along z axis of a POPC/POPE mixed lipid bilayer. Distance and potential are set to be zero in the point of center of mass.

Water also plays a very important role in lipid aggregates systems dynamics including self-assembly and fusion process (8, 9). Although water has been usually considered to be dispensable on large membrane interfaces, spectroscopic data indicated that water molecules near membrane surfaces shows obvious ordered behavior (10). The water

conformation in the lipid-water interface is also highly dependent on the electric charge and dipole moment of the lipid layer surface.

7.3 Simulations of lipid aggregates fusion process

7.3.1 *Fusion of vesicle and vesicle: Spontaneous fusion at stalk position*

Two MARTINI coarse-grained DPPC vesicles were put in a simulation box with a distance between each other less than 1nm (in some references this is referred as “stalk position”) and lined up in the X direction of the simulation box. A non-periodic boundary condition is given on this simulation system. Soon after the simulation begins, the parts of out layers of both vesicles which are close to each other are attracted and get together while the other parts of vesicles are repelled away (Fig 7.5. Panel A). Soon after that, the repelled parts are attracted together (Fig 7.5. Panel B) and an initial contact of fusion are formed (~300ps, different from the previous reported > 50ns). Some of the head groups of different outer monolayer of the two vesicles merge together. It is a localized phenomenon involving only a few lipid molecules (Fig 7.6.). This initial contact soon quickly expands radially, forming a so-called stalk intermediate (Fig 7.5. Panel E). The intermediate continues to fuse radially, forms and stays stable in a hemi-fusion state, which is considered to be a pre-state of complete fusion (Fig 7.5. Panel F).

Figure 7.5 Simulation of vesicle-vesicle fusion. As in Fig 3.7 shows, the colored dots presents the MARTINI interaction center; yellow, blue, pink and cyan dots represent NC3, PO4, GLYC and TAIL interaction centers respectively. Water is not shown for clarity. Snapshots are shown at $t=0$ ps (A), 30ps (B), 90ps (C), 300ps (D), 3ns (E) and 6ns (F).

Figure 7.6 Stalk formation of vesicles fusion (300ps, left and middle). The connected part rapidly expands radially (900ps, right).

7.3.2 Fusion of vesicle and bilayer

The understanding of mechanism of how vesicle and lipid bilayer membrane fuse would be also very important in vesicle carrier transportation research. Therefore, a vesicle-bilayer fusion simulation is performed. The fusion is not complete due to the limitation of the computing power of the computer cluster. However, the system reaches a hemi-fusion state after ~ 11 ns simulation. The fusion process also starts with the fluctuation of the outer layer of both vesicle and the bilayer, the two parts stay stable in a stalk position for a few nanoseconds (~ 8 ns). The bilayer is bended to form a dome-like structure; they then start to radially expand from the connecting part. The outer layers of both vesicle and bilayer membrane are fused together at the end of this simulation performance. It is obvious that if the simulation continues, the fusion would be complete and successful.

Figure 7.7 Simulation of vesicle-bilayer fusion. Water was not shown. Snapshots are shown at $t=0$ ns (A), 8ns (B), 10ns (C), 12ns (D).

7.3.3 Self-assembly of Lipid Vesicle

One of the previous theoretical studies showed that the driving force of vesicle self-assembly is the curvature elastic energy.⁽¹¹⁾ This study proposed a model of mixed surfactants-formed vesicles and indicated that in their model, the pure surfactants would most probably be stable in the state of small micelles or lamellar bilayer phases, meanwhile, the mixture of surfactants would be much more easier to form vesicles and enhanced the stabilization of them. In this study, around 25% percent of the DPPC are substituted into DPPE in downloaded DPPC bilayer. The new mixture lipid bilayer is used for the self-assembly simulation from a bilayer to a vesicle, which was successfully completed and stable after the 200ns simulation.

As Fig 7.8 demonstrates, the structure starts and is continuously performing an inward depression from the center from the time point of ~105 ns. Later on, the structure shows a bowl-like structure as the simulation was going on. After more than 150ns, the bowl-like structure starts to close into a sphere, and the transition is completed after 180ns.

Figure 7.8 Simulation of vesicle self-assembly from mixture lipid bilayer. The CG molecules with green color represent DPPE molecules while the grey-colored ones are DPPC molecules. Water was not shown for clarity. Snapshots are shown at $t=105\text{ns}$ (A), 125ns (B), 160ns (C), 180ns (D)

7.4 The Trigger Event of Fusion – One Special Lipid Molecule

In all of the fusion simulations performed in this report, including vesicle-vesicle fusion as well as the vesicle-bilayer fusion, there is one special lipid molecule in each of the simulation, which played an extremely important role in the first step of the fusion process. The lipid molecule is a randomly chosen one from either one of the vesicles, which locates in the part of one vesicle which is closest to the other. The lipid molecule id numbers in two pure DPPC lipid vesicle-vesicle fusion simulations, one pure DPPC lipid vesicle-bilayer fusion simulation and one DPPC/DPPE mixture lipid vesicle-vesicle fusion simulation are 505, 981, 132 and 547, respectively. (Fig 7.9.)

The special lipid in each case displays an identical pattern, which the two tails of the lipid molecule spread out and each tail inserts into one vesicle or bilayer like a two-sided anchor. This kind of structure pattern connected the two fusing part and made sure they were in a stalk position.

Based on the pattern of the molecule and the frequency it appears in the simulation (100%), it suggests that the appearance of the special anchor lipid molecule would be one of the necessary trigger events of the fusion process.

Figure 7.9 Critical Lipids in A) Vesicle-Vesicle Fusion I of DPPC system; B) Vesicle-Vesicle Fusion II of DPPC system; C) Vesicle-Bilayer Fusion of DPPC system; D) Vesicle-Vesicle Fusion of DPPC/DPPE mixture system

References

1. Hedin LE, Illergard K, Elofsson A. An introduction to membrane proteins. *Journal of proteome research*. 2011;10(8):3324-31.
2. Israelachvili JN, Mitchell, D. J., & Ninham, B. W. . Theory of self-assembly of lipid bilayers and vesicles. *Biochimica et Biophysica Acta (BBA)-Biomembranes*. 1997;470(2):185-201.
3. Rudolph AS. Biomaterial Biotechnology Using Self-Assembled Lipid Microstructure. *Journal of Cellular Biochemistry*. 1994;56:183-7.
4. Ingolfsson HI, Melo MN, van Eerden FJ, Arnarez C, Lopez CA, Wassenaar TA, et al. Lipid organization of the plasma membrane. *Journal of the American Chemical Society*. 2014;136(41):14554-9.
5. Leekumjorn S, Sum AK. Molecular simulation study of structural and dynamic properties of mixed DPPC/DPPE bilayers. *Biophysical journal*. 2006;90(11):3951-65.
6. Fisette O, Paslack C, Barnes R, Isas JM, Langen R, Heyden M, et al. Hydration Dynamics of a Peripheral Membrane Protein. *Journal of the American Chemical Society*. 2016;138(36):11526-35.
7. Gurtovenko AA, Vattulainen I. Lipid transmembrane asymmetry and intrinsic membrane potential: two sides of the same coin. *Journal of the American Chemical Society*. 2007;129(17):5358-9.
8. Chandler D. Interfaces and the driving force of hydrophobic assembly. *Nature*. 2005;437(7059):640-7.
9. Kasson PM, Lindahl E, Pande VS. Water ordering at membrane interfaces controls fusion dynamics. *Journal of the American Chemical Society*. 2011;133(11):3812-5.
10. Cheng JX, Pautot S, Weitz DA, Xie XS. Ordering of water molecules between phospholipid bilayers visualized by coherent anti-Stokes Raman scattering microscopy. *Proceedings of the National Academy of Sciences of the United States of America*. 2003;100(17):9826-30.
11. Safran SA, Pincus P, Andelman D. Theory of Spontaneous Vesicle Formation in Surfactant Mixtures. *Science*. 1990;248(4953):354-6.

Chapter 8

Summary and Future Work

This chapter draws together the threads of this dissertation and discusses the results obtained in previous chapters. It is revealed in this dissertation that statistical coupling analysis could be applied in key residue identification as well as evolutionary study in bioinformatics research. SCA technique, as a powerful tool, could also be used in many other research topics. Hereby, some recommended future work as extensions of current studies is proposed. Further in-depth study on the relationships between aquaporin and diseases could be done as a continuation of this work. Meanwhile, it is also promising to apply SCA method in G-protein-coupled receptor research with similar procedure.

8.1 Summary and discussion

As proposed in the previous chapters, the main objective of this dissertation is to apply SCA method for AQP/AQGP selectivity filter investigation and discover potential mutation sites that could help to manipulate the molecule selectivity. SCA method is further applied for AQP evolutionary study as well as for research on distinguishing feature of other protein families. The goal is achieved successfully and the important findings and implications are summarized as below.

8.1.1 Features of AQP/AQGP selectivity filter and potential mutation sites

As the major work, the goal is to apply a novel bioinformatic technique, statistical coupling analysis, combined with network construction and structure analysis in AQP/AQGP protein study, discover the correlated residues which contribute to the molecule permeation selectivity and discover the molecular structure base of selectivity filter and its adjacent molecule configuration, especially for E. coli AqpZ and GlpF. Overall, the goal has been achieved successfully.

In this work, a set of AQP/AQGP database is established from a BLAST search result using typical aquaporin protein sequences from each subfamily. SCA analysis is conducted on the sequences after MSA and removing redundant and noisy data. A series of networks are established consists of top ranked highly correlated residues after that. By analyzing the network pattern as well as 3D molecule structure observation, and comparing with results get from DQC method, a series of correlated residues are identified to significantly affect the SF pore size thus contribute to the molecule permeation selectivity. To prove the predictive accuracy, the mutant model of AqpZ with above mentioned correlated residues mutated is built. The pore size of mutant is enlarged comparing with that of wild type AqpZ and close to the pore size of GlpF. Furthermore, a series of REMD simulation are performed to prove that the PMF of glycerol to pass through AqpZ mutant is significantly lower comparing with wild type AqpZ but comparable with GlpF. These modelling tests prove that the prediction of residues affect the AQP/AQGP molecule permeation selectivity is relatively credible.

8.1.2 SCA method application in protease/ligase study

The second major goal of this work is to apply SCA method in protease/ligase study to investigate the secret of butelase 1 doing ligation/cyclization as a member of AEP protease family. If the key residues contribute to the function alteration are identified, it would be possible to mutate some other AEP homolog to introduce opposite function to create effective and efficient protein ligase. Because of the lack of effective information, some of the possible key residues are predicted and ligation mechanism is proposed.

As identified ligase from AEP family is rare (only butelase 1 and OaAEP1_b until recently), the information of AEP ligase is very limited. Therefore, the data set is obtained by doing a BLAST search using butelase 1 protein sequence. The procedure of preparing the data set and conducting SCA analysis is similar with the procedure employed in previous work. After that, a network consists of top ranked correlated residues is built and analyzed.

Since the AP region is presumed to be critical in the previous research, the residues of this region which have the possibility to contribute ligating feature and the structures are mainly studied. A group of protein homolog sequences including ligase butelase 1, OaAEP1_b and protease butelase 2, jackbean legumain and human legumain are aligned and compared. Several featured residues are identified and their roles in ligation process are presumed. At last, the mechanism of ligation process of butelase 1 is proposed based on the results.

8.1.3 SCA application in aquaporin evolutionary study

SCA could not only be applied in protein function or structure research, but could employed in protein evolutionary studies as well. Holding this brief, a pilot study on aquaporin evolution is conducted.

The result shows that SCA technique could also be employed as a characterizing tool like orthodox genetic methods by distinguishing different types of networks

patterns. Meanwhile, some analysis of individual networks gives

8.2 Recommended Future Works

SCA methods has been proven to be a powerful biotechnological tool in many research aspects. It has been employed in identifying residues and evolutionary conserved residue networks that mediate allosteric communication in enzymatic studies by its inventors(1). Another research shows the promising application in drug development by studying the metal binding affinity with this method(2). SCA has also been used for computational biology research like artificial protein design(3). Furthermore, it has also been proven that SCA technique could play an important role in in de novo protein structure prediction(4).

In this dissertation, the above-mentioned results imply that SCA analysis is very helpful in feature residue identification and mutation sites prediction in protein engineering. With several research on different kinds of proteins, it suggests that SCA analysis, as a powerful tool, could be employed in more and more bioinformatic research as well as complexity related studies. Meanwhile, network theory, which has been honored to be the foundation of complexity research field, is also employed in protein studies in this work. As network theory has been applied in biological research for many years, including the concept of protein-protein interaction network, metabolism network, as well as residue interaction network.

Herein, the following recommended work is proposed.

8.2.1 Investigation of the relationship between aquaporin/other proteins and diseases

As mentioned, aquaporin has been proved to have significant involvement in many diseases. Thus, it is meaningful to discover the relationship between aquaporin and these diseases. In this study, some of the important correlated pairs of residues were discovered to display their contribution in molecular selectivity. Nevertheless, the study was limited to bacteria proteins. Eukaryotic aquaporins including mammal

proteins could be selected for further SCA study. More critical sites, which are possibly related to some diseases, could be identified. Meanwhile, it is suggested to focus on the correlated residue pairs which have no obvious effects on molecular selectivity. Because they might be important to some other function such as structure stabilization, catalytic reactions, ligand-binding and signal transduction. Mutations of these residues could also play very important role in pathogenesis. Moreover, the pathological study on aquaporin could also give suggestions on drug development.

Recently, research has revealed that the inhibition or over-expression of aquaporin could cause different kinds of diseases. Besides, dysfunction of protein due to some single mutations on aquaporin is also involved in a few diseases. Nevertheless, although not discovered for aquaporin, multiple mutations of proteins might also cause protein dysfunction and lead to disease. For instance, a double mutation in β -amyloid precursor protein was found to be associate with early-onset Alzheimer's disease(5). Moreover, it is estimated that more than 6%-8% of all cancer mutations represent doublet and triplet mutations(6). Therefore, it is meaningful and promising to investigate the relationship between protein mutations and the disease they cause.

Statistical coupling analysis here could play an important role in studying multiple mutations in proteins in the aspect of evolution. By establishing the SCA network and analyze the structure of proteins, it is possible to find out the possible mutation sites which might cause disease and deduct possible pathological mechanisms. Meanwhile, to discover possible drug recognition sites and reveal novel ideas for drug development becomes easier.

8.2.2 Residue interaction network analysis of AQP family proteins

Network theory can be employed in different ways in protein research. Besides SCA network analysis, which was mainly performed in this work, network based on protein structure could also be applied for study the properties of proteins. The network established here is called the residue interaction network (RIN) (7).

Previous RIN analysis application was successful in studying the ribosome and some other enzymes. David-Eden et. al revealed that the ribosome RNA exhibits small-world network pattern in 2008 (8). Significant residues of catalytic reaction centers, mutation-intolerant residues and residues critical to protein folding were identified through this method (9-11). For aquaporin studies, similar procedure could also be employed on aquaporins with determined crystal structures.

A residue interaction network would be established with residues as nodes, while links are presented by physical or chemical interactions. In general, a distance threshold is given to determine if two residues are connected or not. Examining the network parameters such as betweenness/closeness centrality is a promising way to further confirm the significance of the findings in this work. Furthermore, more information could be acquired through this procedure, which will certainly contribute to investigations into aquaporin in-depth research.

8.2.3 SCA application in other proteins function or structural research

The SCA technique's application should not only be limited in AQP/AQGP and protease/ligase research, but can also be employed to research for many other protein families like G protein-coupled receptor (GPCR), polymerase, immunoglobulin and phosphatase. However, plenty of data would be a requirement for better accuracy of the analysis. Fortunately, with the advance of technology, more and more proteins have been sequenced and the size of accessible public database has expanded considerably. Therefore, it is a very good timing for apply SCA technique into protein research.

For example, GPCR is one of the most studied transmembrane protein families. The main function of GPCR is to regulate cellular internal signal transduction pathways and cellular responses by coupling with G proteins (12, 13). Meanwhile, because of the critical role in signaling pathways, GPCR becomes a main target for drug development. And it is estimated that around 50% of the drugs currently being used in the market utilize GPCR as their target(14). Furthermore, with the determination of structure of GPCR recently (15), the research on this specific

protein family becomes more and more intensive. By applying SCA technique, it is possible to discover the co-evolved residues as well as the critical residues for ligand binding, signaling transduction and conformational change. It is also promising to discover the hidden connections between the six subfamilies(16) to investigate how the GPCR evolve to obtain various ligand specificity in different organs and organisms.

8.2.4 Further investigation on butelase 1-facilitated ligation mechanism

The discovery of butelase 1, as a milestone in protein synthesis and proteomics research field, must have been attracting much research interest. It is meaningful and promising to study its molecular mechanism of catalyzing restriction protein ligation.

As mentioned in chapter 6, the insufficient data led to limited information could be obtained from the analysis. It is expected that more ligase-legumains could be identified as more and more plants which are able to produce cyclotides could be used as candidates for the study. Meanwhile, the crystalized protein 3D structure could also be identified. As more plant legumain proteins are purified and proved to have ligation-catalyzing function, more sequences could be employed in SCA analysis although the proteins may not have the efficiency as high as that of butelase 1. Some other correlated pairs, either in AP region or AEP region, could be discovered. With the additional crystalized structure, much more information could be obtained and contribute to the theoretical study of mechanism investigation.

Furthermore, the information of correlated residues as well as structure would also contribute to protein engineering study. Since butelase 1 is currently not able to be expressed via recombinant technique, and OaAEP1b does not have comparable efficiency, it is an alternative way to introduce ligation-catalyzing function to similar proteins by protein engineering. With deep understanding of the mechanism in the future, efficient ligase-legumain, which can be produced by recombinant expression, could be manufactured for industrial purposes.

References

1. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural biology*. 2003;10(1):59-69.
2. Bachega JF, Navarro MV, Bleicher L, Bortoleto-Bugs RK, Dive D, Hoffmann P, et al. Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. *Proteins*. 2009;77(1):26-37.
3. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. *Nature*. 2005;437(7058):579-83.
4. Bartlett GJ, Taylor WR. Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*. 2008;71(2):950-9.
5. Haass C, Lemere CA, Capell A, Citron M, Seubert P, Schenk D, et al. The Swedish mutation causes early-onset Alzheimer's disease by beta-secretase cleavage within the secretory pathway. *Nature medicine*. 1995;1(12):1291-6.
6. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. *Journal of molecular biology*. 2013;425(21):3919-36.
7. Mehta ID, Beck BW. Residue Interaction Network: An Approach to Identify Functionally Important Residues for Protein:Protein Interactions. *Biophysical Journal*. 2013;104(2, Supplement 1):390a.
8. David-Eden H, Mandel-Gutfreund Y. Revealing unique properties of the ribosome using a network based analysis. *Nucleic acids research*. 2008;36(14):4641-52.
9. Emerson IA, Gothandam KM. Network analysis of transmembrane protein structures. *Physica A: Statistical Mechanics and its Applications*. 2012;391(3):905-16.
10. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. *Physical review E, Statistical, nonlinear, and soft matter physics*. 2002;65(6 Pt 1):061910.
11. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanel D, Venger I, et al. Network analysis of protein structures identifies functional residues. *Journal of molecular biology*. 2004;344(4):1135-46.
12. Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, Filipek S. Action of molecular switches in GPCRs--theoretical and experimental studies. *Current medicinal chemistry*. 2012;19(8):1090-109.
13. King N, Hittinger CT, Carroll SB. Evolution of key cell signaling and adhesion

protein families predates animal origins. *Science*. 2003;301(5631):361-3.

14. Cheng Z, Garvin D, Paguio A, Stecha P, Wood K, Fan F. Luciferase Reporter Assay System for Deciphering GPCR Pathways. *Current chemical genomics*. 2010;4:84-91.

15. Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, et al. Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature*. 2011;477(7366):549-55.

16. Bjarnadottir TK, Gloriam DE, Hellstrand SH, Kristiansson H, Fredriksson R, Schioth HB. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics*. 2006;88(3):263-73.

APPENDIX

Table A.1 Cross reference table of positions of MSA treated data set and their corresponding residues in Glp F and Aqp Z.

	Glp F		Aqp Z	
Position	Sequence No.	Residue	Sequence No.	Residue
1	8	K	2	F
2	9	G	3	R
3	10	Q	4	K
4	11	C	5	L
5	12	I	6	A
6	13	A	7	A
7	14	E	8	E
8	15	F	9	C
9	16	L	10	F
10	17	G	11	G
11	18	T	12	T
12	19	G	13	F
13	20	L	14	W
14	21	L	15	L
15	22	I	16	V
16	23	F	17	F
17	24	F	18	G
18	25	G	19	G
19	26	V	20	C
20	27	G	21	G
21	28	C	22	S
22	29	V	23	A
23	30	A	24	V
24	31	A	25	L
25	32	L	26	A
26	33	K	27	A
27	34	V	28	G
28	35	A	GAP	-

29	36	G	29	F
30	GAP	-	35	G
31	GAP	-	36	F
32	42	W	37	A
33	43	E	38	G
34	44	I	39	V
35	45	S	40	A
36	46	V	41	L
37	47	I	42	A
38	48	W	43	F
39	49	G	44	G
40	50	L	45	L
41	51	G	46	T
42	52	V	47	V
43	53	A	48	L
44	54	M	49	T
45	55	A	50	M
46	56	I	51	A
47	57	Y	52	F
48	58	L	53	A
49	59	T	54	V
50	60	A	55	G
51	61	G	56	H
52	62	V	57	I
53	63	S	58	S
54	64	G	59	G
55	65	A	60	G
56	66	H	61	H
57	67	L	62	F
58	68	N	63	N
59	69	P	64	P
60	70	A	65	A
61	71	V	66	V
62	72	T	67	T
63	73	I	68	I
64	74	A	69	G

65	75	L	70	L
66	76	W	71	W
67	77	L	72	A
68	78	F	73	G
69	82	D	75	R
70	83	K	76	F
71	84	R	77	P
72	85	K	80	E
73	86	V	81	V
74	87	I	82	V
75	88	P	83	G
76	89	F	84	Y
77	90	I	85	V
78	91	V	86	I
79	92	S	87	A
80	93	Q	88	Q
81	94	V	89	V
82	95	A	90	V
83	96	G	91	G
84	97	A	92	G
85	98	F	93	I
86	99	C	94	V
87	100	A	95	A
88	101	A	96	A
89	102	A	97	A
90	103	L	98	L
91	104	V	99	L
92	105	Y	100	Y
93	106	G	101	L
94	107	L	102	I
95	129	V	116	F
96	130	D	117	A
97	131	L	118	S
98	132	A	119	N
99	133	G	120	G
100	134	T	121	Y

101	135	F	122	G
102	136	S	123	E
103	137	T	124	H
104	138	Y	125	S
105	139	P	126	P
106	146	V	132	L
107	147	Q	133	S
108	148	A	134	A
109	149	F	135	L
110	150	A	136	V
111	151	V	137	V
112	152	E	138	E
113	153	M	139	L
114	154	V	140	V
115	155	I	141	L
116	156	T	142	S
117	157	A	143	A
118	158	I	144	G
119	159	L	145	F
120	160	M	146	L
121	161	G	147	L
122	162	L	148	V
123	163	I	149	I
124	164	L	150	H
125	165	A	151	G
126	166	L	152	A
127	167	T	153	T
128	168	D	154	D
129	169	D	155	K
130	176	G	156	F
131	177	P	160	G
132	178	L	161	F
133	179	A	162	A
134	180	P	163	P
135	181	L	164	I
136	182	L	165	A

137	183	I	166	I
138	184	G	167	G
139	185	L	168	L
140	186	L	169	A
141	187	I	170	L
142	188	A	171	T
143	189	V	172	L
144	190	I	173	I
145	191	G	174	H
146	192	A	175	L
147	193	S	176	I
148	194	M	177	S
149	195	G	178	I
150	196	P	179	P
151	197	L	180	V
152	198	T	181	T
153	199	G	182	N
154	200	F	183	T
155	201	A	184	S
156	202	M	185	V
157	203	N	186	N
158	204	P	187	P
159	205	A	188	A
160	206	R	189	R
161	207	D	190	S
162	208	F	191	T
163	209	G	192	A
164	210	P	193	V
165	211	K	194	A
166	212	V	195	I
167	213	F	196	F
168	214	A	197	Q
169	215	W	198	G
170	233	F	205	L
171	234	L	206	W
172	235	V	207	F

173	236	P	208	F
174	237	L	209	W
175	238	F	210	V
176	239	G	211	V
177	240	P	212	P
178	241	I	213	I
179	242	V	214	V
180	243	G	215	G
181	244	A	216	G
182	245	I	217	I
183	246	V	218	I
184	247	G	219	G
185	248	A	220	G
186	249	F	221	L
187	250	A	222	I
188	251	Y	223	Y
189	252	R	224	R
190	253	K	225	T
191	254	L	226	L
192	255	I	227	L

Table A.2 Degrees and betweenness centrality of highly coupled pairs of positions in AQP SCA network.

Name	Degree	Betweenness Centrality
Trp14	21	0.61327561
Trp209	16	0.21500722
Asn182	7	0.00937951
His56	6	0.0952381
Thr49	6	0.00974026
Pro126	5	0.00108225
Ser125	4	0
Thr191	4	0
Ala27	3	0
His124	3	0
Leu99	3	0.0021645
Met50	3	0.0021645
Phe76	3	0

Ala117	2	0
Cys20	2	0
Gly18	2	0
Pro163	2	0
Ser177	2	0
Val35	2	0
Ser22	1	0
Thr171	1	0
Tyr121	1	0
Tyr223	1	0

Table A.3 Degrees and betweenness centrality of highly coupled pairs of positions in AQP SCA network.

Position 1	Position 2	SCA Score
13	174	4.0837
153	174	3.7596
162	174	3.5778
13	44	3.5545
13	19	3.4734
44	174	3.2929
13	26	3.2071
13	153	3.1571
13	34	3.0347
13	105	3.021
13	162	3.0057
13	51	3.0011
13	17	2.998
105	174	2.9964
134	174	2.9866
13	21	2.9648
13	104	2.919
104	174	2.914
19	45	2.8282
51	174	2.8136
81	174	2.8116
96	174	2.8051
13	96	2.8028
103	174	2.7959
13	45	2.7835
51	70	2.7792
153	162	2.7672

26	174	2.7542
13	70	2.7195
148	174	2.7005
13	148	2.6912
13	134	2.669
51	188	2.6196
13	142	2.6151
17	174	2.572
13	100	2.5554
104	105	2.5544
13	81	2.5405
44	153	2.536
105	153	2.52
44	51	2.5147
70	174	2.5092
34	174	2.4666
45	91	2.462
104	153	2.4497
44	105	2.4267
13	103	2.4245
44	162	2.4237
103	153	2.4212
26	51	2.4159