ON DEVELOPING A HIGH-THROUGHPUT GENE-TO-PROTEIN-TO-STRUCTURE PIPELINE AND ITS APPLICATION FOR THE PRODUCTION OF 10 CRYSTAL STRUCTURES

by

MIN ZHAO

(Under the Direction of Bi-Cheng Wang)

ABSTRACT

By now, genome-sequencing projects have contributed to the systematical identification of a large number of proteins with unknown functions. A cost effective, high-throughput method for going from gene to structure, including cloning, protein production and crystallization, will help greatly to characterize these proteins. While a general high-throughput pipeline, which was required for the structural genomics program, was developing in our laboratory, it was soon realized that individual attention were needed in working with a large portion of the targeted protein samples. For example, some crystal-directed “salvaging and rescue” procedures were needed for those targets that would not yield satisfactory results by high-throughput methods. These procedures included additional purification, protein modification and stable domain isolation. After salvaging, 4 structures have been produced successfully out of 34 selected targets that had failed in the original high-throughput pipeline. A general salvaging strategy was then developed. In the next stage, these various rescue procedures were integrated into the general high-throughput gene-to-structure pipeline to enhance the
capability of the pipeline. For an additional 35 targets, the improved pipeline produced 5 new structures. Our success rate was 14%, higher than the average success rate (7%) of Protein Structure Initiative centers by August 2004. The improved high-throughput gene-to-structure pipeline was also applied to a non-structural genomics but a functionally important target, the human p100 protein. The structure of the human p100 protein tudor domain helped to delineate the molecular basis of its function.

INDEX WORDS: high-throughput gene-to-structure pipeline, salvaging, crystal structures, structural genomics
ON DEVELOPING A HIGH-THROUGHPUT GENE-TO-PROTEIN-TO-STRUCTURE PIPELINE AND ITS APPLICATION FOR THE PRODUCTION OF 10 CRYSTAL STRUCTURES

by

MIN ZHAO
B.S., Wuhan University, China, 1994
M.S., Zhongshan (Sun Yat-sen) University, China, 1999

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA
2007
ON DEVELOPING A HIGH-THROUGHPUT GENE-TO-PROTEIN-TO-STRUCTURE PIPELINE AND ITS APPLICATION FOR THE PRODUCTION OF 10 CRYSTAL STRUCTURES

by

MIN ZHAO

Major Professor: Bi-Cheng Wang
Committee: John Rose
            William Lanzilotta

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007
DEDICATION

To my parents and my wife, for their endless love and support, no matter how far we are apart. I love you and miss you.
ACKNOWLEDGEMENTS

It is a great pleasure to thank all the people who helped to make this dissertation possible. First of all, I would like to express my gratitude to my supervisor, Dr. B.C. Wang, for his ongoing support and enthusiastic guidance. With his patience, inspiration and encouragement, he helped to make the past four years my life’s most treasured. I really appreciated the chance to learn from him and work in his Lab. I wish to thank Dr. Zhi-jie Liu for his great efforts to explain things clearly and simply, his ideas and kindly assistance with all types of crystallography problems. It was my honor to be a member of his data collection team. I sincerely thank my committee members, Dr. John Rose and Dr. William Lanzilotta for their patience, suggestion and criticism. I am grateful to Miss Jessie Chang who worked with me for almost a year and assisted me in numerous ways. Thank you very much to Dr. Hao Xu for his numerous stimulating discussions, for most of the proofreading, and for being my friend. Thanks to Dr. Yanli Wang for help with the structure refinement of the human p100 tudor domain. My thanks also go to Dr. Irina Kataeva, Dr. Min Zhang, Dr. Jie Yang and members of Dr. Adams’ Lab, who provided clones used in this research. I heartily appreciate the help of Dr. Gary Newton for his valued criticism and proofreading of the first pages of the dissertation. I am much indebted to all my labmates past and present for providing a stimulating and fun environment. Special thanks go to my friends, Mr. Jinyi Zhu for his software support and Mr. Dayong Zhou for his discussions. I cannot end this without thanking my parents and my dearest wife, Jing, for their caring, understanding and love.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td></td>
<td>LIST OF ABBREVIATIONS</td>
<td>xiii</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Specific aims and significance</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Brief history of the traditional protein crystallography and structural genomics</td>
<td>3</td>
</tr>
<tr>
<td>1.3</td>
<td>Development of high-throughput pipeline and salvaging procedures</td>
<td>6</td>
</tr>
<tr>
<td>1.4</td>
<td>Three tiers of the UGA high-throughput pipeline</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>Materials and methods</td>
<td>48</td>
</tr>
<tr>
<td>2.1</td>
<td>Surface mutagenesis and gene cloning</td>
<td>48</td>
</tr>
<tr>
<td>2.2</td>
<td>Protein expression and solubility test</td>
<td>50</td>
</tr>
<tr>
<td>2.3</td>
<td>Protein purification</td>
<td>51</td>
</tr>
<tr>
<td>2.4</td>
<td>Reductive methylation of surface lysine residues</td>
<td>54</td>
</tr>
<tr>
<td>2.5</td>
<td>Crystallization</td>
<td>54</td>
</tr>
<tr>
<td>2.6</td>
<td>Data collection and processing</td>
<td>55</td>
</tr>
<tr>
<td>2.7</td>
<td>Structure determination</td>
<td>56</td>
</tr>
</tbody>
</table>
2.8 Function prediction of hypothetical proteins..............................................57

3 Salvaging/rescue trials of 34 Pyrococcus furiosus proteins that failed at first trials ..........................................................................................................................62
  3.1 Salvaging/rescue targets ...........................................................................63
  3.2 Methods employed in the salvaging/rescue trials ..................................63
  3.3 Results of salvaging trials .......................................................................64
  3.4 Function prediction of PF0907 and PF0725 ............................................66

4 Treating these salvaging/rescue procedures as part of the UGA high-throughput pipeline ..................................................................................................................81
  4.1 Application of the improved high-throughput pipeline to 37 new protein targets .........................................................................................................................81
  4.2 Structure determination of AF0160 through the high-throughput pipeline .................................................................................................................................84
  4.3 Function prediction of five solved structures ...........................................91

5 Structure determination of the tudor domain of the human p100 protein, a non-structural genomics project using the improved high-throughput pipeline .........................................................................................................................122
  5.1 Different methods used to purify or modify the human p100 TSN region ........................................................................................................................................123
  5.2 Crystallization .............................................................................................127
  5.3 Data collection and processing ..................................................................128
  5.4 Structure solution of the human p100 tudor domain ...............................129
5.5 The structure of the human p100 tudor domain and its putative function

6 Conclusions

REFERENCES
LIST OF TABLES

Table 1.1a: Some commonly used tags .................................................................................. 44
Table 1.1b: Some commonly used proteases and their cleavage sites .............................. 44
Table 2.1: Destination vectors used in the genetic trunction of the human p100 TSN region ............................................................................................................................................. 59
Table 3.1a: Salvaging efforts for 20 *Pyrococcus furiosus* proteins failed in the native form .......................................................................................................................................... 75
Table 3.1b: Salvaging efforts for 14 *Pyrococcus furiosus* proteins failed in the Se-Met form .......................................................................................................................................... 76
Table 3.2: Statistics from the crystallographic analysis of PF0907 ..................................... 77
Table 3.3: Results of DaliLite Pairwise comparison of the structures of PF0907 and its homologues ..................................................................................................................................... 78
Table 3.4: Statistics from the crystallographic analysis of PF0725 ..................................... 79
Table 3.5: Results of structure similarity screening of PF0725 using the DALI server ...... 80
Table 4.1: Production of 5 structures from 35 new protein targets by the improved high-throughput pipeline ...................................................................................................................................... 113
Table 4.2: Statistics from the crystallographic analysis of AF0160 ................................. 114
Table 4.3: Common results of structure similarity screening of AF0160 using the DALI and VAST server ........................................................................................................................................ 115
Table 4.4: Statistics from the crystallographic analysis of Cth393 ................................. 116
Table 4.5: Results of structure similarity screening of Cth393 using the DALI server .... 117
Table 4.6: Statistics from the crystallographic analysis of Cth2968................................. 118
Table 4.7: Results of DaliLite Pairwise comparison of the structures of Cth2968 and its homologues.......................................................................................................................... 119
Table 4.8: Statistics from the crystallographic analysis of Cth95........................................ 120
Table 4.9: Statistics from the crystallographic analysis of Cth383........................................ 121
Table 5.1: Statistics from the crystallographic analysis of human p100 tudor domain .. 134
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Some concepts of X-ray crystallography</td>
<td>45</td>
</tr>
<tr>
<td>1.2</td>
<td>A schematic diagram of ISAS/ISIR</td>
<td>46</td>
</tr>
<tr>
<td>1.3</td>
<td>Flowchart of the UGA high-throughput pipeline</td>
<td>47</td>
</tr>
<tr>
<td>2.1</td>
<td>Example of the two-step purification using ÄKTApurifier 3D for automatic</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>purification of three proteins simultaneously</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>SDS-PAGE of PF1835, PF0147 and PF1098, purified by the process</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>described in Figure 2.1</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>SDS-PAGE results of PF1183 and PF0385 after each purification step</td>
<td>71</td>
</tr>
<tr>
<td>3.2</td>
<td>The ribbon representation of PF0907 (A) and the structure-based alignment of</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>PF0907 and its homologues (B)</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of the aa-tRNA binding region of the EF-Tu from Thermus Aquaticus</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>(PDB ID, 1TTT) and the corresponding region of PF0907</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>The structure of PF0725 with CoA</td>
<td>74</td>
</tr>
<tr>
<td>4.1</td>
<td>Distribution of the amounts of purified proteins</td>
<td>101</td>
</tr>
<tr>
<td>4.2</td>
<td>Purification of AF0160</td>
<td>102</td>
</tr>
<tr>
<td>4.3</td>
<td>Crystallization of AF0160</td>
<td>103</td>
</tr>
<tr>
<td>4.4</td>
<td>SCA2Structure pipeline input and output webpage</td>
<td>104</td>
</tr>
<tr>
<td>4.5</td>
<td>Three-fold NCS of AF0160 molecules in the asymmetric unit</td>
<td>105</td>
</tr>
<tr>
<td>4.6</td>
<td>Identification of the position of a Se atom at a loop region</td>
<td>106</td>
</tr>
</tbody>
</table>
Figure 4.7: The structure of AF0160 ................................................................. 107
Figure 4.8: AF0160 is structurally homologous to the TorD-like proteins .......... 108
Figure 4.9: Cth393 is a member of the HIT protein family .............................. 109
Figure 4.10: Cth2968 is a YjgF/YER057c/UK114 family protein ............... 110
Figure 4.11: Cth95 is similar to EF-P ............................................................ 111
Figure 4.12: Cth383 is an YbaB family protein ............................................ 112
Figure 5.1: SDS-PAGE of the native (A) and Se-Met substituted (B) human p100 TSN region (p100td) after purification steps ............................................. 135
Figure 5.2: Reductive methylation of the native human p100 TSN region (p100td)...... 136
Figure 5.3: Surface mutation of the human p100 TSN region (p100td) ........... 137
Figure 5.4: Solubility testing of truncated regions (td1-td6) of the human p100 TSN region ........................................................................................................ 138
Figure 5.5: SDS-PAGE of p100td2-544, 566 and 598 after Tev protease digestion and Ni affinity purification ......................................................... 139
Figure 5.6: Genetic truncation based on the information of limited proteolysis ....... 140
Figure 5.7: Crytallization of the human p100 tudor domain ............................ 141
Figure 5.8: Identification of the degraded human p100 TSN region .................... 142
Figure 5.9: The structure of the human p100 tudor domain ............................ 143
Figure 5.10: The tudor domain of the human p100 protein is similar to that of SMN ... 144
Figure 5.11: The hydrophobic pockets (green) from tudor domains of the human SMN, 53BP1, JMJD2A and p100 protein, and their ligands (raspberry) ............ 145
Figure 5.12: Conserved residues of the tudor domains of the human p100 protein and its homologues ................................................................................... 146
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>three-dimensional</td>
</tr>
<tr>
<td>ANL</td>
<td>Argonne National Laboratory</td>
</tr>
<tr>
<td>APS</td>
<td>the Advanced Photon Source</td>
</tr>
<tr>
<td>β-ME</td>
<td>β-mercaptoethanol</td>
</tr>
<tr>
<td>CCD</td>
<td>charge-coupled device</td>
</tr>
<tr>
<td>CGS</td>
<td>cystathionine γ-synthase</td>
</tr>
<tr>
<td>CoA</td>
<td>Coenzyme A</td>
</tr>
<tr>
<td>DALI</td>
<td>Distance Matrix Alignment</td>
</tr>
<tr>
<td>DMAB</td>
<td>dimethylamine-borane complex</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EF-P</td>
<td>translation elongation factor P</td>
</tr>
<tr>
<td>EntK</td>
<td>Enterokinase</td>
</tr>
<tr>
<td>FOM</td>
<td>figure of merit</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>GST</td>
<td>glutathione-S-transferase</td>
</tr>
<tr>
<td>GXRCC</td>
<td>the Georgia X-Ray Crystallography Laboratory</td>
</tr>
<tr>
<td>HAT</td>
<td>histone acetyltransferase</td>
</tr>
<tr>
<td>HCA</td>
<td>hydrophobic cluster analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>HIT</td>
<td>histidine triad</td>
</tr>
<tr>
<td>HMM</td>
<td>the Hidden Markov model</td>
</tr>
<tr>
<td>HTH</td>
<td>helix-turn-helix</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl-beta-D-thiogalactopyranoside</td>
</tr>
<tr>
<td>ISAS/ISIR</td>
<td>iterative single anomalous scattering/iterative single isomorphous replacement</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani broth</td>
</tr>
<tr>
<td>MAD</td>
<td>multiple anomalous dispersion</td>
</tr>
<tr>
<td>MIR</td>
<td>multiple isomorphous replacements</td>
</tr>
<tr>
<td>MWCO</td>
<td>molecular weight cut-off</td>
</tr>
<tr>
<td>NCS</td>
<td>non-crystallographic symmetry</td>
</tr>
<tr>
<td>NIGMS</td>
<td>National Institute of General Medical Sciences</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institute of Health</td>
</tr>
<tr>
<td>OD$_{600}$</td>
<td>optical density under 600 nm wavelength</td>
</tr>
<tr>
<td>PA0.5G</td>
<td>P- 100mM phosphate; A- 200 µg/ml of each of 18 different amino acids (no Cys, Tyr); 0.5G- 0.5% glucose.</td>
</tr>
<tr>
<td>PASM5052</td>
<td>P- 100mM phosphate; A- 200 µg/ml of each of 18 different amino acids (no Cys, Tyr); SM- selenomethionine; 5052- 0.5% glycerol, 0.05% glucose, and 0.2% lactose.</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate buffered saline buffer</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PLP</td>
<td>pyridoxal 5’-phosphate</td>
</tr>
<tr>
<td>PMSF</td>
<td>phenylmethylsulphonyl fluoride</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PSI</td>
<td>Protein Structure Initiative</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>SAS</td>
<td>single anomalous scattering</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification of Proteins</td>
</tr>
<tr>
<td>SCS</td>
<td>succinyl-CoA synthetase</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SECGSG</td>
<td>Southeast Collaboratory for Structural Genomics</td>
</tr>
<tr>
<td>SeMet</td>
<td>selenomethionine</td>
</tr>
<tr>
<td>SER-CAT</td>
<td>the Southeast Regional Collaborative Access Team</td>
</tr>
<tr>
<td>SIR</td>
<td>single isomorphous replacement</td>
</tr>
<tr>
<td>SIRAS</td>
<td>single isomorphous replacement anomalous scattering</td>
</tr>
<tr>
<td>SMN</td>
<td>Survival of Motor Neuron</td>
</tr>
<tr>
<td>SN</td>
<td><em>staphylococcal</em> nuclease</td>
</tr>
<tr>
<td>SSM</td>
<td>Second Structure Matching</td>
</tr>
<tr>
<td>Tev</td>
<td>Tobacco etch virus protease</td>
</tr>
<tr>
<td>Thr</td>
<td>Thrombin</td>
</tr>
<tr>
<td>TLS</td>
<td>translation/libration/screw</td>
</tr>
<tr>
<td>TSN</td>
<td>tudor and SN region</td>
</tr>
<tr>
<td>UGA</td>
<td>the University of Georgia</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>VAST</td>
<td>Vector Alignment Search Tool</td>
</tr>
</tbody>
</table>
CHAPTER ONE

Introduction

The diverse biological functions of proteins in living cells depend on their three-dimensional (3D) structure. Determining protein structures will be helpful for the understanding of their functions, and will have a significant impact on life science, biotechnology, and drug discovery. For the drug discovery process, about 50% of the cost could be saved using structure-based methods (Stevens, 2004). At the present time, X-ray crystallography, which determines structure by studying the X-ray diffraction of a crystalline sample, is the main method to obtain atomic-resolution structure information of proteins.

1.1 Specific aims and significance

The short-term aim of the present study is to assist in the development and optimization of salvaging/rescue procedures, and to integrate them into the general high-throughput gene-to-structure pipeline to enhance the capability of the pipeline. The long-term goal is to apply the principles of the high-throughput method to study structures of interesting proteins. To achieve these goals, three specific aims were designed. These specific aims are: 1) to test the salvaging/rescue procedures with 34 proteins that were not successful by the general method; 2) to treat these salvaging/rescue procedures as part of the UGA high-throughput pipeline (Tier 3) and apply this improved pipeline to 37 new protein targets; and 3) to use the improved pipeline for the structural study of the human p100 tudor domain, a non-structural genomics target.
To determine protein structures by X-ray crystallography, diffraction-quality crystals of proteins should be produced. The reason that certain proteins failed in yielding diffraction-quality crystals include: the protein is not pure enough; the protein adopts variable conformations; and specific crystallization conditions need to be found. Therefore, additional purification methods to improve the protein purity, protein modification methods to stabilize certain conformations, and alternative crystallization set-up methods may improve the chance to obtain diffraction-quality crystals.

To address these hypotheses, various methods including additional ion-exchange chromatography, reductive methylation, surface mutagenesis, selenomethionine (SeMet) substitution, and modified microbatch method for crystallization trials were used to salvage problem targets, or incorporated into the high-throughput pipeline (improved high-throughput pipeline) for new targets. A set of 34 *Pyrococcus furiosus* proteins that failed in the initial trials was selected as salvaging/rescue targets. A set of 37 new targets, including 31 *Clostridium thermocellum* targets and six *Archaeoglobus fulgidus* targets, were selected for improve high-throughput pipeline.

These methods improved the success rate of our high-throughput pipeline. As a direct result, this research is involved in the production of 10 structures, most of which belong to hypothetical proteins of unknown function. These 3D structures provide clues to help in the understanding of their functions. These structures could also be used for homology modeling and evolution study, and to reveal remote relationships with other proteins.

In addition, we also have a chance to evaluate these methods and deeply understand the high-throughput pipeline. When those “low-hanging fruit” targets are
solved, these methods will become vital for those “high-hanging” targets. High-throughput pipeline advancement will not only increase the structure output of structural genomics but also benefit the entire biological science community.

1.2 Brief history of the traditional protein crystallography and structural genomics

1.2.1 Traditional protein crystallography

It has been over 160 years since the first crystals of a protein, hemoglobin, were grown (McPherson, 1999, Hünefeld, 1840). Over 90 years ago, the diffraction experiment of a crystal of CuSO₄·5H₂O (Friedrich, 1912) helped to demonstrate that X-rays could be used as a tool for crystal structure determination. The first diffraction pattern of a protein crystal was recorded over 70 years ago (Bernal, 1934). More than twenty years later, Max Perutz in Sir W.L. Bragg’s laboratory developed the isomorphous replacement method to solve the phase problem for protein structures which led to the successful determination of the first protein crystal structure of hemoglobin in 1960 (Perutz, 1960). Initially, protein targets were selected by the availability of protein and not necessarily their biological significance. Later, with the development of improved protein crystallographic methods, people could study important proteins or complexes of one system at a time. As a result, the structures of many important proteins have been solved. (Blundell & Johnson, 1976, Dauter, 2006)

1.2.2 Structural genomics

Various genome-sequencing projects have provided new information for many proteins, including many hypothetical proteins of unknown function. To understand more about their biological functions and molecular interactions, the information of their 3D structures are required. An international effort to determine 3D macromolecular
structures in genome-scale is underway, with a primary focus on proteins. These studies are termed structural genomics.

The goal of structure genomics is “the understanding of protein structural families, structural folds, and the relationship of structure and function” (RFA-GM-99-009, NIGMS). By determining the representative protein structures for every protein family, structure genomics will fill in protein fold space and provide a comprehensive view of the protein structure universe. Domains that have a clear common evolutionary origin will be classified into the same family. Families with common structural features or functions that imply a common evolutionary origin are grouped into superfamilies. For example, a fold could have one or more superfamilies for which there is little or no evidence to suggest a common evolutionary origin, but share similar secondary structural features and topology. According to Govindarajan (Govindarajan et al., 1999), approximately 4,000 possible folds exist. However, by November 1, 1999--just before the United States established its structural genomics pilot centers at the beginning of 2000--only 520 folds had known structures in the context of the Structural Classification of Proteins (SCOP) database. (Todd et al., 2005)

The structures solved by structural genomics can be used as models for the homology modeling of closely related proteins. Currently, a reasonably accurate homology model can be obtained if it shares a minimum of 30% sequence identity and 80% overlap with the template structure, considering the impacts of insertion and deletion.

The conclusions of structure genomics will also aid in the study of protein function, the design of new medicines by identifying novel drug targets and advance our
understanding of protein and organism evolution. Overall, structure genomics will likely
greatly affect life science, biotechnology, and medicine. (Stevens et al., 2001, Blundell &
Mizuguchi, 2000, Skolnick et al., 2000, Mittl & Grutter, 2001)

Since the late 1990s--after several years of planning, international meetings and
small-scale studies--structural genomics initiatives have sprung up all over the world,
including the United States, Canada, Brazil, Japan, China, Korea, Israel, France,
Germany, the United Kingdom, Sweden, Italy, Australia, and Switzerland. In the United
States, the Protein Structure Initiative (PSI) funded mainly by the National Institute of
General Medical Sciences (NIGMS), the National Institute of Health (NIH), is the largest
structural genomics initiative in the world and the only project that aims to provide a
large structure database for the entire community (http://www.nigms.nih.gov/psi/). In
September 2000, seven pilot structural genomics consortia were awarded five-year
research grants. Two new centers were supported in 2001. The Southeast Collaboratory
for Structural Genomics (SECSG) was one of the original seven pilot centers. It involves
five partner institutions: the University of Georgia (UGA), the University of Alabama at
Birmingham, the University of Alabama at Huntsville, Georgia State University, and
Duke University Medical Center (Adams et al., 2003).

By February 1, 2005, toward the end of the pilot phase of PSI, PSI centers had
solved 1,032 protein structures. These structures account for about two-thirds of all
structures contributed by structure genomics centers worldwide. Almost 60% (597) of
PSI structures were novel structures (share no more than 30% sequence identity with any
known structures) (Todd et al., 2005). About 20% of PSI structures represented new
structurally characterized families (according to Pfam, a manually curated database of
protein families from sequenced genomes), whereas non-structural genomics structures amounted to only 5% (Chandonia & Brenner, 2006). During that period, PSI centers also contributed 74 new folds or superfamilies, according to the SCOP database. In addition, most of those proteins, which structures were solved by PSI centers, belong to superfamilies that can be found in all three kingdoms of life. They may be essential for life. For the remaining proteins, many are bacteria-specific and thus might be important as drug targets. (Todd et al., 2005)

1.3 Development of high-throughput pipeline and salvaging procedures

The high productivity of PSI centers depends on the high-throughput gene-to-protein-to-structure pipelines. Developed by each PSI center, the high-throughput pipelines share some common features. They usually consist of six steps, 1) target selection, 2) gene cloning and expression, 3) protein purification, 4) crystallization and optimization, 5) data collection and processing and 6) crystal structure determination. The application of robotics, automation, miniaturization, and parallelization for each step in the process can significantly improve the rate of protein structure determination by the high-throughput pipelines.

However, the average success rate of PSI centers from purified proteins to crystal structures was only 7% as of August 2004 (Liu, Shah et al., 2005). The two major bottlenecks of the high-throughput pipeline are the solubility of expressed proteins and the production of diffraction-quality crystals. Therefore, salvaging/rescue procedures should be applied for the failed targets. These procedures are crystal-directed, which means proteins are purified or modified to produce diffraction quality crystals.
The steps of the high-throughput pipelines and salvaging/rescue procedures of structure genomics are discussed in detail below.

1.3.1 Target Selection strategy

The first crucial step in any structural genomics project is target selection--principally a computational process of choosing and prioritizing proteins for experimental structure characterization. Since a large number of proteins have been identified and many protein structures can be modeled by homology modeling methods, structural genomics consortia cannot and need not determine structures of all proteins by experimental methods. In order to construct reasonable models for all proteins, 16,000 carefully selected structures might be sufficient. Otherwise, it would require up to seven times as many structures to cover 90% protein fold representation if targets were selected randomly (Stevens et al., 2001). In addition, if the targets had a high probability of failure, most likely it will be a waste of time, expense and effort to study them. Therefore, carefully and strategically selected targets would lower the cost of studies and improve the chances of obtaining valuable structures.

The two factors that affect the strategy of target selection are the goal of structural genomics and the limitation of high-throughput pipelines.

As described above, the goal of structure genomics is to cover structure space. One can select new-fold targets and representatives of each known family of proteins based on sequence analyses. For the accurate homology modeling of all members of a diverse protein family, multiple structures may need to be solved by experimental methods (Weiner & Lu, 2002). Moreover, some structural genomics projects focus on the complete repertoire of proteins in an organism or a specific pathway, and therefore may
finally help in the long term to build cellular protein-protein interaction networks (Linial & Yona, 2000). On the other hand, some consortia indiscriminately pass an entire proteome of a model organism through their high-throughput pipelines to evaluate the individual production steps and to populate the database that will be used to develop more efficient methods (Weiner & Lu, 2002).

Another factor that affects target selection is the limitation of high-throughput pipelines. For example, membrane protein structure determination is extraordinarily challenging and currently is not suitable for the high-throughput pipelines, even though membrane proteins present a high proportion of human drug targets (Brenner, 2000, Watson et al., 2003). Proteins containing surface regions of low-complexity represent another difficulty in structure determination. These regions have comparatively little variation in residue content (e.g., polyproline stretch) and thus have a high degree of conformational flexibility, which may interfere with the crystallization of proteins (Watson et al., 2003). Another example of the limitations of a high-throughput pipeline approach concerns the *E.coli* expression system. As the most efficient expression system in different pipelines, *E.coli* expression system has a limitation on the size of recombinant target proteins, of which the ideal molecular weights are between 30 and 60 kDa. Therefore, ideal targets for high-throughput study are proteins that have appropriate sizes for corresponding expression systems, no low-complexity regions and no transmembrane domains. For thermostable target proteins, heat treatment of the crude cell homogenate could improve their recovery efficiency. (Weiner & Lu, 2002)
The target selection process can be divided into four distinct stages. These are: 1) the target identification, 2) the target exclusion, and 3) the target prioritization (Brenner, 2000).

The target identification is composed of three steps. First, organisms/genomes are selected based on the goals of structural genomics projects. Second, all proteins of these organisms/genomes are comprehensively organized into families. Once this organization is obtained, the representatives of each family can be selected as potential candidates for structural determination. There are four kinds of protein family database (Watson et al., 2003), 1) motif-based, such as PROSITE (Hulo et al., 2006), PRINTS (Attwood et al., 2003), and Blocks (Henikoff et al., 1999); 2) domain-based, such as Pfam (Bateman et al., 2002), SMART (Schultz et al., 1998), TIGRFAMS (Haft et al., 2001), ProDom (Servant et al., 2002), and SUPERFAMILY (Gough, 2002); 3) gene-based, such as SWISS-PROT (Bairoch & Apweiler, 2000), COG (Tatusov et al., 2001) and SYSTERS (Krause et al., 2000); and 4) others, such as GO (Ashburner et al., 2000), and KEGG (Kanehisa et al., 2002). Several sensitive and profile-based techniques, such as PSI-BLAST (Altschul et al., 1997) and the Hidden Markov model (HMM) method SAM-T99 (Karplus et al., 1998), are used to detect distant relationships of proteins of the same family (Watson et al., 2003).

The second stage is target exclusion. Due to the limitation of high-throughput pipelines, protein targets such as membrane proteins, proteins with low-complexity regions, proteins with large size, and proteins with post-translational modifications are removed from the target list. Some consortia keep some of those targets that are important for their studies, but remove the regions that are difficult to crystallize or select.
their homologs from other organisms. Additionally, protein targets are removed from the list if their structures are known or can be computationally modeled from homologous structures.

After target exclusion, the prioritization step focuses on the most valuable targets, depending on the goals of the structure genomics projects. General criteria include family size, taxonomic diversity and functional importance of targets.

A prioritized target list is produced at the end of the target selection process. As technology advances and structural information accumulates, target selection criteria must constantly be refined and reassessed, depending on project goals.

At the beginning of SECSG, three genomes were selected: *Pyrococcus furiosus* (*P. furiosus*)--a model prokaryote, *Caenorhabditis elegans* (*C. elegans*)--a model eukaryote, and selected proteins for the human (*Homo sapiens*) genome. *P. furiosus* is a slowly evolving archaeon that has a relatively small and well-defined genome of approximately 2,200 ORFs. *C.elegans* was predicted to have at least 19,000 ORFs. The goal of SECSG was to study all of these ORFs of *P. furiosus* and *C.elegans* (Adams et al., 2003). Finally, 2,182 *P. furiosus* targets and 14,442 *C.elegans* targets were selected (Wang et al., 2005). For human proteins, 446 targets that had less than three predicted transmembrane domains and shared less than 30% sequence identity to any Protein Data Bank (PDB) entry were finally selected, based on interest. Later, 328 *Aeropyrum pernix* targets, 223 *Clostridium thermocellum* targets, 268 *Archaeoglobus fulgidus* targets, 205 *Pyrococcus horikoshii* targets, and 261 *Thermus thermophilus* targets were included in the analysis.
1.3.2 Gene Cloning and Expression

After target selection, selected genes are taken from their original organisms and put into appropriate vector-host systems for protein expression. Resources for genes include genomic DNA, cDNA libraries, and pre-existing sequence-validated full-length cDNA clones. Genomic DNA is a good resource if the organism has little or no splicing (e.g., prokaryotes and some simple eukaryotes). Otherwise, cDNA resources are preferred. (Marsischky & LaBaer, 2004)

Genes of interest are obtained by PCR and constructed into appropriate expression vectors. Plasmids, identified as extra-chromosomal self-replicating cytoplasmic DNA elements, are the most commonly used expression vectors. To be replicated in host cells during cell division, plasmids should have appropriate replication origins. Other key features of expression plasmids usually include multiple cloning sites, promoter, ribosome binding site, transcriptional terminator, selectable marker, and tags. (Lodge et al., 2007)

Tags are proteins or short stretches of amino acids (peptides) that are fused to the ends of proteins of interest. First, tags such as His\textsubscript{6} tag or glutathione-S-transferase (GST) tag can be used for affinity purification. Second, some tags can improve expression levels of target proteins. Third, some highly soluble tags (e.g., MBP or Sumo) can be used to improve the solubility of target proteins. Finally, some tags, such as green fluorescent protein (GFP), can be used to monitor protein folding in solution. Some commonly used tags are shown in Table 1.1a. However, tags may alter the properties of target proteins. To remove tags after purification, the most common method is to incorporate a protease
cleavage site between the tag and the target protein. Some commonly used proteases and their cleavage sites are shown in Table 1.1b. (Esposito & Chatterjee, 2006, Stevens, 2000b)

To clone target genes into expression vectors, cloning methods include restriction enzyme-based cloning, homologous recombination-based cloning, and site-specific recombination-based cloning. For restriction enzyme-based cloning, target genes that contain corresponding restriction sites at both ends (added by PCR) and expression vectors that have the same corresponding restriction sites only at multiple cloning sites are cut by restriction enzymes and ligated together by DNA ligase. For homologous recombination-based cloning, target genes are flanked with recombination regions by PCR and inserted into the vector by homologous recombination between the homologous regions of the vector and the gene fragment. For site-specific recombination-based cloning, modified versions of site-specific recombination systems of bacteriophages are used to transfer target genes to master clones and different expression clones. The site-specific recombination sequences are embedded in the PCR products of target genes and vectors to be recombined.

To choose a cloning system for high-throughput applications, several parameters should be considered: fidelity of the cloning process, ease of use, reliability of the cloning system, validation of the cloned products, flexibility of use of the cloned products, lack of undesirable properties on the clones, and costs. Some structural genomics consortia are using site-specific recombinational cloning systems, such as the Gateway system (Invitrogen), which exceed requirements but are comparatively expensive. (Marsischky & LaBaer, 2004)
High-throughput pipelines use three types of expression hosts/systems: the prokaryotic host, the eukaryotic host, and the cell-free expression system. Bacterial cells such as *E.coli* have been widely used as a prokaryotic host because they represent the cheapest, quickest and easiest means of heterologous protein expression. However, to form the final active products, some target proteins from higher-order eukaryotes need substantial posttranslational modifications, including proteolytic cleavage, glycosylation, and amino acid modifications. Most of these modifications cannot be performed in bacterial cells, but are available in appropriate hosts of eukaryotic systems, including mammalian, insect, and yeast cells (Balbâas & Lorence, 2004). The cell-free expression system that has coupled *in vitro* transcription-translation reactions has been successfully applied to produce proteins prone to aggregation or toxicity *in vivo* (Yokoyama *et al.*, 2000).

In the high-throughput pipelines, the process of the cloning and expression of target proteins can be divided into five steps: 1) PCR to obtain the target genes, 2) cloning the target genes into the appropriate vectors, 3) transforming the vectors with target genes into the appropriate host cells, 4) validating the cloned products, and 5) small-scale testing of the expression and solubility of target proteins. PCR primers can be designed in a 96-well format. Additional sequences should be incorporated into primers based on the cloning techniques to be used next. For example, the recombination site should be attached for the site-specific recombination method. The PCR products are confirmed by agarose gel analysis. However, additional screening steps are required to confirm the proper insert direction. Methods of transformation include chemical transformation, electroporation, particle bombardment, and lipofection. Chemical
transformation and electroporation are usually used for *E.Coli* as host cells. DNA sequencing is one method of cloned product validation. For expression and solubility tests, cells are grown in 3-5ml media and induced for expression as required. Cells are lysed by chemicals, sonication, or other mechanical forces. The soluble and insoluble components are separated by centrifuge and analyzed by SDS-PAGE to determine the expression level and solubility of target proteins. The samples can be analyzed further by ELISA assay with antibodies of specific tags. (Weiner & Lu, 2002)

Taking the cloning technology in SECSG as an example, the Gateway cloning system (Invitrogen) is used for many target genes. The T7 or Trc (Tac) promoter has been chosen for over-expression of target proteins in *E.coli*. Cells are usually grown in 1L medium at 37ºC. Target proteins are expressed by IPTG induction or autoinduction. His$_6$ tags are fused with target proteins for the affinity purification in the next step.

At this step of high-throughput pipeline, two common problems are the insolubility and no or low expression of target proteins. Several methods can be used as salvaging/rescue procedures. A variety of tags, such as MBP and NusA, can improve the solubility of target proteins. Different expression conditions (e.g., temperature, inducer concentration, and media) may improve the solubility and expression of target proteins. Alternative expression host/system and promoters can also be used to salvage target proteins. (Wang *et al.*, 2005)

**1.3.3 Protein Purification**

After testing the target gene’s expression, the next step is to scale up protein expression for the purification step. Although several methods can be used to secrete proteins out of the cell, in most cases target proteins have to be expressed inside the host
cells and released in soluble form by cell lysis. Cell lysis methods include mechanical or non-mechanical lysis. Non-mechanical lysis disrupts cells using physical forces, including osmotic shock, freezing and thawing, or such chemicals as alkali, enzyme, or detergent. Mechanical lysis uses sonication, French press, a homogenizer, or a bead beater to disrupt cells. However, one should control the heat produced by mechanical lysis and avoid foaming, surface denaturation, and oxidation. (Ahmed, 2005)

Such factors as pH, proteolysis, reducing agents, and contamination should be considered in making the choice of lysis and purification solutions.

Since small changes in the pH may severely affect the properties of proteins, the protein solution should be maintained at a constant pH to ensure reproducible experimental results. A buffer solution is the ideal solution for protein purification. Buffer salts should be at the lowest reasonable concentration to avoid “non-specific” ionic strength effects. Several other factors -- the aim of the protein purification, the pKa of the protein, the effects of temperature, interactions with other components, compatibility with different purification and analyses methods, and cost -- should be considered when choosing a buffer.

The proteolysis of target proteins can be a serious problem during protein purification. Proteolysis may generate degraded proteins, affect the stability of the target protein, and completely or partially inactivate the target protein. The simplest way to avoid the protein proteolysis is to work fast at a low temperature. Additionally, various protease inhibitors can inhibit different classes of protease in cells. An inhibitor cocktail may be used for new protein extracts. In some cases, the buffer pH can be adjusted to a value at which the proteases are inactive while the target proteins are still stable.
Bacterial intracellular proteins are produced in a reducing environment where the redox potential is lower than that of the purification buffer. To protect the exposed thiol groups from oxidation, reducing agents are usually added to the protein buffer.

The simplest way to prevent bacterial or fungal growth in the protein solutions and columns is to sterile-filter the buffer solutions. Other methods include storing the buffer solutions at 4°C, adding antimicrobial agents such as sodium azide to the buffer solutions, and--if appropriate--using buffers with the pH lower than 3 or higher than 9, which usually prevent bacterial growth.

Contamination with nucleic acids can cause high viscosity in solutions. Nucleic acids can be precipitated by streptomycin sulfate, protamine sulfate, manganese salts, and polyethylenimine. Alternatively, they can be digested by nucleases. (Ahmed, 2005, Janson & Rydåen, 1998)

Following cell lysis, several extraction and precipitation methods can be used as preliminary steps of purification to obtain clarified or more concentrated protein solutions. It is possible to precipitate the molecular contaminants by adding salts, organic solvents, or polymers, or by varying the temperature or pH of the solutions. Protein solutions will then be clarified by centrifugation or membrane filtration before further purification.

Liquid chromatographic techniques are the main methods used for protein purification. Similar to other purification methods, these techniques separate target proteins based on their specific properties (e.g., molecular weight, charge, and hydrophobicity). Most commonly used are gel filtration (size exclusion) chromatography (based on protein size and shape), ion-exchange chromatography and hydroxyapatite
chromatography (based on net charge and distribution of charged groups), chromatofocusing (based on isoelectric point), hydrophobic interaction chromatography and reversed-phase chromatography (based on hydrophobicity), covalent chromatography (based on content of exposed thiol groups), and affinity chromatography (based on biospecific affinities for metal ions, ligands, inhibitors, receptors, antibodies, etc.).

In gel filtration (size exclusion) chromatography, small molecules can enter the pores of gel particles and move slowly through the column. In contrast, large molecules cannot enter the gel pores and travel fast through the interstitial spaces of the gel particles. By this mechanism, gel filtration chromatography separates proteins of different sizes and shapes. However, gel filtration is theoretically not suitable for separation of proteins from crude extracts because of column size and resolution. Instead, gel filtration is usually performed as a polishing step at the end of a purification procedure. To prevent hydrophobic interactions between the packing material and the protein molecules, 50 to 100 mM sodium chloride in the solution provides a suitable ionic strength. For the best resolution, the volume of the loading sample should be less than 3% of the total column volume.

Ion-exchange chromatography can be used to separate proteins with different net charges. Target proteins interact with opposite charges in the stationary phase, and are eluted or displaced by counter ions with greater affinities for the charges of the stationary phase. The charges of the stationary phase are fixed charges, whereas the charges of the counter ions used to displace proteins are displaceable charges. Ion-exchange chromatography can be classified into two types: anion-exchange chromatography with negative displaceable charges, and cation-exchange chromatography with positive...
displaceable charges. The buffer ions should have similar charges with the stationary phase. The two ways to elute target proteins are step elution and gradient elution. Step elution increases the ionic strength of the elution buffer stepwise, while gradient elution increases it gradually. Step elution is simpler and more rapid, whereas gradient elution is more efficient for the separation of proteins with close pIs.

In affinity chromatography, target proteins specifically interact with ligands that are conjugated to the stationary phase. Similar to ion-exchange chromatography, in affinity chromatography the target proteins bind the immobilized ligands and are displaced by adding free ligands, varying the pH, or increasing the ionic strength. The metal-binding proteins can bind transition metal ions by the electron donor groups of some amino acids of the proteins (e.g., histidine and tryptophan). Like other ligands of affinity chromatography, the metal ions are immobilized onto the stationary phase via the chelating groups. Due to the high specificity of the binding, affinity chromatography is one of the most powerful methods of protein purification.

As a calcium phosphate crystal, hydroxyapatite comprises a mosaic of positive and negative sites. The hydroxyapatite columns can be used to bind basic or acidic proteins under different conditions. For the binding of basic proteins, the column is equilibrated with the phosphate buffer. NaCl, MgCl₂, or CaCl₂ solutions are used to equilibrate the column for acidic protein binding. (Ahmed, 2005, Janson & Rydåen, 1998)

With the proper adjustment of such parameters as ionic strength, pH, and various additives, chromatographic conditions can be optimized. Normally, the complete
purification of a protein from a crude mixture requires the combination of several different chromatographic methods.

In addition to chromatographic techniques, some other protein purification techniques may be used. Ultrafiltration is a gentle, fast, and comparatively inexpensive method that uses a membrane with a specific molecular weight cut-off (MWCO). Protein solutions are forced through this kind of membrane by centrifugal force or nitrogen pressure. Proteins and other small molecules with molecular weight lower than the MWCO pass through the pores of the membrane into the filtrate fraction, whereas those with higher molecular weight are retained in the retentate fraction. This method can be used to concentrate or desalt the target proteins. Other purification methods include preparative electrophoresis, preparative isoelectric focusing, etc.

The protein analysis is essential to monitor the progress of protein purification and characterize the final product. The most accurate method of protein analysis is total amino acid analysis after hydrolysis; however, it is time-consuming. Ultraviolet (UV) absorption is generally used to measure protein content. In addition, electrophoretic analyses such as SDS-PAGE are used to monitor purification, check the purity of the target protein, verify protein concentration, detect proteolysis, and detect protein modification. Other methods of protein estimation include colorimetric and fluorescent methods. Mass spectral analysis can be used to check protein degradation. Protein stability may be checked by light scattering, circular dichroism, and NMR line width analysis. (Janson & Rydäen, 1998, Ahmed, 2005, Weiner & Lu, 2002)

In the high-throughput pipelines, heat treatment is usually used at the initial stage for the purification of hyperthermophilic proteins. Buffers that have temperature-
dependent pH (e.g., the Tris buffer) should not be used in this step. (Goulding & Perry, 2003) Metal chelating affinity chromatography is usually used at the first step. The His$_6$ tags that are fused with the target proteins can bind immobilized nickel ions. In addition, running the affinity step first can improve the accuracy of peak finding. (Braun & LaBaer, 2003) The purity of the target proteins after affinity purification is typically in the range of 80% to 95% (Lesley & Wilson, 2005). Prior to the second chromatography step, the affinity tags are removed if necessary. Gel filtration chromatography can purify homogeneous target proteins in a common buffer. For structure studies, the target proteins should be concentrated to more than 10mg/ml. Mass spectroscopy and SDS-PAGE are used to confirm molecular weight and the presence of contaminating, oligomeric, or degraded protein. (Goulding & Perry, 2003) The final target protein solutions are stored at -70°C or below to minimize the risk of inactivation or denaturation.

However, many target proteins are insoluble or do not produce crystals of reasonable diffraction quality. The salvaging/rescue procedures at this level include affinity chromatography of different tags, ion-exchange chromatography, hydrophobic chromatography, and hydroxyapatite chromatography.

1.3.4 Protein modification for target salvaging

Proteins that do not crystallize usually contain flexible regions that may lead to multiple conformations. (Rayment, 1997) Such protein modification methods as reductive methylation of lysine residues (Rayment, 1997), surface mutagenesis (Derewenda, 2004) and SeMet substitution (Smith & Thompson, 1998) can change the entropy of proteins; others, such as limited proteolysis and genetic truncations, can isolate
the stable domains of target proteins. These methods can be used as salvaging/rescue procedures to get proteins crystallized or to improve the diffraction qualities of protein crystals.

Large flexible side chains of hydrophilic residues have high conformational entropy under normal conditions. Therefore, surface hydrophilic residues with large flexible side chains (e.g., Lys and Glu) may prevent crystal formation. (Derewenda, 2004)

The reductive methylation of lysine residues may change the surface properties of proteins and facilitate their crystallization. This method initially forms a Schiff base between the ε-amino group of lysine residue and formaldehyde, which is then reduced to a secondary or tertiary amine. The reaction proceeds rapidly to a dimethylated product, where the hydrogen atoms of the lysine NZ are replaced by methyl groups. Evidences has shown that reductive methylation does not significantly alter the structure but can affect the crystallization properties of target proteins. However, methylated lysine residues of crystallized proteins are rarely involved in crystalline contacts. (Rayment, 1997, Rypniewski et al., 1993, Shaw, Cheng et al., 2007) Therefore, reductive methylation may function to stabilize one of several conformations of target proteins, or simply reduce the solubility. Such reducing reagents as dimethylamine-borane complex are needed to prevent the reformation of the Schiff base from lysine and formaldehyde, which will facilitates reactions with adjacent tyrosine, tryptophan, histidine, asparagines, and cysteine residues. (Rayment, 1997)

Lys and Glu residues are found predominantly on the surface of proteins, especially when two or more of these residues are found in close proximity as a cluster in
the sequence. The replacement of these residues with smaller amino acids such as Ala may facilitate the crystallization of proteins—a process called surface mutagenesis. Previous experiments indicated that crystal contacts usually occur on the mutated site, perhaps because main chain carbonyls and amides can be involved in crystal contacts without excessive entropy loss when these side chains are removed. Because the N-terminal is usually very flexible, mutations on this region may not be helpful for crystallization. The mutations of residues in helices may be less useful because their main chain carbonyls and amides are involved in intrahelical H-bonds and are not free for crystal contacts. The mutations of Lys/Glu residues to Ala can be performed by method of site-directed mutagenesis. Surface mutagenesis requires more time and effort than reductive methylation. (Derewenda, 2004)

The method of SeMet substitution is originally used to incorporate selenium (Se), an excellent phasing probe, into proteins to facilitate structure determination. Furthermore, like reductive methylation and surface mutagenesis, SeMet substitution can also be deemed as a modification method to get proteins crystallized or improve the diffraction qualities of protein crystals. The substitution of Met with SeMet may affect the solubility and stability of proteins. Both Met and SeMet have hydrophobic sidechains, whereas SeMet has a larger surface and hence larger hydrophobicity. Differences in the hydrophobicity of these two residues affect the solubility or stability of proteins. If these hydrophobic side chains are solvent exposed, the substitution will decrease protein solubility. If these hydrophobic side chains are buried, the change will improve protein stability. Since the solubility and stability of proteins have a large impact on protein crystallization, SeMet substitution may be tried if crystals of native proteins are not good
enough. The problem with using this method is that SeMet residues of some target proteins may be oxidized during purification and crystallization trials. The oxidation of SeMet residues usually does not prevent crystallization or result in the formation of cross-linked products; however, it may cause crystallographic phase determination to fail because of the heterogeneous chemical and electronic changes at the Se atoms. For the SeMet proteins that tend to be oxidized spontaneously, methods that may be employed include crystallizing in an inert atmosphere, preventing and reversing oxidation by using reducing agents and freezing crystals in liquid N\textsubscript{2} or in a gaseous N\textsubscript{2} cold stream as soon as possible. (Smith & Thompson, 1998)

Sometimes, naturally degraded proteins can be crystallized because of the removal of their flexible segments. The N-terminus of the degraded proteins can be defined by Edman sequencing, and the C-terminus be determined by accurate molecular weight measurement. Based on this information, the segments of degraded proteins that give rise to crystals can be cloned, expressed, and purified to achieve diffraction quality crystals. The limited proteolysis of target proteins can achieve results that are similar to or even better than those of natural degradation. Previously published studies revealed that the proteolytic cleavage of proteins is usually restricted to the non-conversed or disordered region, whereas proteins yielding diffraction quality crystals are typically resistant to proteolysis. (Gao et al., 2005) Under limiting conditions, only the most kinetically accessible cleavage sites of target proteins are cleaved. These sites are usually surface exposed and located in the flexible loop regions. These solvent-accessible flexible regions are usually unstructured segments that can spoil the formation of crystals and cause nonspecific aggregation. After removal of these regions, the remaining globular
domains are more stable and less conformationally variable. Therefore, these globular domains defined by combining limited proteolysis with mass spectrometry are good candidates for crystal growth. (Kriwacki, 1998) After limited proteolysis, stable segments of target proteins may be further purified for crystallization trials. Moreover, these stable segments can be further cloned, expressed and purified. (Gao et al., 2005)

Genetic truncation involves truncating proteins into desired domains by molecular genetic methods. These methods include cloning DNA sequences that encode desired protein domains into expression vectors, and expressing and purifying desired domains for further application. Re-engineering of stable domains based on the information of natural degradation or limited proteolysis is a kind of genetic truncation. Another type of genetic truncation is based on the information of sequence analysis or structure prediction. Although they cannot accurately identify the domain boundaries, the currently used protein sequence analysis/structure prediction softwares can predict the domains and their possible boundaries, especially when the 3D structures of homologue proteins are known. Based on this information, the target proteins are genetically truncated into putative stable domains for structure studies.

1.3.5 Methods of crystallization

The process of the formation of crystals from a solution is called crystallization. Protein crystals are generally grown from supersaturated protein solutions that contain proteins in excess of their solubility limits. (McPherson, 1999) However, not only may protein crystals form from supersaturated protein solutions, but also precipitates, oils, gels, or crystals of other chemical compounds. Crystallographers aim to find appropriate chemical and physical conditions, such as pH, precipitants, temperature, and method of
crystallization, for the growth of diffraction quality protein crystals. (Bergfors, 1999) It is believed that pH is the most important determinant of protein solubility. (Delucas et al., 2005) Every protein has a specific pH value for its crystallization in a specific condition. Precipitants can precipitate proteins out of solution into crystals by altering the protein-solvent or protein-protein contacts. The four kinds of precipitants generally used in crystallography are salts, organic solvents, polymers, and surfactants. The crystallization of many proteins is also sensitive to temperature. (Bergfors, 1999) Crystallization methods (e.g., vapor diffusion, microbatch, dialysis, and controlled evaporation) can affect the kinetics of crystal growth (Delucas et al., 2005).

A crystallization screen searches through the crystallization space, a multidimensional set of possible chemical components, concentrations, and physical conditions such as temperature, set-up technique, or pH. The situation is further complicated, given that the stability and the availability of the protein are limited. Currently, several screen methods (e.g., the sparse matrix screen (Jancarik, 1991), the random screen (Rupp, 2003), and the incomplete factorial screen (Carter & Carter, 1979)) are based on different theories. (Page & Stevens, 2004)

The sparse matrix screen selects the crystallization conditions from known or published successful crystallization trials. It tests many parameters with a limited amount of protein and may identify different crystal forms. Many diffraction quality crystals grow directly out of the screen without any need for optimization. Currently, the sparse matrix screen is the most widely used screen and has contributed to the increased number of solved protein structures in past years. Nevertheless, the sparse matrix screen is biased towards conditions that have produced protein crystals previously. (Jancarik, 1991, Wooh
et al., 2003) It is likely that totally novel conditions are needed for the crystallization of some protein families that have never been crystallized. The random screen, which covers the crystallization space by using each crystallization condition only once, is the most efficient method of screening novel protein targets. It is estimated that 288 random crystallization trials should be sufficient to find whether the purified protein can be crystallized. (Rupp, 2003) However, as with the sparse matrix screen, the results of the random screen cannot provide more information because the screens are not balanced. In the incomplete factorial screen, previous knowledge is still considered; however, the occurrence of factors and their pair-wise combinations during the sampling process is balanced, identifying the most important factors and interactions that facilitate crystal formation from an initial set of screening experiments. (Carter & Carter, 1979, DeLucas et al., 2003)

These screening methods are demonstrated as equally effective. However, no screens are sufficient for the successful crystallization of all protein targets. (Page & Stevens, 2004, Hui & Edwards, 2003) Moreover, some other crystallization screens, such as an incremental adjustment of the pH and the salt concentrations, may produce protein crystals that would not appear in above the prepared crystallization screens. (Shah et al., 2005)

Some initial crystallization screens developed by these methods are commercially available and used in high-throughput pipelines. The UGA high-throughput pipeline combined seven commercial sparse matrix screens for initial screening. In addition, a locally developed MP1 screen, which contains 48 conditions and emphasizes on the pH, was also included. In the MP1 screen, the pH value ranged from 3.6 to 9.6, and several
popular precipitants (e.g., PEG-400, PEG-4000, and ammonium sulfate) were used. Some salts (e.g., sodium chloride, NaKPhosphate, lithium chloride, and magnesium chloride) were also included in the conditions. These three components (the precipitant, the salt, and the pH) were varied in each condition so that the spread was optimized. (Shah et al., 2005)

It is a laborious and time-consuming process if the initial crystallization screens, which may contain hundreds of conditions, are set up by hand. In addition, the availability of proteins is usually limited. The automation and miniaturization of crystallization experiments can improve the speed with relatively small amounts of proteins. (Pusey et al., 2005) Automated small-volume pipetting systems and standardized premixed crystallization screen solutions are usually used for the automation and miniaturization of crystallization experiments (Stevens, 2000a). As mentioned above, some commercial crystallization screening kits are available. For automated small-volume pipetting systems, the first commercial macromolecule crystallization robot was designed by Cox and Weber and marketed by ICN as the Accuflex. Douglas Instruments designed the Impax and ORYX systems for microbatch under oil. (Pusey et al., 2005) Widely used nanovolume crystallization instruments include the Cartesian Honeybee and the TTP Mosquito. (Berry et al., 2006)

The UGA high-throughput pipeline uses a Cartesian Honeybee crystallization robot (Genomic Solutions) to set up sitting-drop vapor diffusion with Greiner Crystalquick plates for the initial screening. The drop volume used is 400nl, consisting of equal amounts of protein and reservoir solution. Using the Crystalquick plate, three independent samples can be screened at the same time. About 60 samples can be
screened per 40 h week. To manipulate the commercial screens for initial screening or optimization, a Genesis RSP robot (Tecan) is used. (Wang et al., 2005)

Although many diffraction quality crystals are obtained at the initial screening, many others are in the form of microcrystals, spherulites, crystalline precipitates, needles, thin plates, and microcrystals or polycrystals. In addition, the phase separation or crystalline precipitate may also indicate conditions for crystal growth. It is necessary to obtain these leads and optimize the conditions. Once a lead is obtained, optimization may be performed in four ways: 1) scaling up the volume of the trial to obtain larger crystals; 2) fine-tuning the crystallization conditions around the hits to obtain better crystals; 3) adding inhibitors, substrate analogues, or additives to improve the diffraction qualities of crystals; and 4) actively controlling the crystallization environment (e.g., the temperature gradient) while the trial takes place. (Bergfors, 1999, Chayen, 2003)

For the UGA high-throughput pipeline, crystallization conditions are optimized using a single- or double-grid screen (pH and/or precipitant) centered on the lead conditions. Hampton Research Additive Screens 1 and 2 are used in case no diffraction quality crystals are obtained. (Liu, Tempel et al., 2005)

Several methods can be used as salvaging/rescue procedures at this level. Different crystallization methods (e.g., hanging-drop, sitting-drop, and microbatch) may produce crystals of different proteins in different conditions. Some protein samples crystallize by only one specific method. Variations of the vapor diffusion method, the most popular method, are usually used in the initial screen in the high-throughput pipeline. If this method cannot obtain diffraction quality crystals, microbatch
crystallization under oil is a good choice for the salvaging/rescue procedure. Different initial screens, protein concentrations and temperatures can also be employed to salvage important target proteins.

1.3.6 Data collection and processing

To collect crystals from plates for X-ray data collection, protein crystals must be kept wet because the solvent forms a large portion of the crystal lattice. Thin-walled glass capillaries are usually used to mount crystals for data collection at room temperature. However, the more routinely used method now is loop based cryo-mounting (Teng, 1990). This technique can significantly improve the data quality during X-ray data collection. The crystals of interest are mounted by loops and flash-cooled in liquid nitrogen. During data collection, crystals are kept at cryogenic temperatures (lower than -120°C) by blowing cold nitrogen gas onto them. One of the most serious problems of cryocrystallography is the formation of crystalline ice during crystal freezing, which strongly diffracts X-rays and may disrupt the macromolecular crystal lattice. To solve this problem, the protein crystals should be quickly cooled below 130K before any water molecules nucleate to form ice. To prevent the nucleation of water during this process, some crystals need cryoprotectants, such as PEG, glycerol, and lithium sulfate. (McRee, 1999)

The purpose of X-ray data collection is to collect complete and accurate X-ray diffraction data sets of the target protein crystals, from which 3D structural information about the target protein can be extracted. This process depends on the development of the X-ray sources, the X-ray detectors, and the crystallographic software.
The two kinds of X-ray sources are laboratory sources and synchrotron radiation sources. Laboratory sources, which include the sealed tube and the rotating anode, produce X-ray radiation by accelerating electrons at high voltage against a metal target. (McRee, 1999) As the other source of X-ray diffraction (Rosenbauer et al., 1971), synchrotron radiation is emitted when electrons, traveling at speeds close to the speed of light, are forced to change direction by magnetic fields in high-energy storage rings of synchrotron (Elder et al., 1947). The most important characteristics of synchrotron radiation are the substantial improvements of the quality and brightness of the X-ray beams. These attributes enable the characterization of smaller and weakly diffracting crystals. Currently, more than 70 synchrotrons exist around the world. (Sorensen et al., 2006) The Advanced Photon Source (APS) at the U.S. Department of Energy’s Argonne National Laboratory (ANL) provides the Western Hemisphere’s most brilliant X-ray beams. The Southeast Regional Collaborative Access Team (SER-CAT) has two beamlines (Sector 22, APS, ANL): the 22ID insertion-device beamline and the 22BM bending-magnet beamline, which are accessible to the SECSG. Formed in 1997, the SER-CAT (http://www.ser-cat.org/), which is operated by the UGA and consists of 25 member institutions, aims to provide third-generation X-ray capabilities to macromolecular crystallographers and structural biologists in the Southeastern region of the United States.

As a member of the SECSG, the UGA has approximately 24 h beam time per month of operation on both beamlines. The Georgia X-Ray Crystallography Laboratory (GXRCC) at the UGA has a dual port ultrahigh-intensity copper rotating anode source. It
also has a chromium-rotating anode to produce soft X-rays (\(\lambda = 2.29\text{Å}\)) for sulfur phasing of protein structures. (Liu, Tempel et al., 2005)

To detect and record X-ray diffraction data of crystals, detectors such as the single-counter detector used in diffractometers and the area detectors will be used. Area detectors, such as films, image plates, and advanced charge-coupled device (CCD)-based area X-ray detectors, are most commonly used because they can collect larger volumes of reciprocal space than diffractometers can. (McRee, 1999) Similar to films and image plates, the X-ray exposure is recorded in analog form in the CCD detector and is read out later by a computer. CCD detectors improve data quality and increase the speed of X-ray data acquisition. (Gruner & Ealick, 1995) Coupled with third-generation synchrotron sources, CCD detectors can collect a complete set of data in about 20 to 30 minutes. The MarMosaic™ series of detectors, which are seamless multi-element detectors, make a large active detection area using 3 x 3 (MAR CCD 225mm) or 4 x 4 (MAR CCD 300mm) arrays of tiled fiber optic tapers and CCDs (http://www.mar-usa.com/).

For the high-throughput pipelines, the diffraction data of crystals are usually collected at the third-generation synchrotron with CCD detectors. The SER-CAT beamlines are equipped with MAR CCD 300mm (22ID) and 225mm (22BM) detectors (http://www.ser-cat.org/). For in-house X-ray sources at the GXRCC of the UGA, CCD detectors are equipped with a copper rotating anode source, and an image plate detector is associated with the chromium rotating anode source (Wang et al., 2005).

Several photography methods (e.g., rotation, Weissenberg, and precession methods) may be used to record X-ray diffraction data. Compared to other methods, rotation photography (Rossmann, 1985) is more efficient (no screen and need not know
the orientation of crystals before data collection). By now, it is the most commonly used photography method. By this method, the crystal is rotated around a single axis while the detectors collect the diffraction data. (McRee, 1999)

For data collection, several experiment parameters (e.g., exposure time, minimum total rotation range, maximum oscillation angle, detector distance, beam size, and wavelength) should be adjusted. These parameters depend on crystal quality, radiation damage, the purpose for which the data are to be applied, and such equipment as X-ray sources and detectors. Longer exposure time per frame can help in the collection of weak high-resolution data. However, it may prevent the accurate measurement of strong low-resolution data by detectors and cause an increase in the background. Moreover, longer exposure time per frame may cause longer total exposure time for a complete data set where radiation damage may occur. With a lower radiation dose, the minimum total rotation range provides a complete data set, whereas extending the rotation range increases the redundancy of measurements that helps improve the accuracy of data. The oscillation angle, the rotation range per frame, is needed because beam divergence and crystal mosaicity cause the diffraction of each reflection occurring over a finite time and a finite rotation range. Nevertheless, the maximum oscillation angle should be identified to exclude reflection overlaps. An appropriate detector distance should be chosen to avoid reflection overlap and collect high-resolution data. The beam size should not be much larger than the crystal size, causing an increase in the background and requiring a higher total dose of radiation. Different X-ray wavelengths can be chosen for different phasing methods (e.g., multiple anomalous dispersion (MAD) and sulfur phasing) to solve the structure. (Dauter, 2005, Bourenkov & Popov, 2006)
Therefore, a strategy of data collection should be developed to determine whether a crystal is suitable for data collection, what kind of data should be collected, and the choice of experimental parameters for the data collection (Leslie et al., 2002). Based on the examination of several initial images with different exposure times, detector distances, and rotation angles, the quality of the crystal is determined and the data-collection strategy is chosen. Optimum exposure time, the minimum total rotation range, the maximum oscillation angle and the detector distance can be determined. In addition, since the minimum total rotation range is determined, it is better to start the data collection at the proper crystal orientation, allowing the collection of a complete data set as soon as possible. Additional data can then be collected to increase the redundancy of data if the radiation damage is acceptable. This procedure is especially useful if the radiation damage of the crystal is unknown. Fluorescence scan of appropriate heavy atoms should be performed to decide the wavelengths of data collection for MAD phasing. (Gonzalez, 2003)

The automation of the crystal mounting by a robot can further increase the speed of data collection. Researchers from the GXRCC of the UGA and SER-CAT are making it possible for crystals to be mailed to the synchrotron and automatically screened; then the diffraction data of the best crystal can be collected by remote access. Since the beam time at the synchrotron is limited and many targets need to be studied, it is better to screen crystals using in-house X-ray sources to find the diffraction quality crystals for data collection at the synchrotron if possible. The GXRCC of the UGA screens crystals using the Rigake ACTOR/Director system, which can automatically mount and center the crystals, collect images, and index the crystal. The robot is mounted on the copper
rotation anode source. Diffraction quality crystals are then recovered for data collection in the synchrotron. With this system, more than 100 crystals can be screened within 18 hours. (Liu, Tempel et al., 2005)

The collected raw data should be reduced, integrated and scaled for the next step. The GXRCC at the UGA processes its data using either the d*TREK (Pflugrath, 1999) or the HKL2000 suite (Otwinowski & Minor, 1997). The HKL 2000 suite consists of three programs: XdisplayF to present the diffraction pattern visually, Denzo to reduce and integrate data of individual images, and Scalepack to merge and scale data to the full data set. With the HKL 2000 suite, the data processing consists of seven major steps: visualization of the original diffraction pattern, indexing of the diffraction pattern, refinement of the crystal and detector parameters, integration of the diffraction data (Denzo & XdisplayF), discovery of the relative scale factors between measurements, postrefinement of crystal parameters using the whole data set, and merging and statistical analysis of the measurements (Scalepack). Visualization of the data in the original form and the reduction process by XdisplayF helps to recognize complex visual patterns and possible problems. Denzo performs steps 2 to 4 of the data processing to reduce the raw data to a file containing the indexes, background, corrected intensities, and estimated error/uncertainties. Once Denzo is done with the data set, the output is ready to be scaled using Scalepack. The data from all images are scaled together. The postrefinement of parameters is conducted with the whole data set. Statistics obtained by Scalepack can be used to analyze the quality of the diffraction, such as $R_{\text{merge}}$ and $I/\sigma$. In addition, Scalepack can be used to merge or scale different sets of data together on the basis of common reflections. Scaling different sets of data from the same crystal or crystals of the
same protein increases the redundancy and the success rate in the next step of structure determination. (Otwinowski & Minor, 1997)

At this level, several salvaging/rescue methods can be used to salvage crystals, collect better data, or process data for successful structure determination. For example, crystal annealing by sequential freezing and thawing can sometimes improve the diffraction quality of the crystals (Stevenson et al., 2001). Sometimes, the crystal size is larger than the beam size. For some crystals that have uneven quality or that are slightly cracked, better data can be collected at a different area of the crystal. The data can be processed at different resolutions to avoid the influence of ice ring or satellite crystals. As discussed later, we now have the power to solve structures on-site at the synchrotron. With the guidance of structure determination results, the diffraction data of a native or a heavy-atom-containing crystal can be collected at various wavelengths for different purposes (e.g., sulfur phasing, isomorphous replacement, MAD, or refinement).

1.3.7 Structure determination

As a wave, the X-ray possesses amplitude and phase. In crystals, clouds of the electrons of the atoms can scatter X-rays. Since the size of atoms and the distances between neighbor atoms of a protein molecule are similar to the wavelength of X-rays, diffraction occurs. Therefore, the X-ray diffraction pattern of a molecule reveals the distribution of the electrons in the molecule. However, it is too weak to be observable unless it is reinforced by the repetition of the unit cell, the smallest building block in the crystal. The diffraction pattern of molecules of one unit cell can be reduced from the diffraction pattern of the crystal, which can be recorded by the detector. In a single crystal diffraction experiment, each diffracted X-ray that reaches the detector is recorded
as a reflection. The position of the reflection is assigned coordinates (h, k, l). The geometry parameters of a unit cell can be calculated from the spatial arrangement of reflections. A structure factor $F_{hkl}$ is used to describe a diffracted X-ray, which produces reflection $hkl$. $F_{hkl}$ is composed of the amplitude ($|F_{hkl}|$) and phase ($\alpha_{hkl}$). $|F_{hkl}|$ is proportional to the square root of the intensity $I_{hkl}$ of the reflection $hkl$. Since the diffracted X-ray is the result of diffractive contributions from all atoms, $F_{hkl}$ can be described as the sum of the atomic scattering factors of all atoms in the unit cell (Figure 1.1A). If $F$ of all reflections is known, the electron density of molecules in the unit cell can be calculated by Fourier transform, and the structure information of the molecule will be known. However, during the single crystal diffraction experiment, only the intensities of the diffracted X-rays are recorded, whereas all information of the phases is lost (Figure 1.1B). Hence, to solve crystal structures, these phases must be determined. (Rhodes, 1993, Blundell & Johnson, 1976)

To solve the phase problem, the four methods are Patterson summation, direct methods, heavy atom isomorphous replacement, and anomalous scattering. For protein molecules that usually have more than a thousand atoms, the last two methods are usually used.

Isomorphous replacement means that the conformation of molecules and crystal packing do not change when one atom or a very small number of atoms are added or replace the chemically similar ones at the same locations in all unit cells. When a strong diffractor (e.g., a heavy atom or an element of high atomic number) is used for isomorphous replacement, the change of X-ray diffraction pattern is measurable. If the heavy atom positions are known from the difference of diffraction patterns, the phase
problem can be solved from the intensities of the native and isomorphous derivative diffraction patterns. As will be described later, the underlying principle is the “phase triangle” relationship (Figure 1.1C) that relates the structure factors of native diffraction, isomorphous derivative diffraction, and the heavy atom. (Blundell & Johnson, 1976)

The electrons of an element not only scatter the incident X-rays but also absorb and re-emit them with altered phase due to the resonance if the incident photons have similar or higher energy than that of the re-emitted photons. This phenomenon is called anomalous scattering. Anomalous scattering consists of two perpendicular contributions, one real ($\Delta f'$) and the other imaginary ($\Delta f''$), which is always $\pi/2$ in front of the scattered X-rays. (Rhodes, 1993) If the incident photons do not have enough energy, absorption drops sharply. This sudden change is called the absorption edge of this element. Maximum anomalous scattering can be obtained when the incident X-ray wavelength is close to the absorption edge. Some anomalous scatterers (e.g., heavy atoms (Pt, Hg, etc.), and Se) have absorption edges near the wavelength of X-rays used in crystallography. As a result of anomalous scattering, the $I_{hkl}$ does not equal its inverse $I_{h-k-l}$, unless it belongs to reflections of a centrosymmetric structure or in centric zones of non-centrosymmetric structures. As with isomorphous replacement, these intensity differences can be used to localize the position of the anomalous scatterers and solve the phase problem by a phase triangle relationship (Figure 1.1C). (Blundell & Johnson, 1976) However, these intensity differences are usually small and need to be measured accurately. Furthermore, the noise level should be low. Hence, “strong” anomalous scatterers (e.g., heavy atoms and Se) are usually used in experiments of anomalous scattering.
To incorporate heavy atoms into protein crystals for isomorphous replacement or anomalous scattering experiments, the two methods are soaking the protein crystals in solutions of heavy atom salts, and producing and crystallizing heavy-atom-substituted proteins. The heavy atoms can diffuse along the mother liquor channels of protein crystals and bind to the specific sites on the protein surface. (Blundell & Johnson, 1976) The required time ranges from 10min to 2h if the concentration of the heavy atom solution is higher than 10mM. However, many heavy atom salts are expensive and hazardous. As discussed above, using heavy-atom-substituted amino acids such as Se-Met in the expression of protein targets can incorporate heavy atoms into proteins. Nevertheless, some heavy-atom-substituted proteins are difficult to express. In addition, the above two methods may cause nonisomorphism between the native and derivatized crystals, poor quality of diffraction data, and cracked crystals or failure in crystallization. (Dauter, 2002)

By contrast, sulfur is present in almost all proteins and can be used as an anomalous scatterer for the structure determination (Hendrickson & Teeter, 1981, Wang, 1985). Thus, protein structures can be determined directly from native crystals, by so-called direct crystallography (Wang, 2000; Adams and Wang et al., 2003). However, the sulfur X-ray absorption edge is at a long wavelength (5.02Å), out of the energy range of most X-ray sources currently used. In addition, the long X-ray wavelength causes severe absorption effects. However, the $\Delta f''$ is small at short wavelength (only 0.56 electrons at 1.54Å). Therefore, the optimal wavelength should be identified as the compromise of the sulfur anomalous signal and the absorption effects. (Liu et al., 2000) To monitor the signal-to-noise level, a statistic approach, $R_{ass}$ has been introduced. (Fu et al., 2004)
Collecting more data while the radiation damage is acceptable or merging several sets of data of high quality increases the redundancy and improves the accuracy of data and $R_{as}$. Based on idealized data, B.C. Wang’s simulations led to the conclusion that one sulfur atom in 57 amino acids might provide enough signal to solve the structure. (Wang, 1985) The structure solution of $Pfu$-542154 using sulfur phasing approaches this limit by phasing 50 amino acids per sulfur. (Habel & University of Georgia., 2005)

As mentioned above, to solve the phase problem by isomorphous replacement or anomalous scattering, structure factors or positions of heavy atoms or anomalous scatterers should be known first. Their positions can be determined by the difference Patterson function or by direct methods. The $F_H$ is then calculated from their positions. The magnitudes of both $F_{PH}$ and $F_P$ for isomorphous differences or $F_\pm$ and $F$ for anomalous scattering can be calculated from measured intensities. Thus, magnitude or length of each side of the phase triangle is known, and so is the phase of one side—the heavy atom or the anomalous scatterer. Now the phases of the other two sides can be calculated (Figure 1.1C). However, there are two possible ways to orient the triangle if the phase of only one side is known, which means two equally possible solutions—one false and one true. This is known as phase ambiguity (Figure 1.1D). (Blundell & Johnson, 1976)

To solve this problem, some methods, such as MAD and multiple isomorphous replacements (MIR), use additional sets of data. Another method, called single isomorphous replacement anomalous scattering (SIRAS), uses the anomalous scattering data of the derivative together with the single isomorphous replacement data to solve the
phase ambiguity problem. The true solutions of different sets of data should be consistent, whereas the false solutions are not. (Blundell & Johnson, 1976)

However, additional data sets are not absolutely necessary for protein structure determination. Without additional data sets, some methods depend on non-crystallographic symmetry (Bricogne, 1976) or a relative high contribution of the heavy atom to the total diffraction intensity to solve the phase ambiguity (Hendrickson & Teeter, 1981).

By contrast, iterative single anomalous scattering/iterative single isomorphous replacement (ISAS/ISIR), introduced by B. C. Wang, does not require these kinds of information (Wang, 1985). Using good quality SAS or SIR data, this method depends on an iterative noise filtering process to solve the problem of phase ambiguity. This method can also improve the electron density map. In the electron density map, the noise produced by the false phase are actually the statistical variations that can be effectively summed to a constant, while the electron density of the protein regions produced by the true phase is higher than that of the solvent. Thus, the molecular boundary can be defined by the effective summation of density, and then a mask can be produced to identify both protein and solvent regions. The noise of the solvent region and the negative noise of the protein region are filtered to improve the electron density map. This map will be inverse Fourier transformed to the modified phases. Phase combination of the original phases and the modified phases produces improved phases. The improved phases are then used in the next cycle, when the remaining positive noise of the protein region from the first cycle is redistributed into the whole unit cell. After several iterative cycles (generally four to eight cycles) between the real and reciprocal space, the results converge (Figure 1.2). If the
positions of sulfur atoms are known, the ISAS method can be used to solve and refine the phases of protein molecules (Wang, 1985). Thus, direct crystallography (Wang, 2000) can determine the protein structure using a single high quality native crystal.

After locating the heavy atom or anomalous scatterer sites and phasing, it is possible to compute and improve the electron density map. Then a macromolecular model should be fitted into this map. This process is model building. The model is usually not very accurate; thus, a refinement process should be performed to optimize the parameters of the model according to the experimental data and stereochemical expectations. As a result, the electron density map is improved, and the model should be adjusted or rebuilt. Iterations of the refinement and the model reconstruction are generally required before the final accurate structure model is obtained. The final step of structure determination is the validation of the structure model. (Lamzin & Perrakis, 2000)

Many programs are developed for each step from heavy atom or anomalous scatterer localization to validation. To locate heavy atoms or anomalous scatterers, SOLVE (Terwilliger & Berendzen, 1999) uses the Patterson function, while Shake-and-Bake (SnB) (Weeks & Miller, 1999) and SHLEXD (Uson & Sheldrick, 1999) use direct methods. RESOLVE (Terwilliger, 2000) can modify electron density and automatically interpret the map. ARP/wARP (Perrakis et al., 1999) can automatically build protein structure models at resolutions around 2.3Å or higher. Other model building programs include Coot (Emsley & Cowtan, 2004), XtalView (McRee, 1999), and O (Jones et al., 1991). The most commonly used programs for refinement include CNS (Brunger et al., 1998), REFMAC (Murshudov et al., 1997), TNT (Tronrud, 1997), etc. The various
programs and web services for validation include MolProbity (Davis et al., 2007), PROCHECK (Laskowski et al., 1993), etc. The MolProbity validation web service can be used not only for validation but also for refinement (Arendall et al., 2005). The PDB also provides tools for model validation. Some of these programs are integrated into the CCP4 suite (Bailey, 1994). CNS (Brunger et al., 1998) also has a set of programs for almost every step. They are powerful tools, but may require much user intervention. (Lamzin & Perrakis, 2000)

High-throughput structure determination pipelines need a method to use these programs efficiently and automatically with minimal user intervention. Crystallographers of the UGA high-throughput pipeline integrate different crystallographic programs—including SOLVE/RESOLVE, ISAS, AMORE (for the method of molecular replacement (Navaza, 1994)), density modification (DM,(Cowtan & Main, 1993)), ARP/wARP, and REFMAC--into structure determination pipelines. Three pipelines (Liu, Lin et al., 2005)--SCA2Structure, ISASpipe, and AMOREpipe—have been developed with different emphases. In addition to these programs, other major components of the pipelines include a dictionary-driven web-based user interface, a BioPERL-based workflow-management system, a 128-processor computer cluster and analytical tools for harvesting and visualizing results. The dictionary-driven web-based user interface is used to collect basic information of the protein and the crystal, and the input parameters in a few easy steps. Multiple structure-determination jobs can be run in parallel at the 128-processor computer cluster controlled by the workflow-management system. With this method, the program parameter space (e.g., the number of expected heavy-atom/anomalous scatterer sites, space groups, and high-resolution data cutoff) can be screened in large fraction and
fine increments in a time-efficient way. Different phasing programs such as SOLVE and ISAS can also be tried simultaneously to improve the probability of success. This procedure dramatically improves the success rate of structure determination, and jobs are usually completed in 1 h to 2 h. Then the results are harvested, analyzed, and sorted. Key data items and a link to a tar file that contains output files of a given solution are presented in web-based tables. The tar file can be downloaded and inspected to confirm the solution. Because the work is done so quickly, the pipelines can be used as real-time tools to guide the data collection. If the model or part of it cannot be automatically built, the model will be manually completed with XFIT or Coot. (Liu, Lin et al., 2005)

Salvaging/rescue procedures at this step include the manipulation of data. Different methods, programs, packages and pipelines can be used to locate the heavy atom/anomalous scatters and solve the phase problem. Several sets of data can be merged to improve the probability of success.

1.4 Three tiers of the UGA high-throughput pipeline

Some crystal-directed salvaging/rescue procedures have been applied for the targets that failed in the original UGA high-throughput pipeline since September 2003 (Liu, Shah et al., 2005). These procedures include the use of tags to improve the solubility of proteins, further purification to improve the purity of proteins, and modification of proteins to obtain diffraction quality crystals. These methods were also incorporated into the high-throughput pipeline for new targets. Therefore, the UGA crystallography high-throughput pipeline has three tiers: Tier 1, original high-throughput pipeline; Tier 2, the salvaging/rescue procedures; and Tier 3, the improved high-throughput pipeline (Figure 1.3).
Table 1.1a Some commonly used tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Protein</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>His</td>
<td>Hexahistidine tag</td>
<td>Purification</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-transferase</td>
<td>Purification, solubility-enhancing</td>
</tr>
<tr>
<td>MBP</td>
<td>Maltose-binding protein</td>
<td>Purification, solubility-enhancing</td>
</tr>
<tr>
<td>FLAG</td>
<td>FLAG tag peptide</td>
<td>Purification</td>
</tr>
<tr>
<td>CBP</td>
<td>Calmodulin-binding protein</td>
<td>Purification, detection</td>
</tr>
<tr>
<td>Strep-tag</td>
<td>Streptavidin-binding protein</td>
<td>Purification, secretion</td>
</tr>
<tr>
<td>Trx</td>
<td>Thioredoxin</td>
<td>Purification, solubility- and expression-enhancing</td>
</tr>
<tr>
<td>NusA</td>
<td>N-Utilization substance</td>
<td>solubility-enhancing</td>
</tr>
<tr>
<td>SUMO</td>
<td>Small ubiquitin-modifier</td>
<td>solubility-enhancing</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
<td>Detection</td>
</tr>
</tbody>
</table>


Table 1.1b Some commonly used proteases and their cleavage sites

<table>
<thead>
<tr>
<th>Protease</th>
<th>Source</th>
<th>Cleavage site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tev</td>
<td>Tobacco etch virus protease</td>
<td>Glu-Asn-Leu-Tyr-Phe-Gln-/Gly</td>
</tr>
<tr>
<td>Thr</td>
<td>Thrombin</td>
<td>Leu-Val-Pro-Arg-/Gly-Ser</td>
</tr>
<tr>
<td>Xa</td>
<td>Factor Xa protease</td>
<td>Ile-Glu-Gly-Arg/</td>
</tr>
<tr>
<td>EntK</td>
<td>Enterokinase</td>
<td>Asp-Asp-Asp-Asp-Lys/</td>
</tr>
<tr>
<td>3C</td>
<td>Human rhinovirus 3C protease</td>
<td>Leu-Glu-Val-Leu-Phe-Gln-/Gly-Pro</td>
</tr>
</tbody>
</table>

Figure 1.1: Some concepts of X-ray crystallography

A. The structure factor, $F_{hkl}$, is 1) composed of the amplitude ($|F_{hkl}|$) and the phase ($\alpha_{hkl}$) and 2) the sum of the atomic scattering factors of all atoms.

B. The phase problem. During the single crystal diffraction experiment, only the intensities ($\approx |F_{hkl}|^2$) of the diffracted X-rays are recorded, whereas all the information of the phase ($\alpha_{hkl}$) is lost.

C. The phase triangle relationship. $F_{PH}$, $F_P$ and $F_H$ for isomorphous differences or $F_+$. F. and $2F''_A$ for anomalous scattering form the phase triangle. If the magnitude or length of each side and the phase of one side ($F_H$ or $2F''_A$) are known, the phases of the other two sides can be calculated.

D. The phase ambiguity. There are two possible ways to orient the triangle if the phase of only one side is known, one false and one true.
1) The SAS/SIR data are used to calculate the original phases; 2) The phases are Fourier transformed into the electron density map; 3) A density filter is produced to identify both the protein and the solvent regions; 4) The density filter is used to improve the electron density; 5) The improved electron density map is inverse Fourier transformed into calculated phases; 6) The calculated phases are combined with the original or previous phases. 7) The improved phases are used for the next cycle.

Figure 1.3: Flowchart of the UGA high-throughput pipeline

The UGA Crystallography High-throughput Pipeline has three tiers: Tier 1, original high-throughput pipeline; Tier 2, salvaging/rescue procedures; and Tier 3, improved high-throughput pipeline
CHAPTER TWO

Materials and methods

2.1 Surface mutagenesis and gene cloning

2.1.1 Mutation of surface K/E residues to Ala (surface mutagenesis)

Coding sequences of the surface K/E residues were mutated to those of Ala using the QuikChange® Site-Directed Mutagenesis Kit (Stratagene) as described in the instruction manual. Vectors that contained mutated sequences were transformed into JM109 (Promega Co.) and BL21 (DE3) (Invitrogen Co.) as described in the instruction manual. Results were confirmed by DNA sequencing (Sequencing and Synthesis Facility of the University of Georgia, http://www.ssf.uga.edu/).

Primers for the surface mutagenesis of the human p100 TSN region: SM2, forward (f), 5' - GGT GAT GCC AGT GCT GGC GGC GGC GGC GCG ATC TGC TAG C -3', and reverse (r), 5' - GCT AGC AGA TCG CGC CGC CGC CAG CAC TGG CAT CAC C -3'; SM3, f, 5' - GGC ACC CAG TTG GCG GCG CTG ATG GAG AAC ATG CG -3', and r, 5' - CGC ATG TTC TCC ATC AGC GCC GCC AAC TGG GTG CC -3'; SM4, f, 5' - GGT ACC GTG CCC GAG TAG CGG CAG TCG CTG CTC CTG CC -3', and r, 5' - GGC AGG AGA CGC GAC TGC CGC TAC TCG GGC ACG GTA CC -3'.

2.1.2 Gene cloning using the Gateway cloning system

Genes of target proteins were obtained by PCR and transferred to the destination vectors using the Gateway® cloning system (Invitrogen Co.) as described in the
Vectors that contained mutated sequences were transformed into JM109 (Promega Co.) and BL21 (DE3) (Invitrogen Co.) as described in the instruction manual. Results were confirmed by DNA sequencing (Sequencing and Synthesis Facility of the University of Georgia, http://www.ssf.uga.edu/).

Destination vectors used in the genetic truncation of the human p100 TSN region are shown in Table 2.1. Primers for the genetic truncation of the human p100 TSN region:

- **td1**, f, 5’- GAA AAC CTG TAC TTC CAA GGC GGG TCA GGT ATG CCC GTG GAG GAG GTG-3’, r, 5’- GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT TAG CCC TTG GAA TCT GCA AAC T-3’; td2, f, 5’- GAA AAC CTG TAC TTC CAA GGA GGC GGG TCA GGT ATG CTG CAC TTC TAC GTG C-3’, r, 5’- GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT TAG ACC TCT CGT TTG CCG TA-3’; td3, f, as td2-f, r, 5’- GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT TAG GCA GGT GAT AGG GTA C-3’; td4, f, 5’- GAA AAC CTG TAC TTC CAA GGC ATG GTG CAG GAT GTG GAG AC-3’, r, as td2-r; td5, f, as td4-f, r, as td3-r; td6, f, as td4-f, r, 5’- GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT TAG CGG CTG TAG CCA AAT TC-3’; P1, f, 5’- GAA AAC CTG TAT TTT CAG GGC ATG GTG CAG GAT GTG GAG AC-3’, r, 5’- GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT TAG CCC TTG GAA TCT GCA AAC T-3’; PW, f, as P1-f, r, as td6-r.

### 2.1.3 Other vectors used in this research

The vector used for the expression of *Pyrococcus furiosus* proteins was a SECSG modified pET vector, pET24 dBam (Jenney et al., 2005).

The vector used for the expression of *Clostridium thermocellum* targets was a SECSG modified vector, pET15G vector, which encodes the His<sub>6</sub> tag and a thrombin
cleavage site (MGSSHHHHHSS-GLVPRGSQS-TSLYKKAGL) in the pET15b (Novagen) backbone vector (Luan et al., 2004).

For the targets from *Archaeoglobus fulgidus*, the sequence that encodes the TEV protease cleavage site (ENLYFQGGSG) was added to the 5’ of the target gene during PCR. The vector pDEST527, which contains His$_6$ tag sequence (MRSGSHHHHHHRS-DITSLYKKAG) at the 5’ of the cloning site, was used to express His$_6$-TEV cleavage site-target proteins.

The host strain for these vectors was *E. coli* BL21DE3.

### 2.2 Protein expression and solubility test

#### 2.2.1 Expression of the native proteins

Cells (100µL from freezer stocks) were grown in 50ml LB medium with appropriate antibiotics at 37°C for 8 h and transferred to 1L LB medium. The protein expression was induced with 0.5 mM IPTG at OD$_{600}$~0.5. After 3 h following the IPTG induction, cells were harvested by centrifugation (5000g × 15 min). The cell pellets were stored in -20°C.

#### 2.2.2 Autoinduced expression of Se-Met substituted proteins

Cells (100µl from freezer stocks) were grown in 50 ml PA0.5G medium with appropriate antibiotics at 37°C for 8 h and transferred to 1L of PASM5052 for 20 h in the presence of 125ug/ml Se-Met. Cells were harvested by centrifugation (5000g × 15 min) and cell pellets were stored in -20°C. (Studier, 2005)

The 5L working volume fermentors of the UGA Bioexpression and Fermentation Facility were used for proteins whose expression levels were comparatively low.
2.2.3 Solubility test

For the solubility test, cells were grown in 5ml LB medium or PA0.5G. The protein expression was induced with 0.5 mM IPTG at OD600~0.5 or transferred into 20ml PASM5052 in the presence of 125ug/ml Se-Met after 8h. Cells were collected after 3 h following the IPTG induction or after 20 h in PASM5052. Cells were lysed as described below. Results were analyzed by SDS-PAGE.

2.3 Protein purification

2.3.1 Two-step purification of recombinant proteins that contain N-terminal His6 tags (AHHHHHHHG-) and the workflow of ÄKTApurifier

Isolated cell pellets (2.2) were re-suspended in five volumes of the lysis buffer (20 mM sodium phosphate buffer (pH 7.6), 200 mM sodium chloride, 20mM imidazole, 5mM β-ME and 1mM PMSF), and were disrupted with an ultrasonic homogenizer (6×30 second, on ice). For the hyperthermophilic proteins (e.g., *Pyrococcus furiosus* proteins), the sonicated sample was heat-treated at 70°C for 1 h. DNase (Sigma) was added as 500 U per gram of original cell pellets. All samples were centrifuged at 12,000 rpm for 30 min at 4°C, followed by the filtration of supernatant with 0.45 µm filters. The filtrated supernatant was applied to a 5 mL Ni-Histrap HP column (GE Healthcare) and eluted with a 20–500 mM imidazole gradient at 4°C. Fractions containing the protein were pooled, and concentrated to 5 ml (Amicon Ultra, Millipore). For purification using an ÄKTApurifier (GE Healthcare), proteins were eluted with a single 500 mM step, to ensure that target protein peaks were collected into the 5 ml loop. Samples were then loaded manually or automatically (ÄKTApurifier) onto a Superdex 200 Hiload 16/60 gel filtration column (GE Healthcare) and eluted with 20 mM HEPES buffer, pH 7.6
containing 100 mM sodium chloride, 5 mM β-ME, and 2 mM DTT. Fractions containing the protein were then pooled and concentrated above 10 mg/ml for the crystallization trials. Protein purity was checked by SDS-polyacrylamide gel electrophoresis (SDS-PAGE).

To improve reproducibility and speed, an automated robotic chromatographic platform, ÄKTApurifier (GE Healthcare), was used to couple the Ni affinity chromatography and gel filtration chromatography. The method for this automated two-step protein purification was programmed in UNICORN™ software. With this method, three proteins could be purified simultaneously. Briefly, crude extracts of three proteins were applied sequentially onto three 5-ml Ni columns. The in-line air sensor of the sample pump stopped the loading when air was detected, allowing unattended loading. One by one, the proteins were eluted from the Ni column by 500mM imidazole in the appropriate buffer, collected automatically into a 5 ml loop by peak detection, applied to a gel filtration column, and collected in 2ml fractions by the fraction collector. Such parameters as UV absorbance, fractions, pressure, pH, and ionic strength were recorded (Figure 2.1). The tubing and loop were washed by appropriate buffer after each step to avoid cross-contamination. The result was shown by SDS-PAGE (Figure 2.2). Purified proteins were concentrated by ultrafiltration (Amicon Ultra-15 Centrifugal Filter Units, Millipore) and stored in -80°C for crystallization trials. It usually takes 10 h to 12 h for the purification of three proteins by the two-step method, depending on the initial sample volumes. The matrix of the Histrap column was automatically recharged with Ni²⁺ following the metal-stripping step using methods in UNICORN™.
2.3.2 Additional purification by ion exchange chromatography and hydroxyapatite chromatography

The samples were applied to a 5 ml Q (anion exchange) or SP (cation exchange) ion exchange column (GE Healthcare) and eluted with a 20–500 mM NaCl gradient at 4°C. Fractions were checked by SDS-PAGE, and pure fractions containing the target protein were pooled.

The samples were applied to a 20 ml home-packed hydroxyapatite column and eluted with 1 M K$_2$HPO$_4$ at 25°C. Fractions were checked by SDS-PAGE, and pure fractions containing the target protein were pooled.

2.3.3 Purification of GST-tagged protein

Isolated cell pellets were re-suspended in five volumes of the lysis buffer (1× PBS, pH 7.3, 5 mM PMSF and 1 mM DTT), and disrupted with an ultrasonic homogenizer (6×30”, on ice). Samples were centrifuged at 12,000 rpm for 30 min at 4°C, followed by the filtration of the supernatant. The supernatant was loaded onto a GSTrap HP, 5 ml column (GE Healthcare) previously equilibrated with the binding buffer (1× PBS, pH 7.3 and 1 mM DTT) and eluted with the elution buffer (10 mM reduced glutathione, 50 mM Tris-HCl, pH 8.0). Fractions containing the protein were pooled and analyzed by SDS-PAGE.

2.3.4 Tag removal

The expression vector of Tev protease was a gift from Dr. Ming Luo’s lab. (University of Alabama at Birmingham) His$_6$-tagged Tev protease was expressed and purified as described in 2.2.1 and 2.3.1 respectively.
Target protein samples were subjected to Tev (0.1 OD (A280) of purified Tev protease per mg of target protein) or thrombin (GE Healthcare; 5U/mg) cleavage at room temperature for ~3 h. Samples were then dialyzed against the appropriate buffer overnight at 4°C and purified as described in 2.3.1 or 2.3.3 to remove the tag.

His$_6$-tagged Tev protease was removed from the target protein sample together with His$_6$ tag during Ni affinity chromatography. Thrombin, a serine protease, was removed by HiTrap Benzamidine FF affinity column. The sample was loaded onto a HiTrap Benzamidine FF, 5ml column (GE Healthcare) previously equilibrated with the washing buffer (0.05 M Tris-HCl, pH 7.4, 1 M NaCl). Target proteins were collected from the void volume, while the thrombin was eluted with the elution buffer (20mM p-aminobenzamidine in the washing buffer).

2.4 Reductive methylation of surface lysine residues

In this method, 20 µl of 1 M dimethylamine-borane complex (DMAB) and 40 µl formaldehyde were added to 1 ml protein samples (10 mg/mL). The reaction mixture was incubated at 4°C for 2 h in the dark while shaking. The above steps were repeated two more times. Then, 10 µl DMAB was added into the mixture and the mixture was incubated for another 18 h at 4°C in the dark while shaking (Rypniewski WR et al. 1993). The buffer was exchanged through two buffer exchange columns (Econo-Pac P6 Cartridge; Bio-Rad).

2.5 Crystallization

The initial screening involved the use of seven commercial sparse-matrix screens: Crystal Screen, Crystal Screen II, MemFac, PEG/Ion, Crystal Screen Cryo (Hampton
Research), Wizard I and II (Dedode Genetics), and the MPI screen that was developed by SECSG (Shah et al., 2005).

A Cartesian Honeybee crystallization robot (Genomic Solutions) was used to set up sitting-drop vapor diffusion with Greiner Crystalquick plates. The drop volume was 400 nl, consisting of equal amounts of proteins and reservoir solutions. Using the Crystalquick plate, three independent samples could be screened at the same time. About 60 samples could be screened per 40-h week.

In addition, the method of modified microbatch under oil was also used for the initial screening or optimization by a locally modified Douglas Instruments ORYX robot using 1 µl drops that contained equal volumes of proteins and reservoir solutions on a 72-well Nunc plate (Nalge Nunc International). The 4:1 paraffin/silicone oil mixture was added after set-up to seal the plate. For optimization, crystallization conditions were optimized using a single- or double-grid screen (pH and/or precipitant) centered on the lead conditions. Hampton Research Additive Screens 1 and 2 would be used in case no diffraction quality crystals were obtained.

Once the plate had been set up, it was put into the CrystalFarm incubator (Discovery Partners International) for storage, imaging, and scoring. To manipulate the commercial screens for the initial screening and the grid screens for the optimization, a Genesis RSP robot (Tecan) was used.

2.6 Data collection and processing

Crystals (with dimensions greater than 50µm) were mounted and flash-frozen with or without cryo-protective solutions. Diffraction quality of crystals was then
identified by screening in-house using a Rigaku ACTOR robot. Good crystals were recovered for later data collection.

Data sets of these crystals were collected at -173°C using synchrotron radiation (SER-CAT, 22ID, Advanced Photon Source, Argonne National Laboratory) with MAR CCD225 or MAR CCD300. Data sets were collected at selenium absorption edge if applicable. The crystal-to-detector distance, oscillation angle, exposure time, and scan range depended on individual crystals. The data were processed with the program HKL2000. A .sca file and a .log file were generated after the data processing.

2.7 Structure determination

SECSG web-based SCA2Structure and ISASPipe pipeline were used for the phase calculation. The .sca file and the sequence file of proteins were loaded into the pipeline. Parameters were set to appropriate values. Among them, the solvent content of the crystal was calculated from the unit cell (values obtained from data processing) and the molecular weight using Matthews_coef from the CCP4 suite (Bailey, 1994). The Matthews coefficient (Vm) and the probability of molecules in the asymmetric unit were also obtained from this program. The pipeline results were sorted according to the number of atoms autotraced by RESOLVE, SOLVE Z score and SOLVE FOM (figure of merit). The initial models of these results were then inspected (XFIT XtalView or COOT) to choose the best one. The .pdb file and .mtz file of the best model were used for refinement. An example of the usage of the pipeline is described in 4.2.4.

The model was adjusted manually by Coot and refined by REFMAC. The model was rebuilt automatically using ARP/wARP6.1.1 if applicable. The final model was validated using MolProbity and PROCHECK before it was submitted to the PDB.
2.8 Function prediction of hypothetical proteins

Functions of hypothetical proteins can be generally predicted, based on sequence and structure comparison with proteins of known functions, existence of cofactor, and gene neighbors. The methods used here for function prediction included (a) the amino acid sequence BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1997) to find homologous proteins or SUPERFAMILY server (http://supfam.org/SUPERFAMILY/index.html) (Gough et al., 2001) to find remote relationships with structure-known proteins; (b) structure similarity screening using a DALI (Distance Matrix Alignment; http://www.ebi.ac.uk/dali/) (Holm & Sander, 1996), SSM (Second Structure Matching; http://www.ebi.ac.uk/msd-srv/sss/) (Krissinel & Henrick, 2004), or VAST (Vector Alignment Search Tool; http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html) (Gibrat et al., 1996) server; (c) finding gene neighbors of the protein (NCBI, http://www.ncbi.nlm.nih.gov/sites/entrez; TIGR, http://www.tigr.org/; KEGG, http://www.genome.jp/kegg/kegg2.html; Genome Reviews of EMBL-EBI, http://www.ebi.ac.uk/GenomeReviews/) and operon analysis by MicrobesOnline Operon Predictions (http://www.microbesonline.org/operons/) (Price et al., 2005) to classify further the target protein and predict possible functions; and (d) using the ProFunc server (http://www.ebi.ac.uk/thornton-srv/databases/profunc/) (Laskowski et al., 2005) to identify the possible biochemical function of a protein from its 3D structure by finding possible cation/anion-binding regions, Helix-turn-helix (HTH) DNA-binding motifs, possible enzyme active sites, as well as ligand or DNA binding sites. In general, bioinformatics databases (e.g., NCBI, TIGR, KEGG, and UniProt (http://beta.uniprot.org/)) summarize information of hypothetical proteins predicted from
genome sequences and annotate possible functions for them. However, results of different methods do not always agree with each other if not conflict because of different criteria. Therefore, human interpretation is generally needed in addition.
Table 2.1: Destination vectors used in the expression of genetic truncated human p100 TSN regions

<table>
<thead>
<tr>
<th>Vectors</th>
<th>pDEST 527</th>
<th>pDEST 544</th>
<th>pDEST 565</th>
<th>pDEST 566</th>
<th>pDEST 596</th>
<th>pDEST 598</th>
<th>pDEST 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter</td>
<td>T7</td>
<td>T7</td>
<td>T7</td>
<td>T7</td>
<td>Trc</td>
<td>Tac</td>
<td>T7</td>
</tr>
<tr>
<td>Tag</td>
<td>His6</td>
<td>His6- NusA</td>
<td>His6- GST</td>
<td>His6-MBP</td>
<td>MBP</td>
<td>MBP-His6</td>
<td>No tag</td>
</tr>
<tr>
<td>Selection Marker</td>
<td>Amp</td>
<td>Amp</td>
<td>Amp</td>
<td>Amp</td>
<td>Amp</td>
<td>Amp</td>
<td>Amp</td>
</tr>
</tbody>
</table>
Figure 2.1 Example of the two-step purification using ÄKTApurifier 3D for automatic purification of three proteins simultaneously

A, the sample loading and column washing step; B, two-step (Ni+GFL) purifications of PF1835, PF0147 and PF1098. 1. washing with the loading buffer; 2. elution of target proteins from Ni- affinity column; 3. elution of target proteins from gel filtration chromatography.
Figure 2.2  SDS-PAGE results of PF1835 (22.4 kDa), PF0147 (23.0 kDa) and PF1098 (22.5 kDa), purified by the process described in Figure 2.1

M, protein marker.
CHAPTER THREE

Salvaging/rescue trials of 34 Pyrococcus furiosus proteins that failed at first trials

As discussed in the introduction, several methods can be applied to salvage the failed targets at each step of high throughput pipeline. Additional purification methods, such as ion exchange chromatography and hydroxyapatite chromatography, may improve the purity of target proteins. Protein modification methods, such as reductive methylation and Se-Met substitution, may improve the stability and/or change the solubility of target proteins. These methods and other methods, such as alternative set-up methods for crystallization trails, crystal annealing, and data merging may help to obtain diffraction-quality crystals of target proteins. For an individual target, different salvaging methods may have different impacts on its crystallization or crystal diffraction quality. Combining salvaging methods from different steps may be also helpful.

The UGA structure determination pipelines are powerful for solving structures in a time-efficient way by parameter space screening (Liu, Lin et al., 2005). The UGA structure determination team of the SECSG carefully selects the data collection strategy and processes data for individual protein crystals. More data sets are collected as required. These procedures exploit the limitation of the diffraction data quality for structure determination.

Therefore, the present work mainly focused on the target salvaging trials of protein purification and modification (e.g., ion exchange chromatography, hydroxyapatite chromatography, reductive methylation and Se-Met substitution), and studied their ability
to improve diffraction quality of target protein crystals. These salvaging/rescue procedures belong to Tier 2 of the UGA crystallography high-throughput pipeline.

3.1 Salvaging/rescue targets

A set of 34 Pyrococcus furiosus proteins that failed Tier 1 trials was selected as salvaging/rescue targets. Among these proteins, twenty were originally studied in the native form, while the other fourteen proteins were studied with the Se-Met form. Original reasons for choosing these proteins for salvaging included: (a) lack of crystal formation, (b) lack of crystal diffraction, and (c) insufficient diffraction quality. (Table 3.1)

3.2 Methods employed in the salvaging/rescue trials

Among the 34 target proteins, 25 proteins were re-expressed and re-purified, and 9 originally purified proteins were used.

Because of their comparatively low expression level, three proteins (PF0907, PF0616, and PF0175) were expressed in 5L PASM5052 using 5L working volume fermentors (2.2.2).

As discussed in 2.3.1, ÄKTApurifier was used to purify proteins in a fast and automatic way. Twenty proteins were successfully purified by ÄKTApurifier. As shown in Table 3.1a, the amounts of purified proteins ranged from 0.22 mg to 135.6 mg. The average amount was 37.21 mg. The proteins that had two peaks were assumed to be two different forms and were counted separately. The concentrations of these 20 proteins ranged from 3.5 mg/mL to 113 mg/mL, with an average concentration of 34.49 mg/mL. Purified proteins were submitted for crystallization trials and structure determination.
PF1278 was not submitted because of its low concentration and amount. Therefore, finally crystallization trials were performed for 19 proteins.

For the other five re-purified proteins (PF1183, PF0385, PF0492, PF0619 and PF1025), in addition to Ni affinity chromatography and gel filtration chromatography, ion exchange chromatography was used. Additional hydroxyapatite chromatography was performed for PF0619. (Table 3.1b)

As discussed in 1.2.4, the substitution of Met residues with SeMet residues can be used as a modification method to salvage native targets. 20 proteins originally studied in the native form were purified in the Se-Met form. However, Se-Met substitution may also have negative impact on the crystallization of some proteins as discussed later. Purification of those proteins in native forms might salvage those targets. This work was not included in this research. (Table 3.1a)

Another protein modification method, the reductive methylation of lysine residues, was also used to salvage targets. In this research, it was performed for 11 proteins (Table 3.1b).

In addition of sitting-drop vapor diffusion, the modified microbatch method was used for the crystallization trials of PF1174, PF1044, and PF0421 (Table 3.1b). Optimization was performed for the above targets with different protein concentrations.

3.3 Results of salvaging trials

3.3.1 The impact of salvaging methods on crystallization and diffraction quality of crystals

For five salvaging targets to which additional purification step applied, their crystallization or crystal quality was not improved, although the purity was improved. As
an example, SDS-PAGE of PF1183 and PF0385 after purification of each step is shown in Figure 3.1.

For six proteins that did not grow crystals in the native form, four had hits in crystallization trials in the Se-Met form. For fourteen proteins with no or poor diffraction in the native form, crystal diffraction qualities of six proteins were improved in the Se-Met form. On the other hand, three proteins that had crystals in the native form, failed in crystallization trials in the Se-Met form. (Table 3.1a) The average number of Met in ten proteins that obtained crystals or crystals with improved diffraction quality in Se-Met form is 2.6 per 100 residues, whereas the average number of Met in other ten proteins is 1.8 per 100 residues.

For five proteins that cannot get crystals, crystals of one protein were obtained after the reductive methylation. For other five proteins with poor diffraction, crystal diffraction qualities of these proteins were not improved after the reductive methylation. (Table 3.1b)

Alternative crystallization trials set-up method, the modified microbatch method, produces crystals with higher diffraction resolution for all three targets, but their diffraction quality was not good enough for structure determination (Table 3.1b).

**3.3.2 Four structures were obtained**

As a result, these salvaging methods were involved in the production of four structures (PF0907, PF0725, PF1771 and PF0168). Two structures (PF0907, PF0725) were solved as the result of Se-Met substitution (Table 3.2, 3.4). Two structures (PF1771 and PF0168) were solved after Se-Met substitution and additional salvaging work (reductive methylation) from the other groups of the laboratory.
3.4 Function prediction of PF0907 and PF0725

3.4.1 The predicted function of PF0907

A ribbon representation of the structure of PF0907 is shown in Figure 3.2A. It adopted a β barrel topology consisting of eight anti-parallel β strands.

An amino acid sequence similarity search against the PDB database by BLAST revealed that no structures shared more than 30% identical amino acids with over 80% amino acids of PF0907. It was therefore considered as a unique target. Results of a structural similarity search using the DALI server revealed significant structural homologies with domain II of elongation factors EF-1α, EF-Tu, EF-G, and SelB, as well as domain II and IV of the initiation factor IF2/eIF5B (Table 3.3; Figure 3.2B). All these proteins are GTPases involved in translation and interaction with ribosomes. EF-Tu (in bacteria) and EF-1α (in eukarya and archea) can form ternary complexes with GTP and all 20 naturally occurring aa-tRNAs to carry aa-tRNAs into the A-site of ribosome for protein translation (Nissen et al., 1995, Nissen et al., 1999, Vitagliano et al., 2004). The selenocysteine (Sec)-tRNA^{sec} is carried by a specialized translation elongation factor SelB (Leibundgut et al., 2005). EF-G translocates the peptidyl tRNA from the A-site to the P-site during the elongation phase of protein translation (Aevavarsson et al., 1994). IF2/eIF5B is one of the universal translation initiation factors of all three kingdoms (Roll-Mecak et al., 2000). Domains II of these proteins indicates common structural features.

Domain II of elongation factors (e.g., EF-Tu, EF-1α and SelB) binds the 3’ CCA-aa and the 5’ end of aa-tRNA (Figure 3.3). The aminoacyl group of the single-strand CCA-aa end is docked into a pocket of domain II. In the EF-Tu of *Thermus aquaticus*, this pocket is formed with side chains of Phe229, Asp227, Glu226, and Thr239
(corresponding to Val35, Glu33, Glu32 and Ile44 of PF0907). The free amino group interacts with Asn285 and His273 (corresponding to Arg85 and Asn73 of PF0907) through hydrogen bonds. The 2’-OH group of the terminal ribose hydrogen bonds the conserved Glu271 (corresponding to Glu71 of PF0907). This residue also stacks over the adenine base of 3’ CCA-aa on one side and the conserved residues Val237 and Ile231 (corresponding to Val42 and Ile37 of PF0907) on the other side. The conserved Arg300 (corresponding to Lys99 of PF0907) interacts with the 5’ phosphate of aa-tRNA. (Nissen et al., 1995, Nissen et al., 1999) Many of these residues, which are involved in the interaction between the elongation factors and aa-tRNA, are conserved in PF0907. However, these residues are not conserved in EF-G and IF2/eIF5B whose domains II do not bind aa-tRNAs. Therefore, PF0907 may have the function of aa-tRNA binding (Figure 3.4). Conserved residues in various SelB proteins (e.g., Asp191, His192, and Arg247 in SelB of Methanococcus maripaludis) are unique to SelB for specific interaction with Sec-tRNA^{sec}. Double mutation of the first two residues to other residues reduces the binding activity (Leibundgut et al., 2005). In PF0907, these residues correspond to Glu32, Glu33 and Arg85. Therefore, although the Arg residue is conserved in the SelB proteins and PF0907, PF0907 may not have the ability to bind Sec-tRNA^{sec}.

Gene neighbors of PF0907 were analyzed for further identification of the function of PF0907. The MicrobesOnline Operon Prediction method predicted that PF0907 belongs to an operon composed of PF0905, PF0906, PF0907 and PF0908. In addition to Pyrococcus furiosus, this operon is also present in Pyrococcus abyssi, Pyrococcus horikoshii and partially in Thermococcus kodakarensis (Pyrococcus sp.). PF0905 is a putative ABC transporter ATP-binding protein. PF0906 may function as permease in the
ABC transport system. PF0908 is a TatD-like protein that may have DNase activity. The gene of TatD usually appears in the operon of ABC transporter genes.

Therefore, PF0907 may bind aa-tRNA and play a role in transportation of specific substrates across cell membranes.

3.4.2 The predicted function of PF0725

The structure of PF0725 is composed of a Rossmann-fold domain, which consists of a five-stranded parallel β sheet sandwiched between two bundles of α helices, and two other anti-parallel strands vertical to the β-sheet. Analysis of the electron density maps revealed the presence of bound Coenzyme A (CoA), an enzyme cofactor essential for all living organisms. (Figure 3.4A)

A CoA molecule with extended conformation is accommodated in a long crevice on the surface of the protein formed by loops between α-helices and β-strands (L1-L5) (Figure 3.4B). The protein fixes the position of the adenine head of CoA between its loops by using hydrogen bonds and hydrophobic interactions. The adenosine ring of CoA inserts into the hydrophobic pocket formed by Val12, Gly13, Val39, Pro41, Val69, Val73 and Val77 from loops L1, L2, and the helix H3. In addition, Arg53 and Glu80 interact with the AN1 of the adenosine ring through water-mediated hydrogen bonds. The 3’ phosphate group of the ribose at the ppAp moiety of CoA is hydrogen bonded by Ser15 (OG-AO8), Lys20 (NZ-AO8), Asn40 (ND2-AO7, ND2-AO3’) and Tyr43 (OH-AO7). The 2’-OH of the ribose interacts with Asn40 (OD1-W3-AO2’) and Asn42 (OD1-AO2’) through hydrogen bonds. A hydrogen bond is also formed between Lys20 and the 5’-diphosphate of the ribose (NZ-AO2) (Figure 3.4C). At the other end of CoA, the panthetheine moiety extends toward the C-terminal helix of PF0725. A disulfide bridge is
formed between Cys116 of the protein and the terminal PS1 of CoA. In addition, the carbonyl oxygen atom of Pro93 forms a hydrogen bond with the amide nitrogen atom (PN8) of the first amide group, and NE2 of Gln 92 hydrogen bonds with the carbonyl oxygen atom (PO5) of the second amide group (Figure 3.4D). (Hiyama et al., 2006)

An amino acid sequence similarity search against the non-redundant protein database by BLAST revealed that PF0725 belongs to a family of single-domain CoA-binding proteins that are conserved in bacteria, archaea, and a few eukaryal taxa, and their function are still unknown (Hiyama et al., 2006). Although many of the residues involved in interactions with CoA are not strictly conserved, some of them (corresponding to Asn40, Pro41, Gln92, and Cys116 of PF0725) are absolutely conserved in the family. A structure of this family, TT1466 (TTHA1899) from *Thermus thermophilus* HB8 (PDB ID 1iuk), was solved before the deposition of the structure of PF0725. It was structurally homologous to PF0725 (Z score 17.6, r.m.s.d. 1.6Å and sequence identity 29%). However, since the structure of TT1466 did not contain CoA, structural information is still needed to decipher the interaction between this class of proteins and the CoA molecule. The structure of PF0725 is the first structure that has this information. An *E. coli* ortholog of PF0725, YccU, interacts with the ribosomal stalk protein L7/L12 (RpIL) (Butland et al., 2005). Furthermore, results of DALI structure similarity search (Table 3.5) revealed that the structure of Pf0725 is structurally homologous to the CoA-binding domain of succinyl-CoA synthetase (SCS) from *Escherichia coli* (PDB ID 1eud, Z score 12.3, r.m.s.d. 2.3Å and sequence identity 18%). The CoA binding crevice of single-domain CoA-binding proteins is deeper than that of SCS (Hiyama et al., 2006). In conclusion, single-domain CoA-binding proteins can be
used as carriers of CoA for various purposes (e.g., succinyl-CoA synthesis, as reducing agents (Harris et al., 2005), or translation control).
Figure 3.1 SDS-PAGE results of PF1183 and PF0385 after each purification step
Ni, Ni affinity chromatography; GFL, gel filtration chromatography; IEX, ion exchange chromatography
Figure 3.2 The ribbon representation of PF0907 (A) and the structure-based alignment of PF0907 and its homologues (B).

*, refers to the residues that are involved in the interaction between the elongation factors (EFTu, EF1α and SelB) and aa-tRNAs.
Figure 3.3 Comparison of the aa-tRNA binding region of the EF-Tu from *Thermus aquaticus* (PDB ID, 1TTT) and the corresponding region of PF0907

In the complex of the EF-Tu of *Thermus aquaticus* (1TTT,Cyan) and Phe-tRNA\textsuperscript{Phe} (purple), the aminoacyl group (Phe) of Phe-tRNAPhe docked into a pocket of EF-Tu Domain II, formed by F229, D227, E226, and T239. The free amino group of Phe interacts with N285 and H273 through hydrogen bonds. The 2'-OH group of the terminal ribose hydrogen bonds the conserved E271. E271 also stacks over the adenine base of 3' CCA-aa on one side and the conserved residues V237 and I231 on the other side. The conserved R300 interacts with the 5' phosphate of aa-tRNA. Many of these residues are conserved in PF0907 (Green).
Figure 3.4 The structure of PF0725 with CoA
A. The ribbon representation of PF0725/CoA; B. The surface representation of PF0725/CoA, in the same orientation as in A; A CoA molecule with the extended conformation is accommodated in a long crevice on the surface of the protein. C. The interaction between PF0725 and CoA adenosine and ppAp moiety; The adenosine ring of CoA inserts into the hydrophobic pocket formed by V12, G13, V39, P41, V69, V73 and V77. In addition, R53 and E80 interact with the AN1 of the adenosine ring through water-mediated hydrogen bonds. The 3' phosphate group of the ribose at the ppAp moiety of CoA is hydrogen bonded by S15, K20, N40 and Y43. The 2'-OH of the ribose interacts with N40 and N42 through hydrogen bonds. A hydrogen bond is also formed between K20 and the 5'-diphosphate of the ribose. D. The disulfide bridge between the CoA ligand and the protein. At the other end of CoA, a disulfide bridge is formed between C116 of the protein and the terminal PS1 of CoA. In addition, the carbonyl oxygen atom of P93 forms a hydrogen bond with the amide nitrogen atom of the first amide group and Q 92 hydrogen bonds the carbonyl oxygen atom of the second amide group of the CoA panthetheine moiety.
Table 3.1a Salvaging efforts for 20 *Pyrococcus furiosus* proteins failed in the native form

<table>
<thead>
<tr>
<th>Original protein</th>
<th>MW (KDa)</th>
<th>pI</th>
<th>Expression volume</th>
<th>Salvaging procedures (Amounts, mg)</th>
<th>Crystals (Å, MO)</th>
<th>Structure (PDB ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a. S.S  b. A.P  c. R.M  d. M.M</td>
<td>before  after</td>
<td></td>
</tr>
<tr>
<td>PF1278</td>
<td>13.3</td>
<td>6.5</td>
<td>1</td>
<td>0.2</td>
<td>3.6 cracked  NA</td>
<td></td>
</tr>
<tr>
<td>PF0345</td>
<td>9.2</td>
<td>5.1</td>
<td>1</td>
<td>51.5</td>
<td>poor diffraction  NC</td>
<td></td>
</tr>
<tr>
<td>PF1771</td>
<td>43.1</td>
<td>5.0</td>
<td>1</td>
<td>31.2</td>
<td>no diffraction  3.2 medium  *</td>
<td></td>
</tr>
<tr>
<td>PF1349</td>
<td>39.7</td>
<td>6.2</td>
<td>1</td>
<td>37.2</td>
<td>NC  2.8 High</td>
<td></td>
</tr>
<tr>
<td>PF0907</td>
<td>11.8</td>
<td>9.9</td>
<td>5</td>
<td>30.1</td>
<td>2.1  2.0, medium  1XE1</td>
<td></td>
</tr>
<tr>
<td>PF1085</td>
<td>47.9</td>
<td>5.3</td>
<td>1</td>
<td>1) 7.0; 27.5</td>
<td>7.6  NC</td>
<td></td>
</tr>
<tr>
<td>PF1742</td>
<td>48.0</td>
<td>5.0</td>
<td>1</td>
<td>18.1</td>
<td>3, high  3.8, medium</td>
<td></td>
</tr>
<tr>
<td>PF0219</td>
<td>18.1</td>
<td>5.7</td>
<td>1</td>
<td>88.5</td>
<td>4.6  no good crystal</td>
<td></td>
</tr>
<tr>
<td>PF0584</td>
<td>18.2</td>
<td>9.8</td>
<td>1</td>
<td>75.5</td>
<td>NC  NC</td>
<td></td>
</tr>
<tr>
<td>PF1584</td>
<td>17.8</td>
<td>4.9</td>
<td>1</td>
<td>8.4</td>
<td>NC  NC</td>
<td></td>
</tr>
<tr>
<td>PF1205</td>
<td>17.3</td>
<td>8.9</td>
<td>1</td>
<td>22.6</td>
<td>poor diffraction  NC</td>
<td></td>
</tr>
<tr>
<td>PF0168</td>
<td>19.9</td>
<td>4.2</td>
<td>1</td>
<td>1) 62.4; 2) 96.0</td>
<td>3.3, medium  2.8, high  *</td>
<td></td>
</tr>
<tr>
<td>PF0316</td>
<td>40.2</td>
<td>4.5</td>
<td>1</td>
<td>6.3</td>
<td>10  7.5</td>
<td></td>
</tr>
<tr>
<td>PF0548</td>
<td>83.3</td>
<td>4.6</td>
<td>1</td>
<td>46.2</td>
<td>NC  8.0</td>
<td></td>
</tr>
<tr>
<td>PF0903</td>
<td>35.8</td>
<td>6.8</td>
<td>1</td>
<td>135.6</td>
<td>2.8 medium  4.1 high</td>
<td></td>
</tr>
<tr>
<td>PF0904</td>
<td>28.3</td>
<td>9.9</td>
<td>1</td>
<td>1) 26.3; 53.8</td>
<td>x  x</td>
<td></td>
</tr>
<tr>
<td>PF0725</td>
<td>14.7</td>
<td>9.1</td>
<td>1</td>
<td>56.0</td>
<td>NC  1.7  1Y81</td>
<td></td>
</tr>
<tr>
<td>PF0616</td>
<td>27.0</td>
<td>5.1</td>
<td>5</td>
<td>7.8</td>
<td>6  4.1, medium</td>
<td></td>
</tr>
<tr>
<td>PF0175</td>
<td>25.5</td>
<td>5.2</td>
<td>5</td>
<td>38.2</td>
<td>No diffraction  10</td>
<td></td>
</tr>
<tr>
<td>PF1187</td>
<td>39.0</td>
<td>4.9</td>
<td>1</td>
<td>1) 4.1;2) 8.5</td>
<td>NC  3.6, high</td>
<td></td>
</tr>
</tbody>
</table>

MW, molecular weight; S.S, Se-Met substitution; A.P, additional purification; R.M, reductive methylation; M.M, modified microbatch method; MO, mosaicity; NC, no crystals; NA, not available.

*, structures were solved after Se-Met substitution and additional salvaging work (reductive methylation) from the other groups of the laboratory.
Table 3.1b Salvaging efforts for 14 *Pyrococcus furiosus* proteins failed in the Se-Met form

<table>
<thead>
<tr>
<th>Original protein</th>
<th>MW (KDa)</th>
<th>pI</th>
<th>Expression volume</th>
<th>Salvaging procedures (Final amounts, mg)</th>
<th>Crystals (Å)</th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF1183-Se</td>
<td>14.4</td>
<td>9.1</td>
<td>1L</td>
<td>a. S.S 5.0</td>
<td>poor diffraction</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF0385-Se*</td>
<td>12.9</td>
<td>4.7</td>
<td>1L</td>
<td>1) 26.3 2) x</td>
<td>NC</td>
<td>1) NC; 2) x</td>
<td></td>
</tr>
<tr>
<td>PF0492-Se*</td>
<td>10.9</td>
<td>5.3</td>
<td>1L</td>
<td>1) 6.3 2) x</td>
<td>NC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF1025-Se</td>
<td>40.2</td>
<td>7.5</td>
<td>1L</td>
<td>1.9</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF0619-Se</td>
<td>15.1</td>
<td>5.3</td>
<td>1L</td>
<td>3.6**</td>
<td>poor diffraction</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>PF0280-Se</td>
<td>12.1</td>
<td>4.4</td>
<td></td>
<td>x</td>
<td>poor diffraction</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF1243-Se</td>
<td>11.2</td>
<td>4.7</td>
<td></td>
<td>x</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF1962-Se</td>
<td>14.6</td>
<td>8.5</td>
<td></td>
<td>x</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF1372-Se</td>
<td>9.0</td>
<td>4.7</td>
<td></td>
<td>x</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF0653-Se</td>
<td>20.4</td>
<td>9.3</td>
<td></td>
<td>x</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF1261-Se</td>
<td>21.8</td>
<td>9.6</td>
<td></td>
<td>x</td>
<td>NA</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>PF1174-Se*</td>
<td>15.7</td>
<td>4.3</td>
<td></td>
<td>1) x 2) x</td>
<td>8.0</td>
<td>1) x; 2) 3.5</td>
<td></td>
</tr>
<tr>
<td>PF1044-Se*</td>
<td>17.2</td>
<td>4.8</td>
<td></td>
<td>1) x 2) x</td>
<td>10</td>
<td>1) x; 2) 3.5</td>
<td></td>
</tr>
<tr>
<td>PF0421-Se*</td>
<td>44.1</td>
<td>6.0</td>
<td></td>
<td>1) x 2) x</td>
<td>7.8</td>
<td>1) x; 2) 5.0</td>
<td></td>
</tr>
</tbody>
</table>

MW, molecular weight; S.S, Se-Met substitution; A.P, additional purification; R.M, reductive methylation; M.M, modified microbatch method; MO, mosaicity; NC, no crystals; NA, not available.

*, the reductive methylation and the modified microbatch or additional purification were performed separately.

**, in addition of ion exchange chromatography, hydroxyapatite chromatography was also performed.
### Table 3.2 Statistics from the crystallographic analysis of PF0907

**Crystallographic Conditions**
1.6M magnesium sulfate, 0.1M MES, pH 6.5, modified microbatch, temperature 293K

**Crystallization Conditions**
- pfu80080-001

**Cell dimensions**
- Space group: P6<sub>5</sub>2<sub>2</sub>2
- a, b, c (Å): 88.63, 88.63, 73.27
- α, β, γ (°): 90, 90, 120

**Data collection**
- Source: APS 22-ID
- Detector: MarCCD225
- Wavelength (Å): 1.000
- Distance (mm): 210
- 2θ (°): 0.0
- Phi step (°): 0.5
- Total Rotation (°): 360

**Data processing**
- Resolution range (Å)*: 76.70 – 2.00 (2.07-2.00)
- Reflections: 11,239
- Completeness (%)*: 93.7 (57.4)
- R<sub>sym</sub>(%)*: 6.8 (44.3)
- <I/σ(I)>*: Null

**Structural refinement**
- Program: REFMAC 5
- Resolution range (Å)*: 76.70 – 2.00 (2.05-2.00)
- Completeness (%)*: 93.6 (51.7)
- R<sub>work</sub>(%)*: 22.1 (33.3)
- R<sub>free</sub>(%)*: 24.0 (50.0)

**R.M.S. Deviations from ideality**
- Bond lengths (Å): 0.013
- Bond angles (°): 1.465
- Average B factor (Å<sup>2</sup>): 41.43

**Ramachandran analysis**
- Most favored (%): 100.0 (100.0)
- Disallowed (%): 0.0

**Final model**
- Residues: 18-108
- Solvent atoms: 28
- Other molecules: 1
- PDB ID: 1XE1

---

a, R<sub>sym</sub> = \[ \sum |I - \langle I \rangle| / \sum I \], where I is the observed intensity of reflections;
b, R<sub>work-free</sub> = \[ \sum |F_{\text{obs}} - F_{\text{calc}}| / F_{\text{obs}} \], where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.

* outer shell values in parentheses

** MolProbity percent in all allowed regions shown in parentheses
Table 3.3 Results of DaliLite Pairwise* comparison of the structures of PF0907 and its homologues

<table>
<thead>
<tr>
<th>Source</th>
<th>Protein</th>
<th>PDB ID</th>
<th>%id</th>
<th>lali</th>
<th>r.m.s.d</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Sulfolobus solfataricus</em></td>
<td>Elongation factor (EF) 1α</td>
<td>1SKQ</td>
<td>33</td>
<td>80</td>
<td>1.7Å</td>
<td>13.1</td>
</tr>
<tr>
<td><em>Methanococcus maripaludis</em></td>
<td>Elongation factor SelB</td>
<td>1WB1</td>
<td>29</td>
<td>79</td>
<td>1.6Å</td>
<td>11.7</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>EF-Tu</td>
<td>1EFC</td>
<td>27</td>
<td>79</td>
<td>1.9Å</td>
<td>11.4</td>
</tr>
<tr>
<td><em>Thermus aquaticus</em></td>
<td>EF-Tu</td>
<td>1TTT</td>
<td>29</td>
<td>79</td>
<td>2.0Å</td>
<td>10.0</td>
</tr>
<tr>
<td><em>Thermus thermophilus</em></td>
<td>EF-G</td>
<td>1FNM</td>
<td>22</td>
<td>77</td>
<td>2.7Å</td>
<td>9.2</td>
</tr>
<tr>
<td><em>Methanobacterium thermoautotrophicum</em></td>
<td>IF2/eIF5B domain II</td>
<td>1G7S</td>
<td>17</td>
<td>81</td>
<td>2.3Å</td>
<td>11.0</td>
</tr>
<tr>
<td><em>Methanobacterium thermoautotrophicum</em></td>
<td>IF2/eIF5B domain IV</td>
<td>1G7S</td>
<td>15</td>
<td>84</td>
<td>2.7Å</td>
<td>11.3</td>
</tr>
<tr>
<td><em>Bacillus stearothermophilus</em></td>
<td>IF2</td>
<td>1D1N</td>
<td>24</td>
<td>85</td>
<td>2.5Å</td>
<td>9.6</td>
</tr>
</tbody>
</table>

lali, number of structurally equivalent residues;
%id, percentage of identical amino acids over all structurally equivalent residues;
r.m.s.d., root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.

* (Holm & Park, 2000)
Table 3.4  Statistics from the crystallographic analysis of PF0725

**Crystallization Conditions**
0.3M sodium thiocyanate, 35% W/V PEG 3350, modified microbatch, temperature 291K

**Crystal**  
**pfu723267-001**

**Cell dimensions**

<table>
<thead>
<tr>
<th>Space group</th>
<th>P6₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c (Å)</td>
<td>79.15, 79.15, 36.40</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 90, 120</td>
</tr>
</tbody>
</table>

**Data collection**

<table>
<thead>
<tr>
<th>Source</th>
<th>APS 22-ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detector</td>
<td>MarCCD300</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9791</td>
</tr>
<tr>
<td>Distance (mm)</td>
<td>210</td>
</tr>
<tr>
<td>2θ (°)</td>
<td>0.0</td>
</tr>
<tr>
<td>Phi step (°)</td>
<td>0.5</td>
</tr>
<tr>
<td>Total Rotation (°)</td>
<td>180</td>
</tr>
<tr>
<td>Data processing</td>
<td>HKL 2000</td>
</tr>
<tr>
<td>Resolution range (Å)*</td>
<td>68.52-1.59 (1.65-1.59)</td>
</tr>
<tr>
<td>Reflections</td>
<td>14,818</td>
</tr>
<tr>
<td>Completeness (%)*</td>
<td>83.1 (20.1)</td>
</tr>
<tr>
<td>R_sym(%)*</td>
<td>5.4 (37.1)</td>
</tr>
<tr>
<td>&lt;I/σI&gt;*</td>
<td>Null</td>
</tr>
</tbody>
</table>

**Refinement**

<table>
<thead>
<tr>
<th>Program</th>
<th>REFMAC 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)*</td>
<td>68.52-1.70 (1.75-1.70)</td>
</tr>
<tr>
<td>Completeness (%)*</td>
<td>94.9 (64.3)</td>
</tr>
<tr>
<td>R_work(%)b*</td>
<td>22.3 (34.0)</td>
</tr>
<tr>
<td>R_free(%)b*</td>
<td>24.5 (44.2)</td>
</tr>
</tbody>
</table>

**R.M.S. Deviations from ideality**

| Bond lengths (Å) | 0.015 |
| Bond angles (°) | 1.325 |
| Average B factor (Å²) | 24.27 |

**Ramachandran analysis**

| Most favored (%) | 98.2 (100.0) |
| Disallowed (%) | 0.0 |

**Final model**

| Residues | 6-121 |
| Solvent atoms | 23 |
| Other molecules | 10 |
| PDB ID | 1Y81 |

a, R_sym = Σ|I - <I>|/ΣI, where I is the observed intensity of reflections;  
b, R_work-free = Σ|F_obs - F_calc|/ F_obs, where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.  
* outer shell values in parentheses  
** MolProbity percent in all allowed regions shown in parentheses
Table 3.5 Results of structure similarity screening of PF0725 using the DALI server*

<table>
<thead>
<tr>
<th>Source</th>
<th>Protein</th>
<th>PDB ID</th>
<th>%id</th>
<th>lali</th>
<th>r.m.s.d</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Thermus thermophilus</em></td>
<td>TT1466</td>
<td>IUUK</td>
<td>29</td>
<td>116</td>
<td>1.6Å</td>
<td>17.5</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>Succinyl-CoA synthetase</td>
<td>1EUD</td>
<td>18</td>
<td>110</td>
<td>2.3 Å</td>
<td>12.2</td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em> l2</td>
<td>Tartronic semialdehyde reductase</td>
<td>1YB4</td>
<td>12</td>
<td>108</td>
<td>2.5Å</td>
<td>10.9</td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em></td>
<td>Tartronate semialdehyde reductase</td>
<td>1VPD</td>
<td>21</td>
<td>109</td>
<td>2.5Å</td>
<td>10.3</td>
</tr>
<tr>
<td><em>Pseudomonas fragi</em></td>
<td>Fatty oxidation complex α subunit</td>
<td>1WDL</td>
<td>21</td>
<td>106</td>
<td>2.7Å</td>
<td>10.1</td>
</tr>
<tr>
<td><em>Magnaporthe grisea</em></td>
<td>Saccharopine reductase</td>
<td>1E5L</td>
<td>12</td>
<td>108</td>
<td>2.6Å</td>
<td>10.0</td>
</tr>
<tr>
<td><em>Methanococcus jannaschii</em></td>
<td>Trk system potassium uptake protein trkA homolog</td>
<td>1LSS</td>
<td>16</td>
<td>98</td>
<td>2.9Å</td>
<td>10.0</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>Putative potassium channel protein</td>
<td>1ID1</td>
<td>10</td>
<td>97</td>
<td>2.6Å</td>
<td>9.8</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>yuaA</td>
<td>1LSU</td>
<td>18</td>
<td>102</td>
<td>2.9Å</td>
<td>9.6</td>
</tr>
</tbody>
</table>

lali, number of structurally equivalent residues;  
%id, percentage of identical amino acids over all structurally equivalent residues;  
r.m.s.d., root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.

* (Holm & Sander, 1996)
CHAPTER FOUR

Treating these salvaging/rescue procedures as part of the UGA high-throughput pipeline

As discussed in Chapter Three, salvaging methods successfully rescued some targets that failed in the first trials of the original high-throughput pipeline. These methods were incorporated into the UGA high-throughput pipeline as Tier 3 for new targets. In addition, studies have shown that His$_6$ tag may affect crystallization and crystal diffraction quality (Kimber et al., 2003). To remove the His$_6$ tag, protease cleavage sites were inserted between the His$_6$ tag and the target protein for tag removal using thrombin or TEV protease (2.3.4).

4.1 Application of the improved high-throughput pipeline to 37 new protein targets

4.1.1 Purification of 37 new targets

A set of 31 Clostridium thermocellum targets and six Archaeoglobus fulgidus targets were selected. Thirty-seven Se-Met substituted proteins (Table 4.1) were expressed in PASM5052. Among them, 5 L working volume fermentors of the UGA Bioexpression and Fermentation Facility were used for six targets. The automatic two-step method and ÄKTApurifier were applied for 10 targets. ÄKTApurifier was also used for purifying up to seven proteins at one time by Ni affinity chromatography alone. Fourteen proteins were purified by this method, followed by gel filtration chromatography and/or tag removal. Tag removal of some proteins was performed as described in 2.3.4. In total, tags of six proteins were removed by thrombin or TEV
protease cleavage. Also, Cth393 was purified in two forms, with and without His₆ tag. Ion exchange chromatography was performed for three proteins (AF2094, AF0160 and AF0789). In addition, the reductive methylation of surface lysine residues was performed for two targets (AF2094, AF0160). Two targets (Cth1386, Cth2555) were also purified in the native form.

The distribution of protein amounts is shown in Figure 4.1. The amounts of purified proteins ranged from 1.01 mg to 213.32 mg, with an average amount of 36.85 mg. Again, two or more peaks of the same protein were counted separately. Their concentrations ranged from 1.84 mg/mL to 154.21 mg/mL, with an average concentration of 36.39 mg/mL. Purified proteins were submitted for crystallization trials and structure determination. Cth1550 was not handed because of its low concentration and amount. Cth2607 failed in purification. Therefore, crystallization trials were performed for 35 proteins.

4.1.2 Structures of five targets were solved

As a result, 14 of 35 proteins gave hits in crystallization trials (Table 4.1). The success rate from the purified protein to the protein crystal was 40%, which was a little higher than the average success rate of PSI centers (36%) as of August 2004. Among them, five structures were solved. The success rate from the protein crystal to the structure was 36%, and the total success rate from the purified protein to the structure was 14%. These rates were dramatically higher than corresponding average success rates of PSI centers (19% from the protein crystal to the structure and 7% from the purified protein to the structure) by August 2004.
Cth393 proteins both with and without His$_6$ tag obtained crystals. However, only crystals with the tags were good enough for structure determination, perhaps because the concentration of the sample without tags (25 mg/ml) was lower than that of the sample with tags (57 mg/ml). It was later shown in the structure that Cth393 formed dimers under the experimental conditions. Since proteins of the dimer in the crystal asymmetric unit contact through the N-terminals of the proteins, it is also possible that the additional amino acids at the N-terminal helped to stabilize the dimer, which in turn improved the diffraction quality of the crystals. Furthermore, some other researchers have reported that some proteins yield a crystal only with the His$_6$ tag, and vice versa (Kimber et al., 2003). Therefore, the His$_6$ tag may have different impacts on the crystallization and the crystal diffraction quality of different proteins, depending on the folding property of individual proteins. For 29 proteins purified with His$_6$ tag, 12 (41%) gave hits in crystallization trials and four structures (Cth393, Cth2968, Cth95 and Cth383) were solved. In contrast, three of six proteins (50%) purified without His$_6$ tag obtained crystals during the crystallization screening. Among them, one structure (AF0160) was solved. The differences indicated that the crystallization chance of proteins without His$_6$ tags was a little higher than those with the tags. Once protein crystals were obtained, the chances of solving structure were similar.

For three proteins that were further purified by ion exchange chromatography, crystals of one protein (AF0160) were obtained and the structure solved.

The reductively methylated AF0160 and AF2094 did not yield any crystals, although the structure of AF0160 was solved for Se-Met substituted proteins.
4.2 Structure determination of AF0160 through the high-throughput pipeline

AF0160 is composed of 174 amino acids with calculated pl of 5.99. Its gene was cloned into pDEST527 with sequence encoding TEV cleavage site at its 5’ end (4.1.1). AF0160 is a good example to explain the work of the Tier 3 of the UGA high-throughput pipeline because methods used in the Tier 3 (e.g., tag removal, additional ion-exchange chromatography, crystal salvaging, and reductive methylation) were all performed on it.

4.2.1 Purification and the reductive methylation of Se-Met substituted AF0160

As described in 2.2.2, cells were grown in 1L PASM5052. The amount of collected biomass was 4.20 g. The expression of AF0160 was confirmed by SDS-PAGE (Figure 4.2A). The protein was first purified by Ni affinity chromatography as described in 2.3.1. Three peaks were obtained (Figure 4.2B). Fractions corresponding to peaks 2 and 3 contained the target proteins. His$_{6}$ tag was removed by TEV protease digestion and Ni affinity chromatography as described in 2.3.4 and 2.3.1. The protein was further purified by ion exchange chromatography as described in 2.3.2 (Figure 4.2C). The protein sample was then concentrated and applied to the gel filtration column to obtain homogenous protein as described in 2.3.1. The final concentration of AF0160 was 81.3 mg/ml in a total volume of 700ul.

Reductive methylation was performed for 150 ul AF0160 as described in 2.4. The protein sample was then purified by ion exchange chromatography as described in 2.3.2. The final concentration was 23.5 mg/ml and the volume was 300 ul.

4.2.2 Crystallization of AF0160

The initial screenings of Se-Met substituted AF0160 and its methylated counterpart were performed against 384 reagent mixtures by a Cartesian Honeybee
crystallization robot (Genomic Solutions) as described in 2.5. Crystals (Figure 4.3A) were found in several conditions, including Crystal Screen II-36 (Hampton Research; 0.1M HEPES pH7.5 and 4.3M Sodium chloride); PEG/Ion -7, 13 and 14 (Hampton Research; 0.2 M Calcium chloride dihydrate and 20% w/v Polyethylene glycol 3,350 pH5.1; 0.2 M Sodium thiocyanate and 20% w/v Polyethylene glycol 3,350 pH6.9; 0.2 M Potassium thiocyanate and 20% w/v Polyethylene glycol 3,350 pH7.0); and Wizard II-18 (Emerald BioSystems; 20% w/v Polyethylene glycol 3,000, Tris pH 7.0 and 0.2M Calcium acetate dihydrate ). These conditions contained either high concentrate salt or 20% (w/v) PEG ~3,000 with 0.2M salt.

As described in 2.5, crystallization conditions were optimized using a single- or double-grid screen (pH and/or precipitant) centered on the lead conditions with different protein concentrations (81.3 mg/mL, 40.7 mg/mL and 20.4 mg/mL) at two temperatures (18°C and 4°C). Both the modified microbatch under oil method (for PEG/Ion-13) and the sitting-drop vapor-diffusion method (for the other conditions) were used. Better crystals were obtained in conditions PEG/Ion-13-9 (0.15 M Sodium thiocyanate and 20% w/v Polyethylene glycol 3,350 pH6.9), -10 (0.15 M Sodium thiocyanate and 25% w/v Polyethylene glycol 3,350 pH6.9) and -11 (0.15 M Sodium thiocyanate and 30% w/v Polyethylene glycol 3,350 pH6.9) with 81.3 mg/mL or 40.7 mg/mL proteins at 18°C. Additives (from Hampton Research Additive Screens 1 and 2) were then added to these conditions to improve crystal quality. The best crystal was obtained in PEG/Ion-13-11 with 40.7 mg/mL proteins at 18°C when a linker, 3% w/v 6-Aminohexanoic acid, was used (PEG/Ion-13-11-AD35; Figure 4.3B).
4.2.3 Data collection and processing of AF0160

Ninety-five crystals obtained from the initial screening to the final optimization were mounted, flash-frozen without cryoprotective solutions and screened as described in 2.6. Fifty-four of them were obtained from additive screen. Ten data sets of seven crystals were collected at the selenium absorption edge using the 22ID beamline of SER-CAT as described in 2.6 during three trips to SER-CAT. The scan range of all data sets was 360°. The crystal-to-detector distances ranged from 310 mm to 350 mm. The oscillation angle was 0.5° or 1° with 1 second or 2 second exposure time. Each data set was indexed, integrated and scaled using the HKL2000 suite. Output .sca files were generated for the next step. The crystals belonged to space group P212121, and unit cell dimensions were a=65Å, b=70Å, and c=134Å. The R_{sym} of the data set 0627-1_4_1, which was used for phase calculation in the next step, was 14.8% with the highest resolution 2.79Å. For the data set 0815-6_2_1 that was used for refinement, the R_{sym} was 10.6% and the highest resolution was 2.60Å.

4.2.4 Phase solution, refinement and validation of AF0160

The Matthews coefficient (Vm), the solvent content of the crystal, and the probability of molecules in the asymmetric unit were calculated as described in 2.7. The asymmetric unit contained more than one molecule. They were related by non-crystallographic symmetry (NCS). The probability across all resolution ranges was 68% if three molecules were contained in the asymmetric unit with solvent content=54% and Vm=2.59. It was also 26% possible for four molecules to be contained in the unit with solvent content=39% and Vm=1.94.
Values of both possibilities were used for phase calculation using the SECSG web-based SCA2Structure pipeline. The .sca file and the sequence file of AF0160 were loaded into the pipeline. The resolution increment was set to 0.2 Å. In both cases, the space group was set to P2_12_12_1, the wavelength to 0.9793, the high resolution limit to 2.79 Å, and the heavy atom to Se; and the $f''$ was 3.84 for the used wavelength. The maximum number of sites to be searched was set to 15 or 20, and the number of residues in the asymmetric unit was set to 522 or 696 (one molecular of AF0160 contains 174 amino acids and 5 Met residues). The solvent content was set to 0.54 or 0.39. (Figure 4.4A)

Reasonable phases were obtained only with the input of three molecules per asymmetric unit. As a part of the RESOLVE program integrated into the pipeline, NCS analysis results indicated the existence of a three-fold NCS (Figure 4.5A). This result was confirmed by the program PROFESS of the CCP4 suite, which helped in the identification of NCS related atoms from a list of heavy atom (anomalous scatterer) positions (Figure 4.5B).

The pipeline results of the input for three molecules of AF0160 were sorted as described in 2.7 (Figure 4.4B). The .pdb file and .mtz file of the selected result--which had phase resolution to 3.2 Å, SOLVE FOM 0.34 and Z-score 49.0--were used for refinement.

The .mtz file of the data set 0815-6_2_1 was used for refinement. During refinement, reflections were selected for free-R-value calculations (Brunger, 1992). Since NCS leads to relationships between reflections, it causes a slight model bias in the test set if the test R-free set is selected randomly. R-free may be lower than it should be leading
to improper weighting (Rees, 1983). Therefore, for AF0160 the test set was picked using a thin-shell selection algorithm, lx_DATAMAN (Kleywegt & Jones, 1996), which divided the data into resolution bins and picked a thin shell from each bin as the test reflections. After reselection of the test set, the .mtz file was named thinshell-0815.mtz. The phase from the above pipeline result was combined with the thinshell-0815.mtz file using the program CAD of the CCP4 suite. Phase improvement was performed using the CCP4i interface of DM. The initial model from the pipeline was inspected using COOT. The Cα skeleton of 302 residues (58% of the model) were auto-traced by the pipeline. The Cα skeleton of 147 residues (84%) of one molecule could be built through superposing models of other molecules. However, some were incorrect, and the side chains needed a manual fit. The positions of Se-Met were located according to the positions of the anomalous scatterer, Se. The side chains of residues around the Se-Met were fit into the electron density map after density modification. As shown in Figure 4.6A, the electron density map of the region between Met60 and Met72 (M_{60}PQSLAEVYESVM_{72}) indicated that the lower Se atom (of Met60) belonged to the right-side residue. In that case, however, it would be too difficult to fit 11 residues in the area between these two Se-Met positions. Therefore, this site was left unfit at that time. After the model of one molecule had been rebuilt, the models of the other two molecules were obtained using the program PDBSET of the CCP4 suite with the rotation matrix identified by LSQKAB of the CCP4 suite. The model was refined by REFMAC. During the refinement, the NCS-related molecules were kept restrained. The relative weighting of the X-ray and geometry terms was set to 0.05. Automated model (re-)building was unsuccessful using ARP/wARP6.1.1 at the beginning. However, ARP/wARP could be
used to improve the density map and guide the model building. After several rounds of manual rebuilding and refinement, it was clear that the Se mentioned above belonged to the underneath residue (Figure 4.6B) and the side chains were fit. By that time, ARP/wARP could automatically rebuild part of the model, although the tracing of some residues was incorrect. More rounds of model rebuilding and refinement were performed until most residues were fit, except for residues 161-174 of chain A, 163-174 of chain B, and 110 and 165-174 of chain C, which were not observed in the electron density. Multi-group translation/libration/screw (TLS) refinement was then performed using REFMAC.

Models based on TLS rigid-body motion of large groups of atoms provided information about intrinsic dynamic properties of the protein (Painter & Merritt, 2006a). The optimized TLS groups of AF0160 were calculated using TLS Motion Determination (TLSMD) through its web server (Painter & Merritt, 2006b). The TLSIN input file and a modified PDB file were generated by the TLSMD web server and downloaded for refinement. The water molecules were picked first, using ARP/wARP. The selection criteria include a B-factor smaller than 45Å$^2$ and a peak height greater than 3.2σ. The water molecules were then inspected and adjusted using COOT. The final model was validated using MolProbity and PROCHECK (93). The $R_{\text{work}}$ and $R_{\text{free}}$ values of the final model were 24.7% and 29.7% respectively. In the Ramachandran Plot, 90.9% of 451 non-glycine and non-proline residues fell in the most favored regions and 9.1% in the additional allowed regions. The data collection and model refinement statistics are shown in Table 4.2.
4.2.5 The Structure of AF0160

The structure of AF0160 displays an all-helical architecture that is mainly composed of a bundle of 10 helices (H1-H10) (Figure 4.7A). Two long helices, H7 (residues 90-111) and H8 (residues 114-130), are connected by a turn of three amino acids (QMK) with two helices, the N-terminal H1 (residues 2-20) and the C-terminal H10 (residues 143-160), on one side and two other helices, H5 (residues 62-65) and H6 (residues 68-78), on the other side. H9 bends toward H10 at His131. H5 bends toward H5 of other molecules in the same asymmetric unit at Glu66 and Val67. H2 (residues 27-34) is perpendicular to H8. The interactions between these helices are dominated by hydrophobic interactions of interface residues (Leu, Isl, Val, and Phe) in the helices. In addition, the networks of polar interactions hold the ends of helices (H1, 5, 7, 8, 10) together. Hydrogen bonds and salt bridges among Tyr9, Asp155, Arg5, Gln124, Glu105 and Arg121 hold the end of H1 and H10 together and attach them to H7 and H8 (Figure 4.7B). For H6, Arg104 from H7 forms hydrogen bonds with the carboxylate groups of Gly80 and Tyr79 at the C-terminal of H6. In turn, the hydroxyl group of Tyr79 interacts with NH2 of Arg126 and the carbonyl oxygen of Lys119 (both from H8) through hydrogen bond and water-mediated hydrogen bonds respectively. Tyr68 forms hydrogen bonds with His131, which in turn interacts with Arg126 and the carboxylate group of Asp59 from another molecule (Chain B) through water-mediated hydrogen bonds (Figure 4.7C). In the same asymmetric unit of three molecules, molecule A (Chain A) interacts with molecule B (Chain B) through H5. Moreover, Asp59 from the region between H4 and H5 of molecule B interacts with H5 of molecule A through water mediated hydrogen bonds. Molecule A interacts with molecule C (Chain C) through the regions between H2
and H5 (containing H3 and H4). Molecule C interacts with H6 of molecule B through its N-terminal region of H5 (Figure 4.7D).

4.3 Function prediction of five solved structures

4.3.1 AF0160, a putative tetrathionate reductase maturation protein

The amino acid sequence similarity search against the non-redundant protein database using BLAST revealed that AF0160 is similar (more than 30% identical amino acids with over 80% of total amino acids) to a group of proteins conserved in archaea, the functions of which are still unknown. No other structure of this family was found by searching Pfam, a database of protein families (http://pfam.janelia.org/) (Finn et al., 2006). The SUPERFAMILY web server was used to classify AF0160 into the superfamily of the TorD-like protein with E-value 1.9e^-7 (E-value greater than 0.0001 means weak HMM library classifications). Two structures were found that belong to the TorD-like family (PDB ID 1N1C and 1S9U) in SCOP1.71. Common results of the structural similarity search using the DALI and VAST server revealed that the structure of AF0160 is structurally homologous to three TorD-like proteins from Salmonella typhimurium (PDB ID 1S9U, r.m.s.d. 2.9Å), Shewanella massilia (PDB ID 1N1C, r.m.s.d. 3.0Å) and Archaeoglobus fulgidus (PDB ID 2O9X, r.m.s.d. 3.4Å) (Table 4.3, Figure 4.8A).

The result of MicrobesOnline Operon Prediction indicated that AF0160 belongs to an operon composed of five genes--AF0157, AF0158, AF0159, AF0160 and AF0161. The amino acid sequence similarity search using BLAST revealed that: (a) the protein of AF0157 is similar to the iron-sulfur binding β-subunit (TrtB) of tetrathionate reductase complex from Salmonella typhimurium (sequence identity 43%, over 95% of AF0157);
(b) the protein of AF0158, a predicted membrane protein, has a remote relationship with the integral membrane γ-subunit (TtrC) of the complex, identified by three iterations of PSI-BLAST search (sequence identity 16%, over 92% of AF0158, with E-value $1e^{-5}$); (c) the protein of AF0159 is closely related to the α-subunit (TtrA) of the complex (sequence identity 30%, over 94% of AF0159) containing a molybdopterin guanine dinucleotide cofactor (MGD) active site; and (d) the protein of AF0161 is similar to proteins of MoeA family (cd00887, E-value $5e^{-88}$; COG0303, E-value $6e^{-87}$), which are involved in biosynthesis of the molybdenum cofactor, the molybdopterin that contains a tricyclic pyranopterin and is required for the activation of the complex. Tetrathionate reductase is a twin-arginine translocation (Tat) and molybdenum-dependent oxidoreductase (redox enzyme). TorD-like proteins belong to the redox enzyme maturation proteins (REMPs), which are involved in the maturation and Tat signal peptide binding of α-subunit of Tat-dependent redox enzyme complexes prior to their translocation into the periplasm by Tat translocase. Genes of some TorD-like proteins (e.g., DmsD and TorD) are contained in operons of the complexes’ genes (Tranier et al., 2003). AF0160 also has significant signature sequences of TorD-like proteins. These sequences include enriched Asp and (or) Glu residues over a 10 amino acids (3 Glu) and a 20 amino acids (5 Asp/Glu) span within 60 residues from the C terminus, and conserved Asp and His in the featured ExxDH motif (Figure 4.8B). Together with the information of SUPERFAMILY classification and DALI structural similarity search, AF0160 might be a protein involved in the maturation of a putative tetrathionate reductase complex α-subunit from *Archaeoglobus fulgidus*, AF0159.
However, the operon of the tetrathionate reductase complex from *Salmonella typhimurium LT2* does not contain the gene of a TorD-like protein, although the translocation of TtrA is Tat-dependent (Turner *et al.*, 2004, Hensel *et al.*, 1999). The classification of AF0160 to a TorD-like protein and the appearance of its gene in the operon of the putative tetrathionate reductase complex revealed that the translocation of tetrathionate reductase TtrA might also need REMPs. AF0160 is the first recognized putative REMP of tetrathionate reductase and is named TtrD for the sake of convenience.

The gene arrangement in the operon containing AF0160 is *ttr*B-C-A-D, which is also found in the operon of the putative tetrathionate reductase complex from *Aeropyrum pernix K1* and is similar to that from *Salmonella typhimurium LT2* (*ttr*BCA). This kind of gene arrangement is different from that of the three main groups of TorD-like proteins and their redox enzymes, such as DmsD group (ABCD), TorD group (CAD), and NarJ group (ABDC) (Turner *et al.*, 2004). Here, A corresponds to the gene of the α-subunit of the corresponding redox enzyme, B to the gene of the β-subunit, C to the gene of the γ-subunit, and D to the gene of the REMPs. In addition, the structure-based alignment (Figure 4.8B) of TtrD (AF0160) with the three main groups of TorD-like proteins revealed that TtrD does not have signature sequences specific to the three groups, although several residues are conserved (Figure 4.8B). Therefore, TtrD (AF0160) belongs to a new group of TorD-like proteins, other than DmsD, TorD and NarJ.

**4.3.2 Cth393, a histidine triad (HIT) protein with putative functions as hydrolases and transferases**

Statistics from the crystallographic analysis of Cth393 are shown in Table 4.4. The amino acid sequence similarity search against the PDB database using BLAST
revealed that Cth393 is similar to some HIT family proteins from rabbit (Hint, PDB ID 1RZY; sequence identity 41%, over 91% of total amino acids of Cth393), human (protein kinase C interacting protein 1 (PKCI-1), PDB ID 1KPF; sequence identity 42%, over 92% of total amino acids of Cth393) and *Bacillus subtilis* (HIT family protein, PDB ID 1Y23; sequence identity 40%, over 92% of total amino acids of Cth393). The structural similarity search using the DALI server confirmed that the structure of Cth393 is structurally homologous to the above three proteins (Table 4.5). In *Clostridium thermocellum*, no genes were found to form an operon with Cth393 using MicrobesOnline Operon Prediction.

The members of the HIT protein family have a conserved triad of histidine (His-x- His-x-His; His-Leu-His-Tyr-His for Cth393) sequence motif near the carboxy terminal of the proteins (Lima et al., 1997). HIT proteins work as hydrolases and transferases with the amino group or nucleotide-linked nucleoside monophosphate group substrates (Krakowiak et al., 2004). The HIT protein family is conserved throughout Archae, Prokaryae, and Eukaryae, suggesting a fundamental role of these proteins (Lima et al., 1996). Like other HIT proteins, Cth393 contains a five-stranded anti-parallel β sheet with one helix on each side of the sheet (Figure 4.9A), known as α + β meander fold (Lima et al., 1997). The structure-based alignment (Figure 4.9C) revealed that several residues (e.g., His41, Gly96, His101, His103 and His104 of Cth393) are conserved in Cth393 and other HIT proteins. These residues are involved in the interaction of some HIT proteins with their substrates. Although Thr98 corresponding to Ser107 of rabbit Hint protein is not conserved, it may still perform the same function as Ser because of their similar properties. In the structure of a HIT protein from *Bacillus subtilis* (1y23), Cys7, Cys10,
His49, and His100 (corresponding to Cys7, Cys10, His49 and His101 of Cth393) bind with zinc ion. Without zinc ion, Cys7 and Cys10 of Cth393 form a disulfide bond (Figure 4.9B). However, in the rabbit Hint protein and the human PKCI-1, only His residues are conserved. Corresponding to Cys7 of Cth393 and 1y23, Thr17 of the rabbit Hint protein or human PKCI-1 may be also involved in the interaction with zinc ion. The conserved folding and residues in Cth393 and other HIT proteins indicate that Cth393 may have the similar function of hydrolases and transferases.

4.3.3 Cth2968, an YjgF/YER057c/UK114 family protein with putative function in Cys/Met metabolism

Statistics from the crystallographic analysis of Cth2968 are shown in Table 4.6. The amino acid sequence similarity search against the PDB database using BLAST and the structural similarity search using the DALI server revealed that Cth2968 is similar to some YjgF/YER057c/UK114 family proteins from archaea, bacteria and eukarya with main-chain r.m.s.d from 0.7Å to 1.6Å after superposition (Table 4.7). All these proteins are composed of 120–130 amino acids. Cth2968 also contains the C-terminal signature sequence, [PA]-[ASTPV]-R-[SACVF]-x-[LIVMFY]-x(2)-[GSARK]-x-[LMVA]-x(5,8)-[LIVM]-E-[MI] (PARSCVEVSKLPKGLIEI for Cth2968), of this family (Figure 4.10C) (Deaconescu et al., 2002).

In Clostridium thermocellum, two other genes (Cth2969 and Cth2970) were found to belong to the same operon as Cth2968 by MicrobesOnline Operon Prediction. The protein of Cth2969 is similar to cystathionine γ-synthase (CGS), which is involved in cysteine and methionine metabolism (i.e., catalysing the pyridoxal 5'-phosphate (PLP)-dependent γ-replacement of O-succinyl-L-homoserine and L-cysteine to L-cystathionine)
The protein of Cth2970 is similar to the cystalysin from *Treponema denticola*, a C⁰-S⁷ lyase catabolyzing L-cysteine to pyruvate, ammonia and H₂S (sequence identity 37%, over 99% of Cth2970) (Krupka *et al.*, 2000) and MalY, a bifunctional PLP-dependent enzyme with β-cystathionase and maltose regułon repressor activities (sequence identity 34 over 98% of Cth2970) (Zdych *et al.*, 1995). As a β-cystathionase, MalY can cleave cystathionine to homocysteine, ammonia, and pyruvate *in vitro*—an essential step in the biosynthesis of methionine (Zdych *et al.*, 1995). Therefore, Cth2968 together with Cth2969 and Cth2970 could be involved in the cysteine and/or methionine metabolism.

Like other members of the YjgF/YER057c/UK114 family, Cth2968 is composed of a six-stranded β sheet of mixed polarity and two α-helices packed against one side of the sheet (Figure 4.10A). In each asymmetric unit, three proteins form a homo-trimer. Three β sheets (one per subunit) form the triangular barrel of the trimmer, which is surrounded by six α-helices (two per subunit) (Figure 4.10A). Three equivalent solvent-accessible clefts are located at the interfaces between adjacent subunits. These clefts are conserved in all members of the YjgF/YER057c/UK114 family and represent binding regions of substrates or ligands (Manjasetty *et al.*, 2004, Burman *et al.*, 2007). According to the structure-based sequence alignment (Figure 4.10C), Cth2968 also has seven amino acids (Tyr18, Gly32, Asn58, Asn90, Arg104, Pro113 and Glu119) conserved in all members of the YjgF/YER057c/UK114 family, except YjgH. All these residues line the walls of the clefts (Figure 4.10C). In the TdcF protein from *E.coli*, the Arg105 (corresponding to Arg104 of Cth2968) forms hydrogen bonds with all tested ligands, including serine. Glu120 (corresponding to Glu119 of Cth2968) also forms hydrogen
bonds with the serine ligand. Asn88 (corresponding to Asn90 of Cth2968) help to fix the position of Arg105 through two hydrogen bonds. Tyr17 and Pro114 (corresponding to Tyr18 and Pro113 of Cth2968) form hydrophobic interactions with the backbone of serine (Burman et al., 2007). These conserved residues in Cth2968 take positions similar to those of TdcF (Figure 4.10B). In addition, 2-ketobutyrate is the ligand of Hi0719, another member of the YjgF/YER057c/UK114 family (Parsons et al., 2003). Therefore, Cth2968 may have the ability to bind 2-ketobutyrate and serine. Moreover, cysteine may be a putative ligand of Cth2968 because of its structural similarity with serine. Cys106 may interact with the thiol group of the putative cysteine ligand, which is similar to that of PF0725, as mentioned earlier. Since Cth2969 and Cth2970 may have functions in cysteine and/or methionine metabolism, and use L-cysteine as a substrate, Cth2968 might act as a sensor of cellular L-cysteine levels.

4.3.4 Cth95, a putative translation elongation factor P (EF-P)

Statistics from the crystallographic analysis of Cth95 are shown in Table 4.8. The amino acid sequence similarity search against the PDB database using BLAST and the structural similarity search using the DALI server revealed that Cth95 is similar to EF-P from Thermus thermophilus HB8 (PDB ID 1UEB, sequence identity 44%, r.m.s.d. 4.5Å, Z-score 17.8). Detail analysis revealed that the relatively high value of r.m.s.d. is due to the difference in the relative orientation of domains II and III to domain I. When domain I alone of Cth95 and that of EF-P from T. thermophilus were compared using DaliLite Pairwise comparison of protein structures (http://www.ebi.ac.uk/DaliLite/) (Holm & Park, 2000), the r.m.s.d of the main chain was 1.1Å. For domains II and III alone, the value of
r.m.s.d was 1.9Å (Figure 4.11A). The reason for this difference was the flexible conformation of the loop between domains II and I.

Similar to the structure of the EF-P from *T. thermophilus*, Cth95 is composed of three β-barrel domains (I, II and III) which form an “L” shape structure (Figure 4.11A,B). This shape is similar to that of tRNA molecules and might be important for the protein to enter the ribosome (Hanawa-Suetsugu *et al.*, 2004). Domain I consists six β strands, and the loop region between β3 and β4 (residues 28-36) forms one end of the “L” shape. In this loop region, residues corresponding to Lys29, Gly31, Gly33, and Ala35 of Cth95 are conserved in bacteria EF-P proteins and their homologues in Archaea and Eukarya, eukaryotic initiation factor 5A (eIF5A). Corresponding to Lys32 of Cth95, this Lys residue in eIF5A is modified to hypusine. This modification does not occur for bacterial EF-P. (Hanawa-Suetsugu *et al.*, 2004) Both domains II and III are composed of five strands β-barrel. Although the electron density of some residues cannot be observed, the loop between the first (β11) and the second (β12) β strands of domain III forms the other end of the “L” shape. The asymmetric unit contains two molecules of Cth95. However, the conformations of these two monomers are different. Domain I of molecule B is closer to its domains II and III than that of molecule A. The relative difference between the orientations of domain I from the two molecules is about 40° (Figure 4.11C). This is because of the interaction between these two molecules. Domain III of molecule A interacts with both domains I and III of molecule B. OE2 of Glu24, NE2 of Gln26, and NH1 of Arg38 from domain I of molecule B form hydrogen bonds or salt bridges with the carboxyal group of Thr144, Ala145 and Gly147, as well as OG1 of Thr146 from domain III of molecule A (Figure 4.11D). NZ of Lys161 and OG of Ser159 from domain
III of molecule B interact with OE1 of Glu183 and OD2 of Asp175 from domain III of molecule A, respectively (Figure 4.11D). The flexibility of inter-domain orientations results in the flexibility of the protein shape that might be important for its interaction with other proteins or ribosome, which is true in another example of the release factor 2 that is incorporated into the ribosome through changing its inter-domain orientations (Rawat et al., 2003, Klaholz et al., 2003). The structural similarity between Cth95 and EF-P from *T. thermophilus* indicates the putative function of Cth95 in the formation of the first peptide bond of a protein during the initiation of protein synthesis (Hanawa-Suetsugu et al., 2004).

**4.3.5 Cth383, an YbaB family protein**

Statistics from the crystallographic analysis of Cth383 are shown in Table 4.9. The amino acid sequence similarity search against the PDB database using BLAST and the structural similarity search using the DALI server revealed that Cth383 is similar to the YbaB family proteins of bacteria (For *E. coli* YbaB, PDB ID 1PUG, sequence identity 36%, over 94% of total amino acids of Cth383, r.m.s.d. 1.2Å, Z-score 11.9; For YbaB from *Haemophilus Influenzae*, PDB ID 1J8B, sequence identity 33%, over 70% of total amino acids of Cth383, r.m.s.d. 2.9Å, Z-score 11.6). Like that of other YbaB family members, the gene of Cth383 is in the same operon with recR whose protein product, RecR, is required for the recovery of replication following DNA damage. RecR protects the DNA strands of the replication fork blocked by DNA damage. (Courcelle et al., 1997)

Similar to other YbaB proteins, Cth383 contains two α-helices connected by a three-strand β sheet. Two molecules of Cth383 form a dimer with a tweezer-like shape in the asymmetric unit (Figure 4.12A). According to the structure-based alignment (Figure
4.12C), the most conserved region of YbaB proteins is around the N-terminal of $\alpha_2$, including several negatively charged residues (e.g., Asp72, Glu74, and Asp78 of Cth383). These residues are also conserved in other YbaB family proteins that are not listed here (Lim et al., 2003). The Nest analysis from the ProFunc server revealed two regions (Arg55, Lys56 and Asp57, **RKD**; Gly46, Ala47, and Val48, **GAV**) where the main chain parts of successive residues are enantiomeric (main chain torsion angles alternate between the right and the left regions of the Ramachandran plot). These regions are believed to be putative cation/anion-binding regions and functionally important (Watson & Milner-White, 2002a, 2002b, Pal et al., 2002). The **RKD** region is at the loop between $\beta_2$ and $\beta_3$ located at the inner side of the “tweezer.” The **GAV** region is at the loop between $\beta_1$ and $\beta_2$ located at the outer side of the “tweezer.” (Figure 4.12B) Because the gene of Cth383 is in the same operon of *recR*, Cth383 might also be involved in DNA repair.
The proteins that had two peaks were assumed to be two different forms and were counted separately.
Figure 4.2 Purification of AF0160
A, The expression of AF0160 is confirmed by SDS-PAGE;
B, Three peaks were obtained during Ni affinity chromatography;
   The result of SDS-PAGE showed the fractions corresponding to
   peak2 and 3 contained the target protein; F, flow through; P, pellet.
C, The protein was further purified by ion exchange chromatography after
   His<sub>6</sub> tag removal
Figure 4.3 Crystallization of AF0160

A, Crystals appeared in the mother-liquor solution containing 0.2 M Sodium thiocyanate and 20% w/v Polyethylene glycol 3,350 pH6.9 at 18°C;
B, Crystals of the best qiffrcation quality appeared in the condition of 0.15 M Sodium thiocyanate, 30% w/v Polyethylene glycol 3,350 pH6.9 and 3% w/v 6-Aminohexanoic acid at 4°C
Figure 4.4  SCA2Structure pipeline input and output webpage

A, the input webpage. The .sca file and the sequence file of AF0160 were loaded into the pipeline; The maximum number of sites to search was set to 15 or 20, the number of residues in the asymmetric unit to 522 or 696, and the solvent content to 0.54 or 0.39 for 3 or 4 molecules in the asymmetric unit respectively; Other parameters were set as indicated;

B, the output webpage. Reasonable phase was obtained only with the input of 3 molecules per asymmetric unit. The pipeline results of the input for 3 molecules of AF0160 were sorted according to the number of atoms autotraced by RESOLVE, SOLVE Z score and SOLVE FOM.
Figure 4.5 Three-fold NCS of AF0160 molecules in the asymmetric unit
A, NCS analysis result of the RESOLVE program integrated into the pipeline indicates the existence of a three-fold NCS;
B, It was confirmed by the program PROFESS of CCP4 suite, which identified 3 groups of NCS related Se atoms.
Figure 4.6  Identification of the position of a Se atom at a loop region
A, the original electron density map showed that the lower Se belongs to the right-side residue; B, it was confirmed that this Se atom belongs to the underneath residue (Met60) after several rounds of manual rebuilding and refinement.
Figure 4.7 The structure of AF0160
A, the overall structure of AF0160;
B, hydrogen bond interactions and salt bridges between H1, 7, 8 and 10;
C, hydrogen bonds or water mediated hydrogen bonds between H5, 6, 7, and 8;
D, packing of 3 AF0160 molecules in the asymmetric unit
Figure 4.8 AF0160 is structurally homologous to the TorD-like proteins
A, the overlay of AF0160 (Green), 1s9u (Cyan), 1n1c (Magenta) and 2o9x (yellow);
B, the structure-based sequence alignment of AF0160 and 3 groups of TorD-like proteins; the signature sequences of the three groups are underlined.
*, indicates the conserved Asp and His in AF0160 and the TorD-like proteins.
Figure 4.9 Cth393 is a member of the HIT protein family
A, the overall structure of Cth393;
B, the overlay of Cth393 (Green) and 1rzy (purple); side chains of residues which may be involved in the substrates interaction and zinc ion binding have been shown;
C, the structure-based alignment of Cth393 and other HIT proteins; the conserved triad of histidine is underlined.
Figure 4.10 Cth2968 is a YjgF/YER057c/UK114 family protein
A, the surface and ribbon representation of the homotrimer of Cth2968; the
conserved clefts are indicated by the arrows; B, the structural overlay of Cth2968 (cyan
and purple, which represent two different molecules of Cth2968) and TdcF from E.coli
(green); side chains of conserved residues involved in ligands binding have been shown;
the ligand, Ser, of TdcF has been shown in spheres. C, structure-based alignment of
Cth2968 and some YjgF/YER057c/UK114 family proteins; the signature sequence of the
family is underlined.
Figure 4.11 Cth95 is similar to EF-P

A, the structural overlay of Cth95 (green) and the EF-P from *T. thermophilus* (cyan); the domain I (purple) and domain II & III (yellow) of the EF-P from *T. thermophilus* are also superimposed separately on Cth95;

B, the surface representation of Cth95;

C, the overlay of two Cth95 molecules (molecule A, green and molecule B, brown) in the asymmetric unit; the relative difference between the orientations of domain I from two molecules is about 40°;

D, the polar interactions between the molecule A and B;
Figure 4.12 Cth383 is an YbaB family protein
A, the ribbon representation of the dimer of Cth383 shows a tweezer-like shape;
B, the most conserved regions, GAV regions and RKD regions; the side chains of the residues in these regions have been shown;
C, the structure-based alignment of Cth383 and two other YbaB family proteins; the most conserved region is underlined.
Table 4.1 Production of 5 structures from 35 new protein targets by the improved high-throughput pipeline

<table>
<thead>
<tr>
<th>ID</th>
<th>MW (KDa)</th>
<th>pI</th>
<th>Purification</th>
<th>Tag removal</th>
<th>Amount (mg)</th>
<th>Crystallization (resolution Å, mosaicity)</th>
<th>Structure (PDB ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cth1386</td>
<td>36.7</td>
<td>6.5</td>
<td>Ni+GFL</td>
<td></td>
<td>79.0</td>
<td>no diffraction</td>
<td></td>
</tr>
<tr>
<td>Cth2555</td>
<td>40.4</td>
<td>5.4</td>
<td>Ni+GFL</td>
<td></td>
<td>81.9</td>
<td>4.7, medium</td>
<td></td>
</tr>
<tr>
<td>Cth334</td>
<td>42.1</td>
<td>5.3</td>
<td>Ni+GFL</td>
<td></td>
<td>79.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2775</td>
<td>45.1</td>
<td>6.0</td>
<td>Ni+GFL</td>
<td></td>
<td>231.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth560</td>
<td>49.2</td>
<td>5.6</td>
<td>Ni+GFL</td>
<td></td>
<td>1) 113.4; 48.4*</td>
<td>8.3, medium</td>
<td></td>
</tr>
<tr>
<td>Cth1550</td>
<td>31.4</td>
<td>5.1</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2607</td>
<td>33.1</td>
<td>4.8</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>failed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1749</td>
<td>48.8</td>
<td>4.6</td>
<td>Ni, GFL</td>
<td>x</td>
<td>20.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1211</td>
<td>9.1</td>
<td>4.6</td>
<td>Ni, GFL</td>
<td>x</td>
<td>60.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth731</td>
<td>50.5</td>
<td>6.7</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>36.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2753</td>
<td>33.7</td>
<td>4.8</td>
<td>Ni(3D), GFL</td>
<td>x</td>
<td>10.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth393</td>
<td>12.6</td>
<td>6.8</td>
<td>Ni(3D), GFL</td>
<td>1) x</td>
<td>1) 42.5; 2) 102.6**</td>
<td>1) no good crystals 2) 2.3, medium</td>
<td>2) 1XQU</td>
</tr>
<tr>
<td>Cth2968</td>
<td>13.7</td>
<td>4.9</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>58.5</td>
<td>2.2, medium</td>
<td>1XRG</td>
</tr>
<tr>
<td>Cth1677</td>
<td>10.9</td>
<td>9.9</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>18.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth25</td>
<td>10.7</td>
<td>10.</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>1) 17.8; 15.3*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1399</td>
<td>32.1</td>
<td>6.8</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>7.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth995</td>
<td>28.8</td>
<td>5.2</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>7.0</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Cth56</td>
<td>20.1</td>
<td>4.9</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>22.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1170</td>
<td>17.9</td>
<td>9.5</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>38.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth95</td>
<td>20.7</td>
<td>5.4</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>32.2</td>
<td>2.0, medium</td>
<td>1YBY</td>
</tr>
<tr>
<td>Cth2332</td>
<td>24.4</td>
<td>7.1</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>154.0</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Cth2413</td>
<td>22.0</td>
<td>4.6</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>114.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth848</td>
<td>23.4</td>
<td>6.0</td>
<td>Ni(3D), GFL</td>
<td></td>
<td>1) 30.4; 28.7*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1735</td>
<td>27.4</td>
<td>4.6</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>1) 9.0; 15.8*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2742</td>
<td>37.</td>
<td>7.8</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>20.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2120</td>
<td>44.1</td>
<td>10.</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>7.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth2915</td>
<td>19.5</td>
<td>3.9</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>1) 12.0; 15.8*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1307</td>
<td>21.1</td>
<td>9.0</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>27.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth952</td>
<td>22.6</td>
<td>4.1</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>8.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cth1923</td>
<td>17.9</td>
<td>10.</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>1) 3.8; 2) 3.0**</td>
<td>1) 3.8 high</td>
<td></td>
</tr>
<tr>
<td>Cth383</td>
<td>12.0</td>
<td>4.6</td>
<td>Ni+GFL(3D)</td>
<td></td>
<td>24.0</td>
<td>1.8, medium</td>
<td>1YBX</td>
</tr>
<tr>
<td>AF2094</td>
<td>10</td>
<td>4.5</td>
<td>Ni, GFL, IEX</td>
<td>X</td>
<td>29.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF0160</td>
<td>20.2</td>
<td>6.0</td>
<td>Ni, GFL, IEX</td>
<td>x</td>
<td>44.7</td>
<td>2.7, medium</td>
<td>2IDG</td>
</tr>
<tr>
<td>AF0453</td>
<td>14.8</td>
<td>6.6</td>
<td>Ni, GFL</td>
<td></td>
<td>4.3</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>AF5856</td>
<td>16.2</td>
<td>8.7</td>
<td>Ni, GFL</td>
<td></td>
<td>17.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF0789</td>
<td>12</td>
<td>5.7</td>
<td>Ni, GFL, IEX</td>
<td></td>
<td>1) 7.7; 1.6; 14.9*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF2346</td>
<td>11.3</td>
<td>5.9</td>
<td>Ni, GFL</td>
<td></td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*, different peaks were purified separately; **, the proteins were purified two times.
Table 4.2 Statistics from the crystallographic analysis of AF0160

**Crystallographic Conditions**
30% W/V PEG 3350, 0.15M sodium thiocyanate, 0.01M spermine-HCl, pH 6.9, modified microbatch, temperature 291K

**Crystal**

**AF0160_0815-6_2**

**Cell dimensions**

- **Space group**: P2\(_1\)2\(_1\)2\(_1\)
- **a, b, c (Å)**: 65.19, 69.74, 133.98
- **α, β, γ (°)**: 90, 90, 90

**Data collection**

- **Source**: APS 22-ID
- **Detector**: MarCCD300
- **Wavelength (Å)**: 0.9793
- **Distance (mm)**: 320
- **2θ (°)**: 0.0
- **Phi step (°)**: 1.0
- **Total Rotation (°)**: 360
- **Data processing**: HKL 2000
- **Resolution range (Å)**: 50.00 – 2.69 (2.80-2.69)
- **Reflections**: 18,724
- **Completeness (%)***: 97.0 (89.0)
- **R\(_{\text{sym}}\)(%)***: 10.6 (41.0)
- **<I/\sigma(I)>**: 28.75 (4.12)

**Refinement**

- **Program**: REFMAC 5
- **Resolution range (Å)**: 36.83 – 2.69 (2.77-2.69)
- **Completeness (%)***: 100.0 (100.0)
- **R\(_{\text{work}}\)(%)**: 24.4 (35.8)
- **R\(_{\text{free}}\)(%)**: 29.7 (53.1)

**R.M.S. Deviations from ideality**

- **Bond lengths (Å)**: 0.012
- **Bond angles (°)**: 1.361
- **Average B factor (Å\(^2\))**: 30.71

**Ramachandran analysis**

- **Most favored (%)**: 96.3 (99.2)
- **Disallowed (%)**: 0.8

**Final model:**

- **Residues**: A: 1-160; B: 0-162; C: 0-109, 111-164
- **Solvent atoms**: 81
- **Other molecules**: 0
- **PDB ID**: 2IDG

\( a, R_{\text{sym}} = \frac{\sum|I - \langle I\rangle|}{\sum I}, \) where I is the observed intensity of reflections;

\( b, R_{\text{work-free}} = \frac{\sum|F_{\text{obs}} - F_{\text{calc}}|}{F_{\text{obs}}}, \) where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.

* outer shell values in parentheses

** MolProbity percent in all allowed regions shown in parentheses
### Table 4.3 Common results of structure similarity screening of AF0160 using the DALI and VAST server***

<table>
<thead>
<tr>
<th>Source</th>
<th>Protein Description</th>
<th>PDB ID</th>
<th>%id*</th>
<th>lali*</th>
<th>r.m.s.d*</th>
<th>Z-score (score)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Salmonella typhimurium lt2</em></td>
<td>Putative component of anaerobic dehydrogenases</td>
<td>1S9U</td>
<td>18</td>
<td>133</td>
<td>2.9 Å</td>
<td>10.8 (8.3)</td>
</tr>
<tr>
<td><em>Shewanella massilia</em></td>
<td>TorA specific chaperone</td>
<td>1N1C**</td>
<td>19</td>
<td>125</td>
<td>3.0 Å</td>
<td>10.4</td>
</tr>
<tr>
<td><em>Archaeoglobus fulgidus</em></td>
<td>Reductase, assembly protein</td>
<td>2O9X</td>
<td>16</td>
<td>128</td>
<td>3.4 Å</td>
<td>8.9 (8.7)</td>
</tr>
</tbody>
</table>

lali, number of structurally equivalent residues;  
%id, percentage of identical amino acids over all structurally equivalent residues;  
r.m.s.d., root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.  
* results of VAST were shown in parentheses  
** Compared to other structures of TorD-like proteins, monomers of 1N1C adapt different conformations. However, the globular domains formed by two monomers are structurally similar to a monomer of other TorD-like proteins. (Tranier et al., 2003) The structure was reproduced and compared with AF0160 using DaliLite***  
** DALI, (Holm & Sander, 1996); VAST, (Gibrat et al., 1996); DaliLite, (Holm & Park, 2000)
Table 4.4 Statistics from the crystallographic analysis of Cth393

**Crystallization Conditions**
20% w/v PEG 3000, 0.2M sodium chloride, 0.1M HEPES, pH 7.5, vapor diffusion, sitting drop, temperature 291K

<table>
<thead>
<tr>
<th>Crystal</th>
<th>Cth-393_1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell dimensions</strong></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>P4_2_2</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>78.80, 78.80, 114.89</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 90, 90</td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>APS 22-ID</td>
</tr>
<tr>
<td>Detector</td>
<td>MarCCD300</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9690</td>
</tr>
<tr>
<td>Distance (mm)</td>
<td>300</td>
</tr>
<tr>
<td>θ (°)</td>
<td>0.0</td>
</tr>
<tr>
<td>Phi step (°)</td>
<td>0.5</td>
</tr>
<tr>
<td>Total Rotation (°)</td>
<td>180</td>
</tr>
<tr>
<td>Data processing</td>
<td>HKL 2000</td>
</tr>
<tr>
<td>Resolution range (Å)*</td>
<td>20.00-2.30 (2.40-2.30)</td>
</tr>
<tr>
<td>Reflections</td>
<td>16,265</td>
</tr>
<tr>
<td>Completeness (%)*</td>
<td>98.0 (100.0)</td>
</tr>
<tr>
<td>R&lt;sub&gt;sym&lt;/sub&gt;(%)*</td>
<td>8.0 (26.4)</td>
</tr>
<tr>
<td>&lt;I/σ(I)&gt;*</td>
<td>Null</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>REFMAC 5</td>
</tr>
<tr>
<td>Resolution range (Å)*</td>
<td>20.00–2.30 (2.36-2.30)</td>
</tr>
<tr>
<td>Completeness (%)*</td>
<td>97.4 (95.1)</td>
</tr>
<tr>
<td>R&lt;sub&gt;work&lt;/sub&gt;(%)*</td>
<td>27.8 (36.0)</td>
</tr>
<tr>
<td>R&lt;sub&gt;free&lt;/sub&gt;(%)*</td>
<td>32.6 (50.7)</td>
</tr>
<tr>
<td><strong>R.M.S. Deviations from ideality</strong></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.017</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.329</td>
</tr>
<tr>
<td>Average B factor (Å(^2))</td>
<td>34.57</td>
</tr>
<tr>
<td><strong>Ramachandran analysis</strong>**</td>
<td></td>
</tr>
<tr>
<td>Most favored (%)</td>
<td>97.3 (100.0)</td>
</tr>
<tr>
<td>Disallowed (%)</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Final model</strong></td>
<td></td>
</tr>
<tr>
<td>Residues</td>
<td>A: 4-116; B: 4-117</td>
</tr>
<tr>
<td>Solvent atoms</td>
<td>38</td>
</tr>
<tr>
<td>Other molecules</td>
<td>26</td>
</tr>
<tr>
<td>PDB ID</td>
<td>1XQU</td>
</tr>
</tbody>
</table>

\[ a, \text{R}_{\text{sym}} = \sum |I - \langle I \rangle|/\sum I, \text{ where I is the observed intensity of reflections; } \]
\[ b, \text{R}_{\text{work-free}} = \sum |F_{\text{obs}} - F_{\text{calc}}|/F_{\text{obs}}, \text{ where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5\% of the total number of reflections.} \]

* outer shell values in parentheses
** MolProbity percent in all allowed regions shown in parentheses
Table 4.5  Results of structure similarity screening of Cth393 using the DALI server*

<table>
<thead>
<tr>
<th>Source</th>
<th>Protein</th>
<th>PDB ID</th>
<th>%id</th>
<th>lali</th>
<th>r.m.s.d</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oryctolagus cuniculus</em></td>
<td>Histidine triad nucleotide-binding protein 1</td>
<td>1RZY</td>
<td>38</td>
<td>113</td>
<td>1.2Å</td>
<td>20.2</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>Protein kinase C interacting protein</td>
<td>1KPF</td>
<td>40</td>
<td>111</td>
<td>1.1Å</td>
<td>19.8</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>Histidine triad protein</td>
<td>1Y23</td>
<td>40</td>
<td>106</td>
<td>2.0Å</td>
<td>15.7</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>mRNA decapping enzyme</td>
<td>1ST4</td>
<td>12</td>
<td>106</td>
<td>2.6Å</td>
<td>11.5</td>
</tr>
</tbody>
</table>

lali, number of structurally equivalent residues;
%id, percentage of identical amino acids over all structurally equivalent residues;
r.m.s.d., root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.

* (Holm & Sander, 1996)
Table 4.6  Statistics from the crystallographic analysis of Cth2968

**Crystallization Conditions**  
30 v/v% PEG 400, 0.1M HEPES, pH 7.6, vapor diffusion, sitting drop, temperature 291K

**Crystal**  
Cth-2968_3

**Cell dimensions**
- **Space group**: P3_21
- **a, b, c (Å)**: 80.44, 80.44, 137.20
- **α, β, γ (°)**: 90, 90, 120

**Data collection**
- **Source**: APS 22-ID
- **Detector**: MarCCD300
- **Wavelength (Å)**: 0.9690
- **Distance (mm)**: 380
- **θ (°)**: 0.0
- **Phi step (°)**: 1.0
- **Total Rotation (°)**: 360
- **Data processing**: HKL 2000
- **Resolution range (Å)***: 30.00 - 2.10 (2.18 - 2.10)
- **Reflections**: 30,510
- **Completeness (%)***: 100.0 (100.0)
- **R_sym(%)**: 7.2 (16.9)
- **<I/σ(I)>**: Null

**Refinement**
- **Program**: REFMAC 5
- **Resolution range (Å)***: 19.90 - 2.20 (2.26 - 2.20)
- **Completeness (%)***: 99.6 (100.0)
- **R_work(%)**: 19.3 (18.2)
- **R_free(%)**: 22.5 (29.4)

**R.M.S. Deviations from ideality**
- **Bond lengths (Å)**: 0.016
- **Bond angles (°)**: 1.352
- **Average B factor (Å\(^2\))**: 22.27

**Ramachandran analysis**
- **Most favored (%)**: 98.5 (100.0)
- **Disallowed (%)**: 0.0

**Final model**
- **Residues**: A: 2-126; B: -11-126; C: -10-126
- **Solvent atoms**: 57
- **Other molecules**: 17
- **PDB ID**: 1XRG

\(a, R_{sym} = \sum |I - \langle I\rangle|/\sum I,\) where \(I\) is the observed intensity of reflections;
\(b, R_{work-free} = \sum |F_{obs} - F_{calc}|/ F_{obs},\) where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.

* outer shell values in parentheses

** MolProbity percent in all allowed regions shown in parentheses**
Table 4.7 Results of DaliLite Pairwise comparison of the structures of Cth2968 and its homologues

<table>
<thead>
<tr>
<th>Source</th>
<th>Protein</th>
<th>PDB ID</th>
<th>%id</th>
<th>lali</th>
<th>r.m.s.d</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrococcus horikoshii</td>
<td>Ph0854</td>
<td>2DYY</td>
<td>64</td>
<td>98</td>
<td>1.0Å</td>
<td>22.6</td>
</tr>
<tr>
<td>Aeropyrum pernix K1</td>
<td>Ape1501</td>
<td>2CWJ</td>
<td>53</td>
<td>90</td>
<td>1.0Å</td>
<td>19.8</td>
</tr>
<tr>
<td>Sulfolobus tokodaii</td>
<td>St0811</td>
<td>1X25</td>
<td>52</td>
<td>99</td>
<td>0.7Å</td>
<td>25.2</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>Tm0215</td>
<td>2B33</td>
<td>61</td>
<td>96</td>
<td>0.9Å</td>
<td>24.3</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>YabJ</td>
<td>1QD9</td>
<td>58</td>
<td>97</td>
<td>0.7Å</td>
<td>24.6</td>
</tr>
<tr>
<td>Thermus thermophilus</td>
<td>Ttha0137</td>
<td>2CSL</td>
<td>55</td>
<td>97</td>
<td>0.7Å</td>
<td>24.8</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>Hi0719</td>
<td>1J7H</td>
<td>53</td>
<td>100</td>
<td>1.5Å</td>
<td>19.7</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>TdcF</td>
<td>2UYK</td>
<td>49</td>
<td>98</td>
<td>1.1Å</td>
<td>22.9</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>YjgF</td>
<td>1QU9</td>
<td>50</td>
<td>98</td>
<td>1.1Å</td>
<td>23.0</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>YjgH</td>
<td>1PF5</td>
<td>31</td>
<td>85</td>
<td>1.6Å</td>
<td>18.2</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Yeo7</td>
<td>1JD1</td>
<td>51</td>
<td>79</td>
<td>1.0Å</td>
<td>23.3</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>P14.5</td>
<td>1ONI</td>
<td>51</td>
<td>97</td>
<td>0.7Å</td>
<td>24.2</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>Perchloric</td>
<td>1QAH</td>
<td>49</td>
<td>98</td>
<td>0.7Å</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>Acid Soluble Protein (PSP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capra hircus</td>
<td>Uk114</td>
<td>1NQ3</td>
<td>46</td>
<td>97</td>
<td>0.7Å</td>
<td>23.9</td>
</tr>
</tbody>
</table>

lali, number of structurally equivalent residues;
%id, percentage of identical amino acids over all structurally equivalent residues;
r.m.s.d., root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms
### Table 4.8 Statistics from the crystallographic analysis of Cth95

**Crystallization Conditions**
1.5M dipotassium hydrogen phosphate, 0.1M HEPES, pH 7.6, vapor diffusion, sitting drop, temperature 291K

**Crystal**  Cth-95_1

**Cell dimensions**
- Space group: C2 (C121)
- a, b, c (Å): 114.92, 55.48, 88.46
- α, β, γ (°): 90, 120, 90

**Data collection**
- Source: APS 22-ID
- Detector: MarCCD300
- Wavelength (Å): 0.9793
- Distance (mm): 230
- 2θ (°): 0.0
- Phi step (°): 1.0
- Total Rotation (°): 360
- Data processing: HKL 2000
- Resolution range (Å)*: 50.00-1.85 (1.92-1.85)
- Refinements: 39,625
- Completeness (%)*: 97.0 (94.8)
- Rsym(%)*: 8.3 (54.1)
- <I/σI>*: Null

**Refinement**
- Program: REFMAC 5
- Resolution range (Å)*: 27.73-1.95 (2.00-1.95)
- Completeness (%)*: 99.6 (100.0)
- Rwork(%)*: 22.8 (25.2)
- Rfree(%)*: 27.0 (28.9)

**R.M.S. Deviations from ideality**
- Bond lengths (Å): 0.015
- Bond angles (°): 1.408
- Average B factor (Å²): 22.50

**Ramachandran analysis**
- Most favored (%): 96.6 (99.1)
- Disallowed (%): 0.9

**Final model**
- Solvent atoms: 198
- Other molecules: 5
- PDB ID: 1YBY

a, Rsym = ∑|I - <I>|/∑I, where I is the observed intensity of reflections;
b, Rwork-free = ∑|Fobs - Fcalc|/ Fobs, where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.

* outer shell values in parentheses

** MolProbity percent in all allowed regions shown in parentheses
Table 4.9  Statistics from the crystallographic analysis of Cth383

**Crystallization Conditions**
20% w/v PEG 3350, 0.2M trilithium citrate modified microbatch, temperature 291K

<table>
<thead>
<tr>
<th>Crystal</th>
<th>Cth-383_2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell dimensions</strong></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>P2₁ (P12₁₁₁)</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>49.91, 36.79, 61.82</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 111.93, 90</td>
</tr>
</tbody>
</table>

**Data collection**

| Source | APS 22-ID |
| Detector | MarCCD300 |
| Wavelength (Å) | 0.9794 |
| Distance (mm) | 230 |
| 20 (°) | 0.0 |
| Phi step (°) | 1.0 |
| Total Rotation (°) | 360 |

**Data processing**

| Resolution range (Å)* | 50.00-1.71 (1.77-1.71) |
| Reflections | 16,740 |
| Completeness (%)* | 73.3 (12.1) |
| R<sub>sym</sub>(%)* | 6.3 (39.6) |
| <I/σ(I)>* | Null |

**Refinement**

| Program | REFMAC 5 |
| Resolution range (Å)* | 30.96 – 1.80 (1.85-1.80) |
| Completeness (%)* | 100.0 (100.0) |
| R<sub>work</sub>(%)* | 22.5 (34.1) |
| R<sub>free</sub>(%)* | 25.6 (Null) |

**R.M.S. Deviations from ideality**

| Bond lengths (Å) | 0.015 |
| Bond angles (°) | 1.455 |
| Average B factor (Å²) | 28.26 |

**Ramachandran analysis**

| Most favored (%) | 98.9 (100.0) |
| Disallowed (%) | 0.0 |

**Final model:**

| Residues | A: 1-103; B: 12-103 |
| Solvent atoms | 90 |
| Other molecules | 4 |
| PDB ID | 1YBX |

a, R<sub>sym</sub> = Σ|I - <I>|/Σ|I|, where I is the observed intensity of reflections;
b, R<sub>work-free</sub> = Σ|F<sub>obs</sub> - F<sub>calc</sub>|/ F<sub>obs</sub>, where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.

* outer shell values in parentheses
** MolProbity percent in all allowed regions shown in parentheses
CHAPTER FIVE

Structure determination of the tudor domain of the human p100 protein, a non-structural genomics project using the improved high-throughput pipeline

The aim of our improved high-throughput pipeline is to improve the chance of crystallization of protein targets and/or improve the diffraction quality of protein crystals. Therefore, this pipeline may also be applicable to interesting proteins other than structural genomics targets. The human p100 protein was selected because it is a valuable protein with multiple important functions, and no structure was available before the start of this research.

On the other hand, some other methods that can improve the chance of crystallization (e.g., surface mutagenesis and genetic truncation) are still not incorporated into the pipeline and worth trying. The result of this study will provide information about how to incorporate those methods into the high-throughput pipeline.

The human p100 protein is composed of 885 amino acid residues. Based on sequence analysis by hydrophobic cluster analysis (HCA) method, the human p100 protein can be divided into two regions, the N-terminal *staphylococcal* nuclease (SN)-like region and the C-terminal tudor and SN (TSN) region (Callebaut & Mornon, 1997). The SN-like region, which is 639aa long, consists of a repeat of four SN-like domains. The TSN region (residues 640 to 885) comprises a tudor domain (residues 640-824) that is similar to the repeat segment of the *Drosophila* tudor protein, and a partial SN-like domain (residues 825-885).
The human p100 protein has several distinct functions. Originally identified as a transcription co-activator, the human p100 protein can interact with several transcription factors (e.g., EBNA-2 (Tong et al., 1995), STAT6 (Yang et al., 2002), STAT5 (Paukku et al., 2003) and C-myb (Leverson et al., 1998)) and bridge them to the basal transcription machinery. In addition, the human p100 protein can bind the histone acetyltransferase (HAT) domain of CBP and recruit the HAT activity to STAT6, thus opening the chromatin structure and facilitating transcription of corresponding genes (Valineva et al., 2005). The homologue of the human p100 protein in rice, Rp120, may direct the transport of prolamine mRNA (Sami-Subbu et al., 2001). The human p100 protein is also known as a component of the RNA-induced silencing complex (RISC) (Caudy et al., 2003). Moreover, the human p100 protein is involved in the processing of precursor mRNA (Shaw, Zhao et al., 2007). In conclusion, p100 plays multiple roles in transcription activation and RNA metabolism. The structure of the human p100 protein will help to delineate the molecular basis of its functions.

5.1 Different methods used to purify or modify the human p100 TSN region

The original human p100 gene and the encoding sequences of its SN-like region and TSN region were obtained from Dr. Jie Yang from the University of Tampere, Finland and Tianjing Medical University, China. For protein expression in E.Coli, the human p100 gene was cloned into six gateway destination vectors with different tags as described in 2.1.2. Encoding sequences of the SN-like region and the TSN region were cloned into pGEX-4T-1 vectors to form fusion proteins with GST tag and thrombin cleavage site at the N-terminals. All expression vectors were transformed into JM109 (Promega) and BL21 (DE3) (Invitrogen) as described in 2.1.2. The solubility test was
performed as described in 2.2.3. Except for its TSN region, the whole human p100 protein with different tags and its SN-like region with GST tag were expressed insoluble in *E.Coli*. The encoding sequence of the SN-like region was then transferred into six gateway destination vectors as described in 2.1.2. However, expressed proteins were still insoluble. Therefore, this work focused mainly on the TSN region of the human p100 protein.

### 5.1.1 Purification of the native and Se-Met substituted human p100 TSN region

Briefly, the native and Se-Met substituted proteins were expressed and purified through a GST affinity column, an ion exchange column, a Benzamidine affinity column and/or a gel filtration column as described in the materials and methods (2.1, 2.2, 2.3). The GST tags were removed during the purification (Figure 5.1). Since the vector pGEX-4T-1 contains a Tac promoter which is relatively weaker than the T7 promoter, 0.5 mM IPTG was still added in the auto-induction media at the cell density of OD$_{600}$~0.5 to improve the expression of the Se-Met substituted protein. After the GST tag removal, five residues (GSPEF) were left on the N-terminal of the protein. The native human p100 TSN region was purified from four different preparations. The final concentrations ranged from 5.4 mg/ml to 60.2 mg/ml (1st round, 20 mg/ml; 2nd round 5.4 mg/ml; 3rd round, 24.8 mg/ml; and 4th round, 60.2 mg/ml). The Se-Met substituted human p100 TSN region was purified six times. The final concentrations ranged from 6.4 mg/ml to 53.8 mg/ml (1st round, 31.6 mg/ml; 2nd round 14.8 mg/ml; 3rd round, 6.4 mg/ml; 4th round, 30.3 mg/ml; 5th round, 53.8 mg/ml; and 6th round, 22.3 mg/ml).
5.1.2 Reductive methylation of the native and the Se-Met substituted human p100 TSN region

Reductive methylation was performed for a part of the purified native and Se-Met substituted proteins. Lysine residues in the human p100 TSN region were reductively methylated to dimethyllysine as described in 2.4 (Figure 5.2A). The result of the reductive methylation was confirmed by LC/MS (Liquid Chromatography /Mass Spectrometry, the Chemical and Biological Sciences Mass Spectrometry Facility of the University of Georgia, [http://www.uga.edu/mass-spec](http://www.uga.edu/mass-spec)) (Figure 5.2B). The final concentrations of the reductively methylated native human p100 TSN region were 30 mg/ml and 32 mg/ml. The final concentrations of the reductively methylated Se-Met substituted human p100 TSN region were 11 mg/ml, 25 mg/ml, 29.6 mg/ml, and 15.8 mg/ml.

5.1.3 Surface mutagenesis of the human p100 TSN region

As discussed in the introduction, the mutation of Lys/Glu to Ala might improve the chance of crystallization. Six K/E clusters were found in the sequence of the human p100 TSN region (Figure 5.3A). The first one was near the end of the N-terminal. If the N-terminal was very flexible, mutation of this site might not be helpful for crystallization. According to the sequence analysis by HCA, the last two clusters might be in helices. Mutation of those sites might be less useful because their main chain carbonyls and amides were involved in intra-helical H-bonds and were not free for crystal contacts. (Derewenda, 2004) Therefore, the other three clusters (SM2, SM3, and SM4) were chosen for the surface mutagenesis experiment. Encoding sequences of surface K/E residues in these clusters were mutated to Ala as described in 2.1.1 (Figure 5.3B). The
solubility test (2.2.3) showed that only SM2 was partially soluble (Figure 5.3C). The mutated human p100 TSN regions (SM2) with or without Se-Met labeling were expressed and purified as described above (Figure 5.3D). The final concentrations were 12.7 mg/ml for the mutated human p100 TSN region (SM2) with Se-Met labeling and 13.3 mg/ml for that without Se-Met labeling.

5.1.4 Genetic truncation based on the information of the sequence analysis and the limited proteolysis

As mentioned earlier, the sequence analysis of HCA revealed that the human p100 TSN region contains a tudor domain. Based on this information, clones of six different truncated regions of the TSN region were constructed. The first one (td1; residues 640-824) covered the tudor domain. The second one (td2; residues 669-746) contained the most conserved region with the Drosophila tudor protein and the human Survival of Motor Neuron (SMN) tudor domain. The fifth one (td5; residues 695-758) corresponded to residues 82-147 of the SMN tudor domain, the structure of which had been solved. The second and the fifth truncated regions overlapped. The third one (td3; residues 669-758) covered the outer boundaries of the above two regions. The fourth one (td4; residues 695-746) covered the overlapped region of td2 and td5. The sixth one (td6) contained residues 695-885 from the N-terminus of the td5 to the end of the human p100 TSN region. All encoding sequences were transferred into three gateway destination vectors (pDEST527, 544, and 598), and td2 was transferred to two more vectors (pDEST566 and 596) as described in 2.1.2. Tev protease cleavage sites were inserted between the tags and the encoding sequences.Td1 to Td4 were partially soluble when they were expressed with the His6-Nus tag (pDEST 544) or the MBP-His6 tag (pDEST
Td5 and Td6 were partially soluble only with the His6-Nus tag and the MBP-His6 tag respectively (Figure 5.4). Unfortunately, it was difficult to isolate cleaved proteins after Tev protease digestion (example of Td2 is shown in Figure 5.5).

Limited chymotrypsin proteolysis cut out 33 residues from the N-termini of the human p100 TSN region. Based on this information and the result of sequence analysis mentioned earlier, clones of two truncated regions were made. The first one (PW; residues 673-885) covered the region from residue 673 to the end of the p100 protein. The second one (P1) contained residues from 673 to 824, the end of the tudor domain (Figure 5.6A). Encoding sequences of those regions and the His6 tags were cloned into the pDEST14 vector for expression. However, both protein products were insoluble (Figure 5.6B).

5.2 Crystallization

Crystallization conditions of the native and the Se-Met substituted human p100 TSN region and its surface mutant (SM2-1) before or after reductive methylation were screened against 384 reagent mixtures using the Cartesian Honeybee crystallization robot (Genomic Solutions) and the locally modified Douglas Instruments ORYX robot as described in 2.5. No crystals of the native or Se-Met substituted proteins were obtained. However, protein crystals of the methylated p100 TSN region with Se-Met labeling (11 mg/ml) appeared in the wells of several conditions (WI-13, 1.26 M ammonium sulfate, 0.1 M cacodyloate pH 6.5; WI-47, 1.26 M ammonium sulfate, 0.1 M Tris HCl/NaOH pH 8.5, 0.2 M lithium sulfate; MF-6, 1.0M ammonium sulfate, 0.1 M sodium acetate, pH 4.6; and HI-32, 2.0 M sodium formate, 0.1 M sodium acetate pH 4.6) after one month. In addition, crystals of the p100 SM 2-1 with Se-Met labeling (12.7 mg/ml) also appeared
under the condition of HI-32 after one month (Figure 5.7A). Most of these conditions contained high concentrations of either ammonium sulfate or sodium formate.

As described in 2.5, a single- or double-grid screen (pH and/or precipitant) centering those lead conditions was performed for the optimization at 18˚C by the modified microbatch under oil method. Better protein crystals of the methylated Se-Met substituted human p100 TSN region were found in WI-47-15 (0.1M Tris HCl/NaOH pH7.9, 1.26M ammonium sulfate and 200mM lithium sulfate) and 39 (0.1M Tris HCl/NaOH pH8.4, 1.26M ammonium sulfate and 200mM lithium sulfate) (Figure 5.7B).

Original protein samples for crystallization trials and the protein dissolved from the crystals in the gel filtration buffer were re-checked by SDS-PAGE (Figure 5.8A). The results showed that the proteins had been degraded during the crystallization trials and the crystals were actually formed from the degraded protein. The molecular weight of the degraded protein, checked by LC/MS, was around 11 kDa (Figure 5.8B). The N-terminal residues of the degraded protein were determined by Edman protein sequencing (Emory University Microchemical Facility, http://www.corelabs.emory.edu/mcf/) (Figure 5.8C). The degraded protein (corresponding to residues 673-773) covered most of the tudor domain, and the N terminal was very similar to the result of the limited chymotrypsin proteolysis (Figure 5.8D).

5.3 Data collection and processing

Thirty-nine crystals of methylated, Se-Met substituted and degraded protein (the human p100 tudor domain) were mounted from the wells of the optimized conditions, and flash-frozen without cryoprotective solutions. A total of 30 crystals were soaked with KI (13 crystals), K₂PtCl₄ (7 crystals), HgCl (2 crystals), or HgAc (8 crystals) from 20 min
to 52 h before mounting. Crystals were screened as described in 2.6. After three trips to SER-CAT, nine data sets of six crystals had been collected at the selenium absorption edge or 1.0Å using the 22ID beamline of SER-CAT as described in 2.6. The scan range of all data set was 360°. The crystal-to-detector distances ranged from 180 mm to 270 mm. The oscillation angle was 1° with 2-second or 3-second exposure time. Each data set was processed as described in 2.6. The crystals belonged to the space group P1, and unit cell dimensions were a=31Å, b=38Å, c=40Å, α=90.03°, β=105.32° and γ=99.72°. The values of R_sym ranged from 4.7% to 10.9%, and the highest resolutions ranged from 1.42Å to 2.13Å.

5.4 Structure solution of the human p100 tudor domain

The Matthews coefficient (Vm), the solvent content of the crystal, and the probability of molecules in the asymmetric unit were calculated as described in 2.7 with the molecular weight of the degraded protein (11 kDa). The asymmetric unit contained two molecules. The probability across all resolution ranges was 92% with solvent content=39% and Vm=2.01. Thin-shell .mtz files were made as described in 4.2.4.

Pipelines including SCA2Structure and ISASPipe, Solve/Resolve alone, and SHARP (delaFortelle & Bricogne, 1997) were tried with the above data sets to calculate the phase. However, positions of anomalous scatterers could not be found correctly. A group from the collaborative laboratory in China was successful in structure determination of the methylated native human p100 TSN region (PDB ID 2HQE). Using the corresponding region (residues 673-773) of this structure as a model, the phase of the structure of the degraded human p100 TSN region was obtained using a molecular replacement program, PHASER (McCoy et al., 2007) from the CCP4i suite. The model
was inspected using COOT and adjusted manually. Electron densities of residues 673-679 and 770-773 could not be observed; therefore, these residues were removed from the model. Model refinement was conducted using REFMAC. After several rounds of model rebuilding and refinement, the water molecules were picked using ARP/wARP. The selection criteria included a B-factor smaller than 45 Å² and a peak height greater than 3.2σ. The final model was validated using MolProbity through its web server (93). The Rwork and Rfree values of the final model were 23.2% and 27.2% respectively. MolProbity Ramachandran analysis showed that all residues were in the allowed regions, with 99.4% in the favored regions. The data collection and model refinement statistics are shown in Table 5.1.

5.5 The structure of the human p100 tudor domain and its putative function

The structure of the degraded protein (the human p100 tudor domain, residues 680-769) is composed of a long helix (α1) and a five-strand β-barrel (β1-β5) (Figure 5.9A). The helix (α1) and strands β1-4 are linked by several loops (L1-L4), while a helical turn connects strands β4 and β5. A hydrophobic core that consists of residues Ala713, Val725, Val736, Val747 and Val752 stabilizes the β-barrel (Figure 5.9B). At one end of the β-barrel, several aromatic residues (Phe715, Tyr721, Tyr738 and Tyr741) form a hydrophobic pocket (Figure 5.9C). For the convenience of discussion and comparison later, these residues were labeled r1, r2, r3 and r4.

The structure of the β-barrel of the human p100 tudor domain is similar to that of the SMN tudor domain (PDB ID 1MHN). Comparison of these two structures revealed an r.m.s.d of 1.2Å between Cα atoms. Differences were found mainly at loops, such as L2 (the loop linking β1 with β2), L3 (the loop linking β2 with β3), and L5 (the loop linking
β4 with β5) (Figure 5.10A). The structure-based alignment of the tudor domains from the human p100 protein and some other proteins revealed an absolutely conserved residue, Gly (corresponding to Gly742 of the human p100 protein). On L4 (the loop linking β3 with β4), the Gly742 residue is important for its distinct conformation (Selenko et al., 2001). The hydrophobic pocket is also conserved in the tudor domains of these proteins. (Figure 5.10B) Instead of aromatic residues, the hydrophobic pockets of some proteins contain some hydrophobic nonaromatic residues for r1-3 (e.g., Leu, Ile, or Val). The fourth residues (r4) of the hydrophobic pockets can be grouped into two types. Tudor domains of the human p100 protein, human TDRD2, human TDRD4-1, human SMN protein, and the mouse TDRD3 have aromatic or hydrophobic residues (Tyr, Phe, Val) at this position and belong to the first type. Tudor domains of the human 53BP1-1, JMJD2A-2, and LaminB have hydrophilic residues (Asp) at this position and belong to the second type. Based on this method, multiple-sequence alignment using ClustalW (Chenna et al., 2003) revealed that tudor domains of the human PHD finger protein 20 (PFP20) and the yeast-splicing factor 30 (Spf30) belong to the type II and the type I, respectively (Figure 5.10C). As described later, this may be related to the different functions of these tudor domains.

As shown in Figure 5.11 B and C, the hydrophobic pockets from tudor domains of the human 53BP1 and JMJD2A can bind methylated ligands. Therefore, the tudor domain of the human p100 protein may also interact with methylated ligands. It was revealed that the hydrophobic pocket of the human p100 protein could bind with the dimethyl group of Leu808 from the C-terminal fragment of another monomer packed together in the crystal (Shaw, Zhao et al., 2007). However, the details of the interaction between various
methylated ligands and the hydrophobic pocket of various tudor domains may be different. The tudor domains of 53BP1, JMJD2A, and PFP20 can bind di- or trimethylated Lys residues of histones (Botuyan et al., 2006, Huang et al., 2006, Kim et al., 2006). As discussed above, they belong to type II tudor domains with Asp residues at the positions of r4. The tudor domain of SMN, TRTD3 and Spf30 can bind dimethylated Arg residues at the RG-rich tail of the human Sm core proteins (Kim et al., 2006, Selenko et al., 2001). They belong to type I tudor domains with Tyr residues at the positions of r4. Therefore, the type I human p100 tudor domain may interact with methylated ligands that are similar to the dimethylated Arg residues.

In addition, the human p100 tudor domain structure described here revealed that it can bind Pro59 from L3 (the loop linking β2 with β3) of another monomer in the asymmetric unit (Figure 5.11D). This result revealed that the hydrophobic pocket of the human p100 tudor domain could also bind ligands with pyrrolidine rings or aromatic rings.

For further investigation of the function of the human p100 tudor domain, tudor domains of the human p100 protein and its homologues from 15 other species were aligned by ClustalX (Thompson et al., 1997) (Figure 5.12A). Six residues--corresponding to Asp717, Arg722, Asp740, Tyr741, Gly742, and Glu745 of the human p100 protein--were conserved in all proteins. As described earlier, Gly742 is conserved in almost all tudor domains. The conservation of Tyr741, the r4 residue of the hydrophobic pocket, indicates that it may be important for the specific functions of p100 proteins. The Arg722 forms salt bridges with Asp740 and helps fix the position of Tyr741. Asp717 and Glu745 form hydrogen bonds with Tyr721 and Tyr738, which are the r2 and r3 residues of the
hydrophobic pocket. These hydrogen bonds help fix the positions of Tyr721 and Tyr738. (Figure 5.12B) Therefore, the shape of the hydrophobic pocket and the Tyr residue at the position of r4 might be important for the functions of p100 proteins. In contrast, Tyr738 also fixes the position of Glu745. Mutation of this residue to Lys in the human SMN protein affects the binding of SMN with Sm core protein (Selenko et al., 2001). Therefore, in addition to the hydrophobic binding pocket, Glu745 may be also important for protein binding of the human p100 tudor domain.

In conclusion, the hydrophobic pocket of the human p100 tudor domain could bind methylated ligands and ligands with pyrrolidine or aromatic rings. The shape of the hydrophobic pocket and two conserved residues (Tyr741 and Glu745) may be important for these functions.
Table 5.1  Statistics from the crystallographic analysis of human p100 tudor domain

**Crystallization Conditions:**
1.06M ammonium sulfate, 0.2M lithium sulfate, 0.1M Tris, pH 7.9, modified microbatch, temperature 291K

**Crystal**
- **Space group**: P1
- **a, b, c (Å)**: 30.78, 37.68, 40.20
- **α, β, γ (°)**: 90.03, 105.32, 99.72

**Data collection**
- **Source**: APS 22-ID
- **Detector**: MarCCD300
- **Wavelength (Å)**: 0.9724
- **Distance (mm)**: 180
- **2θ (°)**: 0.0
- **Phi step (°)**: 1.0
- **Total Rotation (°)**: 360
- **Data processing**: HKL 2000
- **Resolution range (Å)***: 38.72-1.42 (1.46-1.42)
- **Reflections**: 25,303
- **Completeness (%)***: 82.2 (23.1)
- **R_sym(%)***: 5.0 (3.0)
- **<I/σI>***: 15.60 (Null)

**Refinement:**
- **Program**: REFMAC 5
- **Resolution range (Å)***: 38.72-1.42 (1.46-1.42)
- **Completeness (%)***: 82.2 (23.1)
- **R_work(%)***: 22.9 (36.4)
- **R_free(%)***: 27.0 (45.0)

**R.M.S. Deviations from ideality**
- **Bond lengths (Å)**: 0.009
- **Bond angles (°)**: 1.210
- **Average B factor (Å^2)**: 16.11

**Ramachandran analysis**
- **Most favored (%)**: 99.4 (100.0)
- **Disallowed (%)**: 0.0

**Final model**
- **Residues**: A: 680-769; B: 680-769
- **Solvent atoms**: 329
- **Other molecules**: 0
- **PDB ID**: 2HQX

a, \( R_{\text{sym}} = \frac{\sum|I - <I>|}{\sum I} \), where I is the observed intensity of reflections;  
b, \( R_{\text{work-free}} = \frac{\sum |F_{\text{obs}} - F_{\text{calc}}|}{\sum F_{\text{obs}}} \), where the R-factor is calculated including and excluding refinement reflections. Free reflections consist of 5% of the total number of reflections.  
* outer shell values in parentheses  
** MolProbity percent in all allowed regions shown in parentheses

134
Figure 5.1 SDS-PAGE results of the native (A) and Se-Met substituted (B) human p100 TSN region (p100td) after purification steps
A, lane 1 and 2, the pellet and supernatant after cell lysis and centrifugation; lane 3, the flow through fraction from the GST affinity column; lanes 4-5, flow through fractions containing p100td after on GST affinity column Thrombin digestion; lanes 6-7, elution fractions containing GST after on GST affinity column Thrombin digestion; lane 8, thrombin; lanes 9-15, elution fractions of ion exchange chromatography; B, lane 1- the whole cell expression of the Se-Met substituted protein after IPTG induction; lane 2, flow through from the GST affinity column; lane 3, the elution fraction containing the GST tagged protein from the GST affinity column; lane 4, after Thrombin digestion; lanes 5 and 6, GST tag removal by GST affinity column; lane 5, the flow through fraction containing p100td-SeMet; lane 6, the elution fraction containing GST; lanes 7-10, the flow through fraction from the Benzamidine FF column containing p100td-SeMet.
Figure 5.2 Reductive methylation of the native human p100 TSN region (p100td) 
A, SDS-PAGE of the methylated protein (p100td-mth) after buffer exchange (lanes 1-3); 
B, the molecular weights of the native proteins before (a, 29694 Da) and after (b, 30104 Da) reductive methylation were determined by LC/MS; the human p100 TSN region contains 12 Lys residues and the N-terminal -NH2 group of backbone might also be reductively methylated. 28 Da is added per methylated lysine residue / N-terminus. (30104-29694)/28 > 13. Therefore, all Lys residues were methylated.
Figure 5.3 Surface mutagenesis of the human p100 TSN region (p100td)
A, six K/E clusters were found in the sequence of the TSN region; B, Agarose Gel Electrophoresis of DNA fragments encoding proteins of mutated K/E clusters (SM2-SM4); C, SDS-PAGE results of protein mixtures after cell lysis and centrifugation indicating only the TSN region (p100td) with mutated K/E cluster 2 (SM2-1) was partially soluble; S, supernatant; P, pellet; D, SDS-PAGE results of SM2-1 after purification steps; lane1, the pellet after cell lysis and centrifugation; lane2, the flow through fraction from GST affinity column; lanes 3-7, elution fractions from GST affinity column containing the GST tagged SM2-1; lane8, after Thrombin digestion; lanes9-12, after GST tag removal; lanes13-15, after gel filtration chromatography.
Figure 5.4 Solubility testing of truncated regions (td1-td6) of the human p100 TSN region SDS-PAGE results of protein mixtures after cell lysis and centrifugation. Td1 to Td4 were partially soluble when they were expressed with the His6-Nus tag (pDEST 544) or the MBP-His6 tag (pDEST 598). Td5 and Td6 were partially soluble only with the His6-Nus tag and the MBP-His6 tag respectively. S, supernatant; P, pellet;
Figure 5.5 SDS-PAGE of p100td2-544, 566 and 598 after Tev protease digestion and Ni affinity purification

p100td2 cannot be separated from the tags or fusion proteins.
Figure 5.6 Genetic truncation based on the information of limited proteolysis
A, sequences of PW and P1, included in the red rectangles;
B, the SDS-PAGE result after expression, cell lysis and centrifugation; both
protein products were insoluble.
S, supernatant; P, pellet.
Figure 5.7 Crystallization of the human p100 tudor domain
A, crystals appeared in the initial screening of the methylated proteins with Se-Met at 18°C (a, WI-47, 1.26M ammonium sulfate, 0.1M Tris HCl/NaOH pH8.5, 0.2M lithium sulfate; b, MF-6, 1.0M ammonium sulfate, 0.1M sodium acetate, pH4.6; c, WI-13, 1.26M ammonium sulfate, 0.1M cacodyloate pH6.5; d, HI-32, 2.0M sodium formate, 0.1M sodium acetate pH4.6) and the surface mutated proteins (SM 2-1) with Se-Met (e, HI-32, 2.0M sodium formate, 0.1M sodium acetate pH4.6);
B, Better protein crystals of the methylated proteins with Se-Met were found in conditions of a. (0.1M Tris HCl/NaOH pH7.9, 1.26M ammonium sulfate and 200mM lithium sulfate) and b. (0.1M Tris HCl/NaOH pH8.4, 1.26M ammonium sulfate and 200mM lithium sulfate) at 18°C.
Figure 5.8 Identification of the degraded human p100 TSN region

A, SDS-PAGE of proteins from dissolved crystals (lane 1) and samples used for crystallization trial (lane 2);
B, the molecular weight of the degraded protein determined by LC/MS;
C, the N-terminal residues of the degraded protein were checked using Edman protein sequencing;
D, possible sequence of the degraded protein.
Figure 5.9 The structure of the human p100 tudor domain
A, ribbon representation;
B, the hydrophobic core consists of residues Ala713, Val725, Val736, Val747 and Val752; side chains of these residues are shown in the color of raspberry;
C, the hydrophobic pocket formed by side chains of Phe715, Tyr721, Tyr738 and Tyr741; side chains of these residues are shown in the color of green;
Figure 5.10 The tudor domain of the human p100 protein is similar to that of SMN
A, the overlay of the structure of the human p100 tudor domain and
that of the human SMN; differences are mainly found at loops (L2, L3 and L5);
B, the structure-based alignment of the tudor domains from the
human p100 protein and some other proteins;
C, the multiple sequence alignment of the tudor domains using
ClustalW;
r1-4, residues of the conserved hydrophobic pocket; *, conserved Gly.
Figure 5.11 The hydrophobic pockets (green) in the tudor domains of the human SMN, 53BP1, JMJD2A and p100 protein, and their ligands (raspberry)
A, the human SMN; B, the human 53BP1; C, the human JMJD2A;
the human p100 protein; r1-4, residues of the conserved hydrophobic pocket; Lys-mth2,3, the di- or tri- methylated Lys residues.
Figure 5.12 Conserved residues of the tudor domains of the human p100 protein and its homologues
A, multiple sequences alignment of the tudor domains of these proteins using ClustalX; *, conserved residues;
B, conserved residues that are related to the hydrophobic pocket
CHAPTER SIX

Conclusions

Crystal-directed methods used in this research included additional purification steps, tag removal, Se-Met substitution, reductive methylation, surface mutagenesis, genetic truncation and alternative crystallization trials set-up methods. The success of each method on different targets was different and depended on the properties of the selected targets.

Ion exchange chromatography was mainly used for additional purification. Although this method alone did not result in any structures in this research, it was successful for targets of other groups (Liu, Shah et al., 2005) or in combination with other methods (AF0160). The His\textsubscript{6} tag had either positive or negative impact on protein crystallization or crystal diffraction quality, depending on individual proteins. For example, the structure of Cth393 was solved with the His\textsubscript{6} tag, while the structure of AF0160 without the tag. Therefore, both forms of proteins are worth trying. As discussed in the introduction, Se-Met substitution could change the entropy of proteins. Results of this research revealed that Se-Met substitution could be used as a salvaging procedure for protein targets that cannot obtain structures in the native form. On the other hand, if the targets failed in the Se-Met form, solving the native structure by Sulfur-ISAS phasing is also a good strategy. An example is the structure of AF1382, which was solved using Sulfur-ISAS method by Jinyi Zhu et al. (PDB ID 2QVO). Reductive methylation and surface mutagenesis could dramatically change the surface entropy of proteins. As an
example, the human p100 tudor domain was crystallized only when the reductive methylation or the surface mutagenesis was performed. However, re-cloning is required for the method of surface mutagenesis, and the solubility problem may emerge for the new protein products. Natural degradation and limited proteolysis helped in the attainment of the stable domains of the proteins, which are more prone to crystallize. Homogeneous proteins could be obtained by genetic truncation based on the result of these methods and the sequence analysis. However, it is still difficult to identify clearly the boundaries of stable domains by current sequence analysis methods (Weiner & Lu, 2002). Again, the solubility problem may emerge because the target has to be re-cloned and re-expressed. The above methods can be combined to salvage targets.

To summarize the efforts and the experience in this research, a general salvaging strategy is concluded here. First, the concentration, the purity, and the integrity of the proteins should be re-checked. In the event of poor diffraction quality of the crystals, different set-up methods, additional purification steps, different forms (Se-Met/native), reductive methylation, and tag removal/adding should be considered. If proteins do not crystallize, different forms (Se-Met/native), reductive methylation, tag removal/adding, surface mutagenesis, and limited proteolysis are good choices. Genetic truncation based on protein sequence analysis is not recommended at the present time, unless no other information is available.

Some salvaging methods discussed above were included in the improved high-throughput pipeline. As described in Chapter 4, the improved high-throughput pipeline achieved a 14% success rate from purified protein to structure—much better than the 7% success rate for PSI centers as of August 2004. However, 86% of the purified proteins
failed to yield structures. More methods are needed, and no one method could dominate the pipeline. It is not surprising that one set of methods is suitable for only one group of proteins. Therefore, at each step of the high throughput pipeline--from cloning to structure determination--several candidate methods should be screened in parallel to find the most suitable methods. This pipeline should be an open method, so that any new useful method can be optimized and integrated into it. In addition, a salvaging strategy should be employed in the high throughput pipeline for failed targets. On the other hand, with the accumulation of information, it may finally become possible to predict suitable methods for individual proteins. Hopefully, this method can finally be used to solve the structures of many interesting proteins whose structure information cannot be obtained at the present time.
REFERENCES


159


Shaw, N., Zhao, M., Cheng, C., Xu, H., Saarikettu, J., Li, Y., Da, Y., Yao, Z.,


