EXPLORING MOLECULAR BIOLOGICAL AND STRUCTURAL ASPECTS OF NON-PFAM PROTEINS

by

JINYI ZHU

(Under the Direction of Bi-Cheng Wang)

ABSTRACT

A non-Pfam protein sequence, by definition, has no significant match to any sequence in the Pfam protein family classification. Non-Pfam proteins are often species-specific and related to evolutional and developmental functions. They have been excluded from most structural genomics projects, which aim to build a complete protein fold space, due to their unique sequences. However, these unique structures could contribute many novel protein folds. The high throughput structural genomics pipelines developed at the SECSG were used to determine the 3D structures of selected non-Pfam targets. In addition, sulfur-SAS phasing methodology was refined and expanded to crystal having moderate low to moderate resolution. In this work, 51 out of 62 selected non-Pfam proteins were expressed in large-scale media and 23 proteins were successfully purified by chromatography methods. Crystallization hits were observed in 8 proteins. Crystal structure of AF1382 was determined by sulfur-SAS phasing using merged medium-resolution data. AF1382 belongs to a winged-helix fold and has putative interactions to DNA. Crystal structure of TT0030 was determined by isomorphous replacement. TT0030 is similar to a Rossmann fold. Their atomic coordinates and structure factors have been deposited into PDB. Meanwhile, 4 non-Pfam protein crystal structures have been determined by other

members of the lab. The structure of PH1580 represents a new SCOP fold and the structure of AF2093 is very likely to represent a new fold. Proteins PF1176 and AF0160 are structurally similar to proteins with identified functions. This research shows that: i) the majority of non-Pfam proteins are real, foldable proteins; ii) non-Pfam proteins can make a significant contribution to the discovery of new SCOP folds; iii) many non-Pfam proteins are biologically important; iv) sulfur-SAS is a viable phasing method with the medium-resolution data.

INDEX WORDS: non-Pfam, protein fold, sulfur-SAS, AF1382, TT0030, structural genomics

EXPLORING MOLECULAR BIOLOGICAL AND STRUCTURAL ASPECTS OF NON-PFAM PROTEINS

by

JINYI ZHU

B.S., Nanjing University, China, 2002

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

2008

EXPLORING MOLECULAR BIOLOGICAL AND STRUCTURAL ASPECTS OF NON-PFAM PROTEINS

by

JINYI ZHU

Major Professor: Bi-Cheng Wang

Committee: John Rose
Robert Scott

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

DEDICATION

This dissertation is dedicated to my parents. They have opened my eyes to the world and always encouraged me to move forward.

ACKNOWLEDGEMENTS

So many people have helped me during my graduate study. First and foremost I would like to thank Dr. Bi-Cheng Wang, my advisor and committee chair, for his guidance and kind support during these years, also for his patience and encouragement in my research. He has created an exhilarating environment of science. I feel fortunate and honored to study and pursue research under his supervision. I also would like to thank my committee members, Dr. John Rose, for his instructions in research and precious time in correcting my dissertation; and Dr. Robert Scott, for his critical thinking about my research and dissertation.

My sincerest thank goes to Dr. Zhi-Jie Liu. He offered me great help during my most difficult time and helped me build confidence in research. Without him teaching me details through every experiment in X-ray Crystallography, I would not be able to complete this research. My deepest thank also extends to Dr. Hao Xu, for the knowledge I have learnt in protein production, a field I have never touched before, also for his proofreading. I heartily thank Dr. Gary Newton for his criticisms on the first chapter.

I am grateful for working with two wonderful ladies successively, Jenny Hwang and Jessie Chang. As a team, we have worked with high efficiency and harmony. I also would like to thank the current and former members of the Wang's group for being a constant motivation to work harder: Ms Lily Li, Dr. Min Zhao, Mr. Dayong Zhou, and Dr. Lirong Chen. Last but not least, I would like to thank all of my friends at UGA who have made the life at Athens a real pleasure.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVATIONS

Abbreviation	**Full name**
3D	three-dimensional
AF0160	*Archaeoglobus fulgidus* ORF 0160
AF1382	*Archaeoglobus fulgidus* ORF 1382
AF2093	*Archaeoglobus fulgidus* ORF 2093
ANL	Argonne National Laboratory
APS	Advanced Photon Source
Coot	Crystallographic Object-Oriented Toolkit
DALI	Distance Matrix Alignment
DF1	Due Ferro 1
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HMM	hidden Markov model
HTH	Helix-turn-helix
IPTG	isopropyl β-D-1-thiogalactopyranoside
ISAS	Iterative Single-wavelength Anomalous Scattering
LB	Luria-Bertani
MAD	Multi-wavelength Anomalous Dispersion
NAD	nicotinamide adenine dinucleotide
NCS	non-crystallographic symmetry

NMR	Nuclear Magnetic Resonance
OD_{600}	optical density under 600 nm wavelength
ORF	open reading frame
PDB	Protein Data Bank
PF1176	*Pyrococcus furiosus* ORF 1176
PH1580	*Pyrococcus horikoshii* ORF 1580
PISA	Protein Interfaces, Surfaces and Assemblies
PMSF	phenylmethylsulphonyl fluoride
PSI	Protein Structure Initiative
RMSD	Root Mean Square Deviation
SAS	Single-wavelength Anomalous Scattering
SCOP	Structural Classification of Proteins
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SECSG	Southeast Collaboratory for Structural Genomics
SeMet	seleno-methionine
SER-CAT	Southeast Regional Collaborative Access Team
SG	Structural genomics
SSM	Second Structure Matching
STIV	*Sulfolobus* turreted icosahedral virus
TT0030	*Thermus thermophilus* ORF 0030
UGA	University of Georgia

CHAPTER ONE

Introduction

Structural genomics (SG) projects have been initiated all over the world to determine structures of a large number of proteins, which will be used to build a complete protein fold space. To effectively reduce the cost of structure determination, structural genomics projects focus on the representatives of sequence homologous proteins. For this reason, the Pfam database has been extensively incorporated in the target selection and Pfam-A proteins have been the major targets of many structural genomics projects.

However, about one million non-Pfam proteins have been excluded from structural genomics analysis because their sequences appear to be unique. It is very possible that these unique structures could contribute many novel protein folds. Thus, the major goals of this study is to (1) use the high throughput structural genomics pipelines developed at the Southeast Collaboratory for Structural Genomics (SECSG) to show that the 3-dimensional (3D) structures of non-Pfam A proteins yield novel folds and (2) to refine and expand sulfur-SAS phasing methodology to crystal having moderate low to moderate resolution using proteins produced for these studies.

1.1 Development of Pfam

A protein domain is evolutionary unit whose coding sequences are capable of undergoing duplication and/or recombination (Chothia *et al.*, 2003), while a protein family is made up of proteins that have evolved from a common ancestor. Proteins within a family usually have significant similarity among their sequences, functions, and 3D structures. The term "protein

family" may refer to either a small group of proteins with almost identical sequence or a huge group of proteins with the lowest possible level of detectable sequence similarity. To diminish this ambiguity, the concept of protein superfamilies (Dayhoff, 1976) was introduced along with the following classification:

superfamily > family > sub-family.

Determined from evolutionary, functional and structural data, a superfamily combines all the domains of different families that share a common evolutionary ancestor.

In addition, the number of protein sequences has been increasing rapidly in the recent decades due to the many genome-sequencing projects and a large portion (between 40% to 65% (Sonnhammer *et al.*, 1997)) of these proteins have similar sequences to proteins of known function. The sequence similarities among these newly found proteins are also significant since they may imply function. Therefore, a system to organize proteins using multiple sequence alignments has become important and necessary.

The Pfam database, developed in 1997, is "a comprehensive database of protein domain families" (Sonnhammer *et al.*, 1997). It consists of three parts: Pfam-A proteins, Pfam-B proteins and other proteins. For this work we define non-Pfam proteins as any proteins whose sequence is not in Pfam A.

Two major factors that had to be considered when developing the Pfam database were how to cluster the protein sequences and how to build multiple alignments within them. Previous methods emphasized either full protein domains or short conserved regions within the protein sequence. Although matches from short conserved regions contain useful information such as functional sites, they often fail to provide information about the domain boundaries.

Thus, whole domain alignment, which is more sensitive to domain annotation, provides better information for family-based sequence analysis.

Automatic methods of sequence alignment are most frequently used for databases containing huge numbers of families. However, automated methods have the following disadvantages: i) poor quality alignments, ii) unreliable domain boundaries, and iii) significantly increased computing time as the number of protein sequences increase. Manually constructed alignment provides more accurate results but is less applicable for maintaining large and growing data sets. A combination of manual and automatic alignment approaches was adopted by the Pfam database. The major concern is how much manual alignment should be applied, which is directly related to the distribution of the protein family sizes. While most proteins are distributed in a relatively small number of common families, they also share a few common folds and superfolds. This situation makes it possible to employ a semi-manual method focusing on the largest families to catch a large portion of proteins. In practical operations, the alignment would be best performed based on both sequence data and structural information at the superfamily or family level.

Pfam-A

In order to maintain an up-to-date record of the manual alignments, especially for the new protein sequences, Pfam adopts two kinds of alignments: i) a manually curated "seed" alignment with few changes between releases, and ii) an automatically generated "full" alignment based on profile hidden Markov models (profile HMMs) (Eddy, 1998). This is done by first, harvesting members of a seed alignment from various databases or published alignments. Alignments against these seeds are then carried out. The alignments will be examined for important features such as active site residues or structurally important residues, as well as

truncation, frameshift, or incorrect splicing errors. Profile HMMs are built from each seed alignment by an hmmb (HMM build) program HMMER suite (http://hmmer.janelia.org/). Then an automatic search will be carried out by aligning each HMM with all sequences in UniProtKB database (http://www.uniprot.org/) (Bairoch *et al.*, 2007) and the NCBI GenPept database (Wheeler *et al.*, 2008). The search result is then compared with existing members of the family. In the case where a seed alignment is very small, newly found sequences are combined with the original one(s) and the search is repeated. Several quality control steps are performed after the construction of seed and full alignments. Functional annotation, literature references and database links are also included for each family. The whole process is outlined in Figure 1.1.

Starting from 175 families and 15610 sequences in Pfam Release 1.0, the current Pfam release (version 22.0) contains 9318 protein families and almost 3 million, accounting for 74% of total sequences in the database. The source of sequence information has now been expanded to include UniProtKB, NCBI GenPept and other sequences from selected metagenomics projects.

Pfam-B

Pfam-B was designed to provide completeness to Pfam-A. In practice, it has become a useful resource for potential new Pfam-A families. Prior to Pfam version 3.4, Pfam-B was automatically built using DOMAINER (Sonnhammer *et al.*, 1994), a computationally expensive program based on an all-against-all BLAST comparison. However, since Pfam version 4.0, Pfam B sequences are selected from the ProDom database of protein domain families (Corpet *et al.*, 1999) that uses an improved (faster) alignment scheme. This approach has proved to be an effective way to detect missing members of families, by comparing Pfam-A with ProDom during the database construction. For example, in Pfam version 4.1, after incorporating some domains

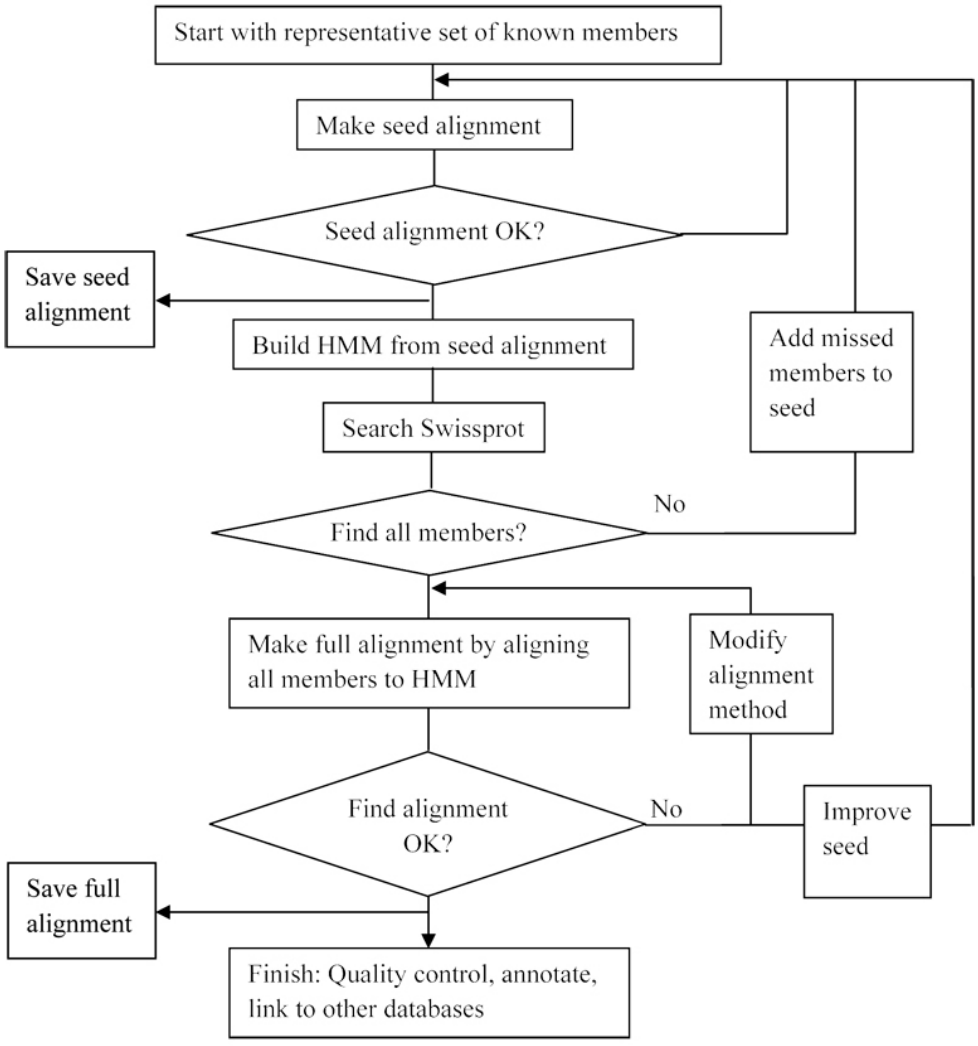


Figure 1.1: The original procedure of a HMM and an alignment construction for a Pfam-A family. Adapted from Sonnhammer, E. L. L., S. R. Eddy, et al. (1997). Proteins-Structure Function and Genetics 28(3): 405-420.

from Pfam-B with the 31 sequences used in the seed alignment, the number of PF00355 family members of increased from 51 to 192 (Bateman *et al.*, 2000).

Others

There are a significant number sequences that fail to meet the criteria to be included in either Pfam A or Pfam B. These sequences are classified as others since they lie outside the Pfam definitions.

Pfam clans

A domain in a given sequence can only exist in one Pfam family to avoid overlap between families. Because many families are highly related, the concept of clan was introduced into Pfam in 2005, to illustrate the relationships between different families (Finn *et al.*, 2006). A clan consists of two or more Pfam-A families that arise from a single evolutionary origin. Clans are manually constructed based on many factors including structures, functions, profile–profile comparisons and other databases such as SCOP(Andreeva *et al.*, 2004). The profile–profile comparisons were performed by tools such as PRC (http://supfam.mrc-lmb.cam.ac.uk/PRC/), HHsearch (Soding, 2005) and SCOOP (Bateman *et al.*, 2007).

Clans provide a hierarchical classification of Pfam families and improve the annotation of families. They offer help for better identification of hidden structural homologues and more accurate functional and structural predictions for protein families. In Pfam version 22.0, there are 283 clans, covering 1808 Pfam-A families and 43% of the Pfam domains. This analysis has indicated that many families are related and many large families are included in clans.

The major difference between a Pfam-A family and a Pfam-B family is the quality of alignment. HMM searches are carried out to classify new domain sequences into Pfam-A families, which are carried out by HMMER. E-values (expectation values) are calculated like

BLAST (Altschul *et al.*, 1997). Good matches are indicated by E-values much less than 1. Most new Pfam-A families were built using entries from the PROSITE database (http://ca.expasy.org/prosite/) (Hulo *et al.*, 2008) and PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/) (Mulder *et al.*, 2007). Only information from large PROSITE families (> 15 members) was used to construct Pfam-A families. Moreover, in the recent release of Pfam version 22.0, over 500 new Pfam-A families were manually built based on PDB sequences that were not previously covered.

It has been more than ten years since the Pfam database was introduced. The Pfam database has provided advances in domain analysis and classification over traditional databases and has become an important part in many genomic sequencing projects. The coverage and quality of families have been consistently improving as well as the annotation system and access to the database.

1.2 Pfam and Non-Pfam proteins in Structural Genomics

Over the past two decades, genome-sequencing projects have provided a large number of proteins whose function or structure is unknown. In 2001 Mittl and Grutter suggested that with present state of knowledge the about one-third structures and two-thirds of the function of the proteins for the whole genome sequences could be predicted (Mittl *et al.*, 2001). Since then structural- and functional-genomics projects have been initiated all over the world to better define structure and function for the remaining annotated proteins.

Contrary to traditional biological research, which identifies function first, structural genomics projects determine the proteins 3D structure to drive the functional investigation. High-throughput methods and new strategies of target selection have been developed by a number of structural genomics centers aimed at lowering the average cost of structure

determination. If successful, structural genomics should produce two major benefits: i) a complete description protein fold space and ii) identification of distant evolutionary relationships not recognized from sequence (Brenner *et al.*, 2000). Thus two major directions have emerged: one focused on solving a representative structure for all existing undocumented folds and the focused on the using structure to infer the function of a particular protein.

Currently, identifying structural features that can be related to possible function is the main objective of most projects. In addition, a protein structural model can be accurately predicted by programs, for example, MODELLER (Sali *et al.*, 1993), if it shares more than 30% sequence identity with a template structure (an experimentally determined 3D structure). Thus, it is reasonable to prioritize target selection based on representatives from each homologous family or superfamily since the first protein structure in a family could be used to discover function, mechanism, fold or even uncharacterized evolutionary relationships in other family members (Chandonia *et al.*, 2006).

Generally, a target selection process of a structural genomics project can be described as following (Brenner, 2000).

Stage 1: *Initial protein selection*. Proteins in the realm of interest (depending on the Center's focus) are selected and may come from one or several organisms or genomes, or a group of proteins with a particular function (e.g. enzymes). These initial targets are then organized into families and representative member of each family selected would be select for structural characterization.

Stage 2: Family exclusion. In this step, proteins that present extreme challenges to structural studies by X-ray or Nuclear Magnetic Resonance (NMR) are excluded. For example, structural determinations of transmembrane proteins have been found to have a low success rate

(both in protein production and structure determination) and thus, are not suitable for high-throughput projects. Low complexity regions have been confirmed to affect the successful determination of structures by NMR and X-ray crystallography (Bannen *et al.*, 2007). In addition, proteins with known 3D structures or whose structures could be modeled based on structural templates are removed.

Stage 3: Remaining families prioritized. Although the ultimate goal is covering the entire protein fold space, prioritization could ensure rapid coherence and relevance within the given information. The process in each structural genomics group may vary because of the different focus of the Center or project. A popular approach would focus on proteins distributed over a large number of organisms. Those proteins are ancient and conserved, thus are important to cellular function in general. Solving their structures would provide a good chance at understanding important biological functions. Other groups may have interests in: i) determining structures for a complete genome, ii) determining structures for the easiest to study proteins in specific genome, and iii) determining structures of the proteins who have few or no sequence homologues, the so called ORFans, also known as orphan ORFs (open reading frames).

Step 4: Identification of specific proteins to be studied. Factors such as molecular size, thermostability, pI, and methionine counts are generally evaluated during this stage. The final targets might often be homologs of the original proteins of interest.

At least 8 databases and 7 programs are incorporated in the first two stages above. Sequences are clustered into families with other homologous ones using methods like BLAST (Altschul *et al.*, 1997), PSI-BLAST (Altschul *et al.*, 1997), and HMMs (Eddy, 1998). Possible functional relationships could be revealed using databases such a Pfam and SMART (Schultz *et al.*, 2000), which contain multiple sequence alignments of individual protein and domain families;

PROSITE, which contains profiles generated from multiple sequence alignments; PRINTS, TIGERFAMs (Haft *et al.*, 2001), and ProDom. The SCOP (Hubbard *et al.*, 1999) and CATH (Orengo *et al.*, 1997) databases have been extremely valuable resources for structure/function relationships between proteins. For example, structural alignments are obtained by different algorithms including DALI (Holm *et al.*, 1993), which is based on Cα contact distances; SSAP (Orengo *et al.*, 1996), VAST (Madej *et al.*, 1995), and PrISM (Yang *et al.*, 2000), which are based on secondary structure information.

In the United States, the Protein Structure Initiative (PSI) (http://www.nigms.nih.gov/psi/) is the largest structural genomics initiative in the world. The pilot stage of PSI started in 2000 with seven centers organized for developing methods and technologies for a future production phase. Two centers joined in the next year. As one of the original seven pilot centers, the Southeast Collaboratory for Structural Genomics (SECSG) 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.

At SECSG, three genomes were chosen for study: *Pyrococcus furiosus* (*P. furiosus*), *Caenorhabditis elegans (C. elegans)*, and selected proteins for the human (*Homo sapiens*) genome (Adams *et al.*, 2003; Wang *et al.*, 2005). *P. furiosus* is an extremophilic species of Archaea, with a small genome of about 2,200 ORFs. It was chosen because it is a slowly evolving archaeon with a small genome of about 2,200 ORFs. The *C. elegans* genome was predicted to have at least 19,000 ORFs. SECSG target selection included most of the *P. furiosus* (2,182 ORFs) and *C. elegans* (14,442 genes) genomes and 446 human genes. In addition, targets that shared more than 30% sequence identity to any Protein Data Bank (PDB) entry or contained more than three predicted transmembrane domains were excluded from the initial target list.

Later a number of non-Pfam targets were added to the SECSG target pool, which form the basis of the work presented here. These include 328 sequences from *Aeropyrum pernix*, 223 sequences from *Clostridium thermocellum,* 268 sequences from *Archaeoglobus fulgidus,* 205 sequences from *Pyrococcus horikoshii,* and 261 sequences from *Thermus thermophilus*.

As a result of structural genomics efforts by February 1, 2005, 36% of the Pfam families contained one or more members with known structures. PSI centers had solved 1,032 protein structures, almost two-thirds of all structures solved by worldwide structural genomics centers (Chandonia *et al.*, 2006). A total of the 597 PSI-structures (58%) shared less than 30% sequence identity with any known structure. About 20.4% of all PSI structures represented a new Pfam family, while the rate from non-SG structures was only 5%. Every year, half of new structurally characterized families come from structural genomics centers worldwide, although they only account for 20% of new structures. PSI centers also contributed 74 new folds or superfamilies. By May 15, 2004, 16% of domains from PSI centers represented a new fold or superfamily, which was much higher than 4% from non-SG structures.

Non-Pfam proteins are not included in most structural genomics projects because of their lack of sequence homology. In addition, many non-Pfam proteins are ORFans which means that they have (1) no homolog in any genome (a singleton ORFans) or (2) have homologs only in the same genome (a paralogous ORFan) or a closely related genome (an orthologous ORFan). ORFans usually account for 20-30% of sequences in newly sequenced genomes (Siew *et al.*, 2003b). Their origins remain unknown. During the first few years of discovery, ORFans were thought to be results of the sparse sampling of the sequence space and would disappear with completeness of genome sequences. Some considered ORFans as non-essential proteins (Schmid *et al.*, 2001) or expressed proteins related to errors or incorrect genes (Goffeau *et al.*,

1996). However, gene analysis of 60 complete genomes (Siew *et al.*, 2003a) showed that the total number of ORFans continued to increase as the number of sequenced genomes grew. In addition, although the overall fraction of ORFans had been diminishing, it still remains high in newly sequenced genomes. Furthermore, a majority of longer ORFans have been shown to be expressed as folded proteins. A more recent structural analysis of nineteen ORFans structures (Siew *et al.*, 2004) suggested that many ORFans could be real, foldable proteins rather than sequencing errors. The nineteen ORFans came from organisms spanning all kingdoms of life and thirteen ORFan corresponded to proteins with experimentally derived function, which also suggest that ORFan could be biologicaly significant.

Based on this, ORFans may be considered as either a new and previously unseen protein or distant relatives of known families whose sequence has diverged beyond recognition by sequence comparison tools in use today. In the future, with the growth of the genome sequence database and more sensitive computational tools today's ORFans may be assigned to their proper families.

1.3 SCOP Fold

Protein folds refer to spatial arrangement of regular secondary structural elements in the proteins. They are important for structural classification and could be used to interpret structures of proteins with similar sequences. Proteins structural similarity also leads to implications of their evolutionary origin and possible function.

To help access structural similarities between rapidly increasing protein structures deposited in Protein Data Bank (PDB; (Berman *et al.*, 2000)), a databse, Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) was constructed in 1994. It attempts to classify all known protein structures by visual inspection and comparison. Other than the basic

classification of protein structural domain, SCOP is constructed on four levels of hierarchic structural classification: i) *Family*, ii) *Superfamily*, iii) *Common fold* and iv) *Class*.

Family. Proteins will be grouped into a family if they meet either of two criteria: i) 30% or greater residue identities, ii) similar functions and 3D structures but have lower sequence identities. Those criteria are indicative of a common evolutionary origin of the family members.

Superfamiliy. Families are classified into a superfamily when structures and/or functional features indicate a possible common evolutionary origin. The proteins in a superfamily usually have low sequence identity.

Common fold. Several families and superfamilies will have a common fold if their members have the same major secondary structures in similar arrangement and topology, but no evolutionary relationships among them.

Class. Different folds are grouped into seven classes: i) all alpha, whose structures are basically formed by helices; ii) all beta, whose structures are basically formed by β-sheets; iii) alpha and beta, those with largely interspersed α-helices and β-strands; iv) alpha plus beta, those with largely segregated α-helices and β-strands; v) multi-domain, for those structures with domains of different fold; vi) membrane and cell surface proteins and peptides, which do not include proteins in the immune system; vii) small proteins, usually dominated by metal ligand, heme, and/or disulfide bridges.

In the latest release of SCOP v1.73 (Andreeva *et al.*, 2008), 92,927 domains from 34,495 PDB entries are classified into 3,464 families, 1,777 superfamilies and 1,086 folds. Comparinged to its first release, the number of families, superfamilies and SCOP folds have increased 7-fold, 5-fold, and 4-fold, respectively. In addition, a new updating protocol has been incorporated in order to manage the large number of new structures from structural genomics

projects and to enhance the discovery of new and distant relationships. Nearly half of the families and superfamilies contain at least one structural genomics target domain and half of these domains represent a new SCOP family at the time they were first released. These data indicate the impact of structural genomics on the discovery of protein relationships.

1.4 Contributions to SCOP folds by Pfam and Non-Pfam

The number of protein folds in nature is predicted to be limited to somewhere between 1,000-10,000 (Coulson *et al.*, 2002; Koonin *et al.*, 2002; Leonov *et al.*, 2003; Grant *et al.*, 2004). Because of this limitation, the yearly growth of new folds and superfamilies does not occur at the same rate as the growth of new structures. Biased protein target selection could also be a cause for the slow increase. For example, projects focusing on a particular biological pathway or enzymatic activity would mainly contribute homogenous structures from different mutations or binding ligands.

In addition, since Non-Pfam proteins have unique sequences they generally are not included in structural genomics target selection. However, these unique sequences may have new folds. So the question arises: will these non-Pfam proteins provide significant contributions to the discovery of complete protein fold space?

A recent analysis (Che, unpublished data) of contributions to SCOP folds by Pfam and non-Pfam proteins has been carried out. Briefly, a total of 86875 chain sequences from 28945 structures were collected from the PDB (As released of 05/30/2006) and merged as one file in FASTA format. HMMER was used to search against the Pfam 20.0 installed on a local cluster. Protein chains with "No hits found" suggested no significant similar to 8291 HMMs in Pfam version 20.0 and therefore were considered as non-Pfam sequences. Next, representatives for

SCOP folds or superfamilies were harvested based on SCOP release 1.69, divided into non-Pfam and Pfam categories, using HMMER, and grouped into each year of their releases.

Statistics showed that 93.5% and 6.5% chain sequences belonged to Pfam-A and non-Pfam respectively, while their population accounted for 74% and 26% in Pfam version 20.0, respectively. Less than 10% of the annually deposited structures in the PDB correspond to non-Pfam proteins (Figure 1.2). However, SCOP folds contributed by non-Pfam targets (118) accounts for 13.1% of total number. About 2.7% of non-Pfam domains represent a new SCOP folds while the rate is 1.3% for Pfam domains. The annual contributions to new SCOP folds and superfamilies by non-Pfam proteins are much higher than their contribution to PDB targets. Additionally, an increasing trend of contribution has been observed in the recent years (Figure 1.3).

A statistical conclusion from this study cannot be drawn due to the limited number of non-Pfam entries in the PDB and it is unknown whether non-Pfam proteins share divergent evolutionary links to existing Pfam folds. However, the preliminary analysis has indicated that non-Pfam target selection strategy would likely make a significant contribution to the new fold discovery.

1.5 Specific Aims and Significance of This Work

The current structural genomics target selection strategies based on Pfam-A has effectively increased not only the number of newly solved protein structures, but also the number of new folds and new families. However, restricting targets to Pfam-A sequences will miss a significant portion of sequence/fold space since approximately one-fourth (1 million) of highly-unique sequences will not currently be explored resulting in an incomplete description of protein fold space.

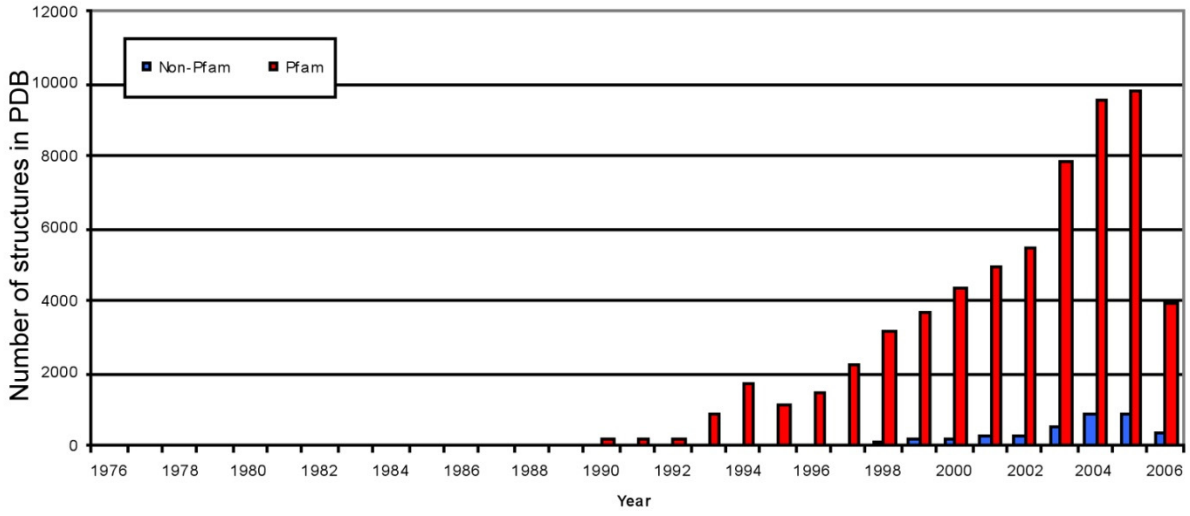


Figure 1.2: Annual contribution of Pfam and non-Pfam targets in PDB. Red: Pfam; blue: non-Pfam. The non-Pfam targets stayed under 10% of the total deposited proteins every year.

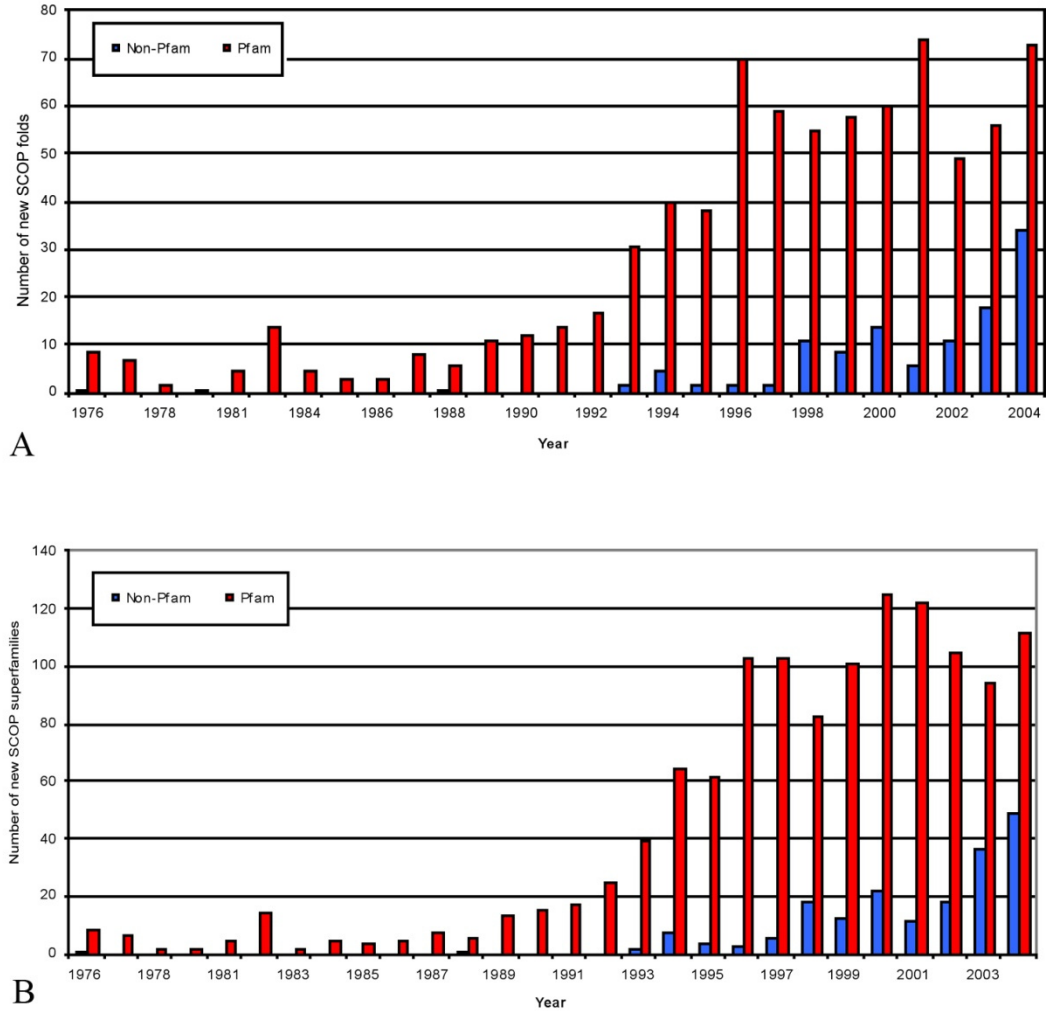


Figure 1.3: Growth of new SCOP folds and superfamilies in the selected time frame. Red: Pfam; blue: non-Pfam. A) Growth of new SCOP folds per year contributed by Pfam and non-Pfam targets. B) Growth of new SCOP superfamilies per year contributed by Pfam and non-Pfam targets. The contribution of new SCOP folds and superfamilies from non-Pfam targets have become significant in the recent years

The work presented here represents a pilot study on applying a structural genomics approach to the characterization of non-Pfam A proteins by:

I) Use high-throughput pipelines at UGA to clone, express, and purify non-Pfam proteins

II) Use X-ray crystallography to determine the structures of non-Pfam proteins and identify new SCOP folds

III) Use bioinformatics to predict possible biological functions for these non-Pfam proteins based on their structure

IV) If possible, use proteins produced by the pilot study to extend the experimental envelope for the sulfur-SAS phasing method to crystals of moderate resolution.

There is concern about potential difficulties in expression and purification of non-Pfam targets (i.e. can real foldable proteins be produced?). High-throughput pipelines have been developed at SECSG and proved to increase success rate of protein production and crystallization (Wang *et al.*, 2005). Using these high-throughput pipelines and salvaging techniques (Liu *et al.*, 2005b), we believe that the success rate of protein production and crystallization for non-Pfam proteins could be similar to that observed for Pfam proteins.

Given their unique sequences, the function of most non-Pfam sequences remains unknown. Thus, the structural characterization of non-Pfam proteins would give the first clues about possible function and form the basis for further functional studies. In addition, non-Pfam sequences may entail an intrinsic phenomenon in evolution, and play important roles in the uniqueness of the organism. Therefore it is wise to consider non-Pfam proteins in a global view of the protein world.

X-ray crystallography and Nuclear Magnetic Resonance are the routine methods for the discovery of 3D structures of macromolecules in structural genomics projects. X-ray crystallography is able to determine protein structures at the atomic level without limitations in

protein size, whereas NMR is less accurate and restricted to relatively small proteins. Thus X-ray crystallography was the technique used to determine non-Pfam protein structures described in this work.

1.6 Exploring SAS Structure Determination Method

In X-ray crystallography, X-rays are diffracted by the well-formed protein crystal lattice. Based on the diffraction pattern an electron density map can be computed (if phases are known) and the protein's amino acid sequence fitted into this map giving the final structure. The electron density at a given point (x, y, z) in the unit cell (here x, y, z are fractional coordinates), can be calculated by the summation shown in Figure 1.4A. Two components are needed in this equation, the amplitude of a structure factor, $|F_{hkl}|$, and the phase difference, α_{hkl}. The structure factor F_{hkl} of a reflection *hkl* is the sum of the atomic scattering factors of all atoms in the unit cell (Figure 1.4B). Its amplitude $|F_{hkl}|$ is proportional to the square root of intensity I_{hkl}, which is measured by the detector during data collection. However, the phase, α_{hkl}, cannot be measured directly from the diffraction data. This problem is generally referred to as the Phase Problem in X-ray crystallography.

To solve this problem, several methods have been developed: direct method, isomorphous replacement, molecular replacement and anomalous scattering.

Direct methods estimate the phases of the Fourier transform of the scattering density from the relationships among the reflections and their intensity magnitudes in the data. The technique is generally applied to determine structures of small molecules or proteins having up to 1000 atoms in the asymmetric unit, but it requires ultra high (< 1 Å) resolution data. For proteins, direct methods are used routinely to find the heavy-atom or anomalous substructure since high resolution data is not required in this case.

A

$$\rho(xyz)=\frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F(hkl)|\cos 2\pi(hx+ky+lz-\alpha_{hkl})$$

B

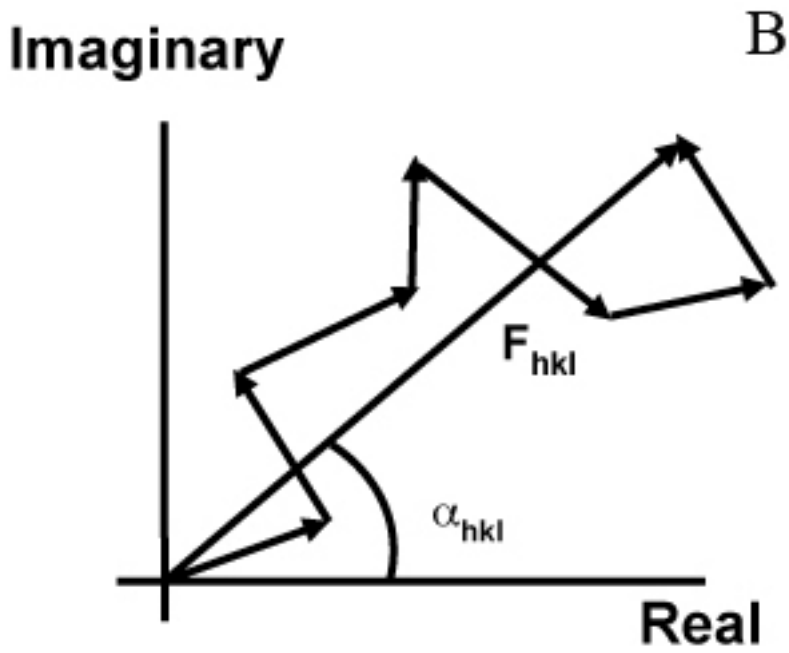


Figure 1.4: The electron density equation and the structure factor. A) The Fourier transform used for calculating electron density (ρ) at any point (x,y,z). V is volume of unit cell; $|F_{hkl}|$ is amplitude of structure factor, and α_{hkl} is phase difference. B) The structure factor, F_{hkl}, shown in vector notation. It is the sum of the atomic scattering factors of all atoms in the unit cell.

In isomorphous replacement, heavy atoms whose Z-numbers are large (e.g. Hg, Au, Pt), are introduced into the unit cell an isomorphous manner. The measured changes of diffraction intensities between the native and derivative data are used to solve the phase problem. The incorporation of heavy atoms can be carried out by soaking the crystal in a heavy atom-containing solution or by co-crystallization. The heavy atom salts used for soaking and co-crystallization are usually hazardous and can be expensive. In some cases heavy atom soaking destroys or severely changes the crystal lattice leading to non-isomorphism and failure of the technique.

Molecular replacement is used when a homologous structure to the target protein (> 30% sequence identity) is available. This technique can not be used for *de novo* structure determination.

Anomalous scattering, in this method in addition to the normally scattered photons and fluorescence emitted at lower energy, some photons can be absorbed and immediately re-emitted at the same energy level though resonance scattering. This naturally occurring phenomenon has long been misinterpreted as "abnormal" or "anomalous" scattering since it was first observed by Bijvoet (Bijvoet, 1954). Anomalous scattering splits the structure factor into two components; the original structure factor, F_{normal}, and the complex anomalous structure factor $F_{anomalous}$ as shown in Figure 1.5A. Here, $\Delta f'$ is the real anomalous component (dispersive term) with a phase of either 0 or 180° and $\Delta f''$ is the imaginary component (absorption term) that is always 90° ahead of the real component. In addition, $\Delta f''$ is a wavelength dependent discontinuous function (usually used to indicate the strength of anomalous scattering) that reaches a maximum value at an absorption edge.

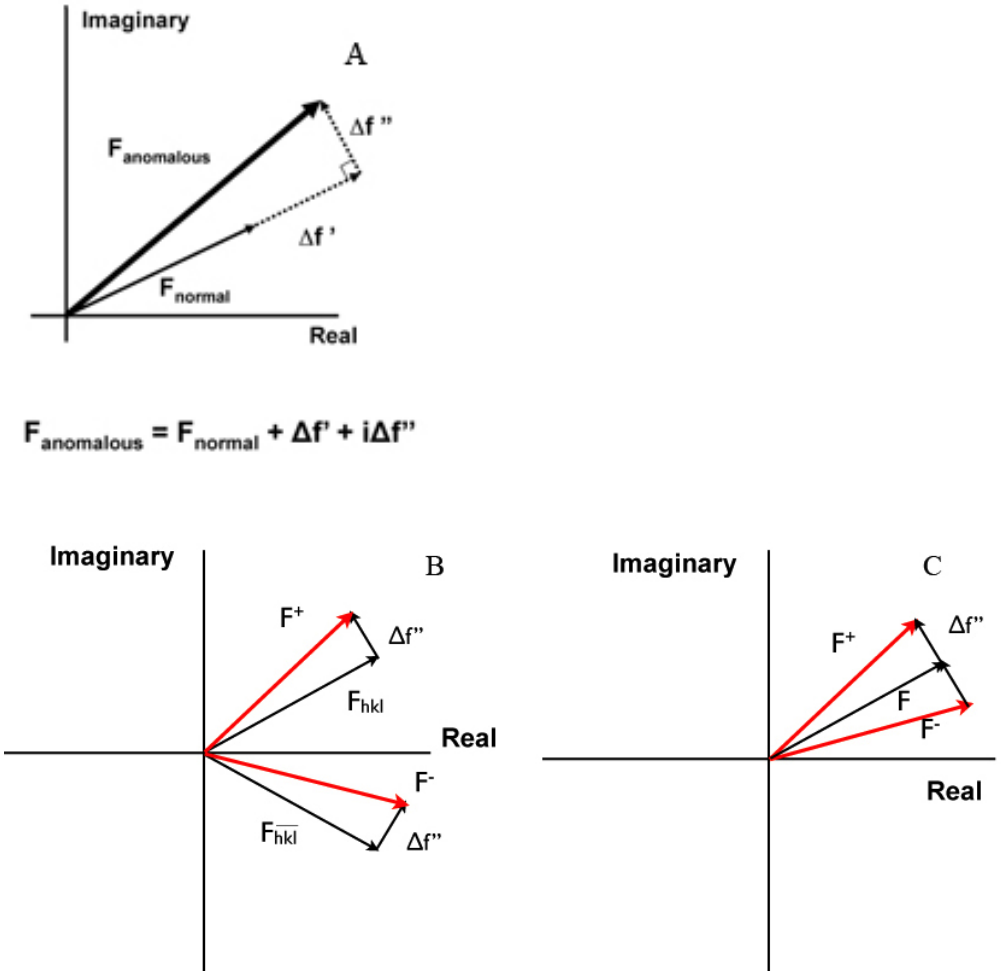


Figure 1.5 Anomalous scattering and breaking of Friedel's law. A) Vector and equation summation of anomalous scattering compared to normal Thompson scattering. Δf' is the real anomalous component with a phase of either 0 or 180° and Δf'' is the imaginary component that is always 90° ahead of the real component. B) The break of Friedel's law by anomalous scattering. New structure factors (F^+ and F^-, shown in red) of a previous Friedel pair (F_{hkl} and $F_{\overline{hkl}}$) formed by anomalous scattering have different phases. C) Phase triangle formed by F^+, inversed F^- and $2\Delta f''$. If F^+ and F^- can be accurately measured, with known Δf'' for each element, the amplitude of F can be calculated.

The observation of wavelength-dependent anomalous scattering made the Multi-wavelength Anomalous Dispersion (MAD) possible. In a MAD experiment, several sets (at least three) of diffraction data are collected at different wavelengths near the absorption edge in order to maximize the absorption and dispersive effects (Hendrickson *et al.*, 1990). Typically, wavelengths are chosen at the peak of $\Delta f''$, at the point of inflection on the absorption curve where the dispersive term reaches its lowest value, and at a remote wavelength with almost no anomalous scattering. The structure factors from peak data and inflection data can be thought to serve as derivative data in the MIR (multiple isomorphous replacement) experiment with the remote data serving as native data. This method is limited to proteins/crystals containing ordered anomalous scattering atoms whose absorption edge is within tunable synchrotron X-ray radiation such as seleno-methionine (SeMet) labeled proteins.

Single wavelength anomalous scattering is a subset of the MAD method requiring that only the peak ($\Delta f''$ = maximum) data be collected. Since in anomalous scattering data Friedel's law (Friedel, 1913), which states that $|F_{hkl}| = |F_{-h-k-l}|$ is no longer valid due to the phase shift from $\Delta f''$ (Figure 1.5B). The difference in F^+ and its inverse F^- is $2\Delta f''$, as indicated in the "phase triangle" shown in Figure 1.5C. The values of F^+ and F^- can be measured in the single-wavelength scattering experiments and $\Delta f''$ recorded for each element, we may obtain two solutions of $F_{protein}$ (Figure 1.6), with the same amplitude but different phases. Thus the central point in SAS experiments is to "find the orientation of a phase triangle from one of its side" (Wang, 1993-2001).

In 1985, B. C. Wang developed the Iterative Single-wavelength Anomalous Scattering (ISAS) method, part of the ISIR/ISAS program system, to break the phase ambiguity in SAS phasing (Wang, 1985). The philosophy of the ISAS method (Figure 1.6) is to consider the SAS

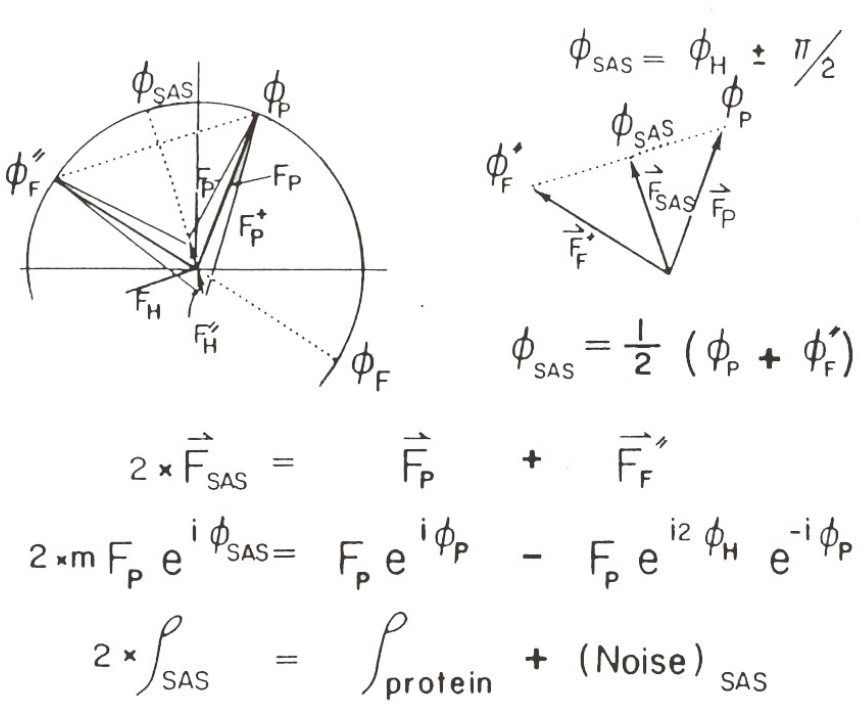


Figure 1.6 An SAS map can be considered as a superposition of two maps. The Harker construction for phase calculation by SAS shows that the structure factor of SAS (F_{SAS}) can be considered as the sum of structure factors from two possible solutions of protein phase, while F_p stands for the correct phase and F_F stands for the false phase. The Fourier map produced by the false phases contains no structural information and shows as a general background. When the background noises are gradually filtered from the electron density map, the SAS map will become close to the protein map. Adapted from: Wang, B. C. (1985). Methods Enzymol 115, 90-112.

electron density map as the superposition of two maps contributed by the protein and the noise, respectively. When the noise is filtered out in direct space, an enhanced map from the protein could be used as a partial structure to resolve the phase ambiguity in turn.

A flowchart of the ISAS method is shown in Figure 1.7. For a given set of SAS data, an electron density map is calculated from the SAS phase, α_{SAS} (Figure 1.6). When the projection of $F_{protein}$ is in the same direction as F_{SAS}, α_{SAS} will be $\alpha_H + \pi/2$, in which α_H.is the phase of anomalous scatterer. When the projection of $F_{protein}$ is in the opposite direction to F_{SAS}, α_{SAS} will be $\alpha_H - \pi/2$. Then the molecular boundary is located by summation of the density around each grid point. Normally proteins will have higher strength of signal than the solvent so a mask can be automatically calculated. Outside the molecular boundary, the electron density is flattened to a new lower value. Inside the protein region, the remaining negative density after adding of a constant density is removed. After filtering of almost all the noise in the solvent region and part of the noise in the protein region, the enhanced map is inverse Fourier transformed to calculated phases, which are then combined with the original phases to create improved phases. The phase filtering process is a modification of phase probability between the original phases and calculated phases. Improved phases could be used to start a new cycle without creating another electron density mask. The iterative cycling is usually carried out four to eight times and followed by a phase extension. The final phases are used to calculate the final electron density map for the protein.

The ISAS method can be considered as a signal improvement process through error reduction. It requires accurate measurements in the experiments, as the anomalous signal is about one order of magnitude smaller than the difference caused by isomorphous replacement. Thus, to measure the anomalous scattering signal one could either design better instruments,

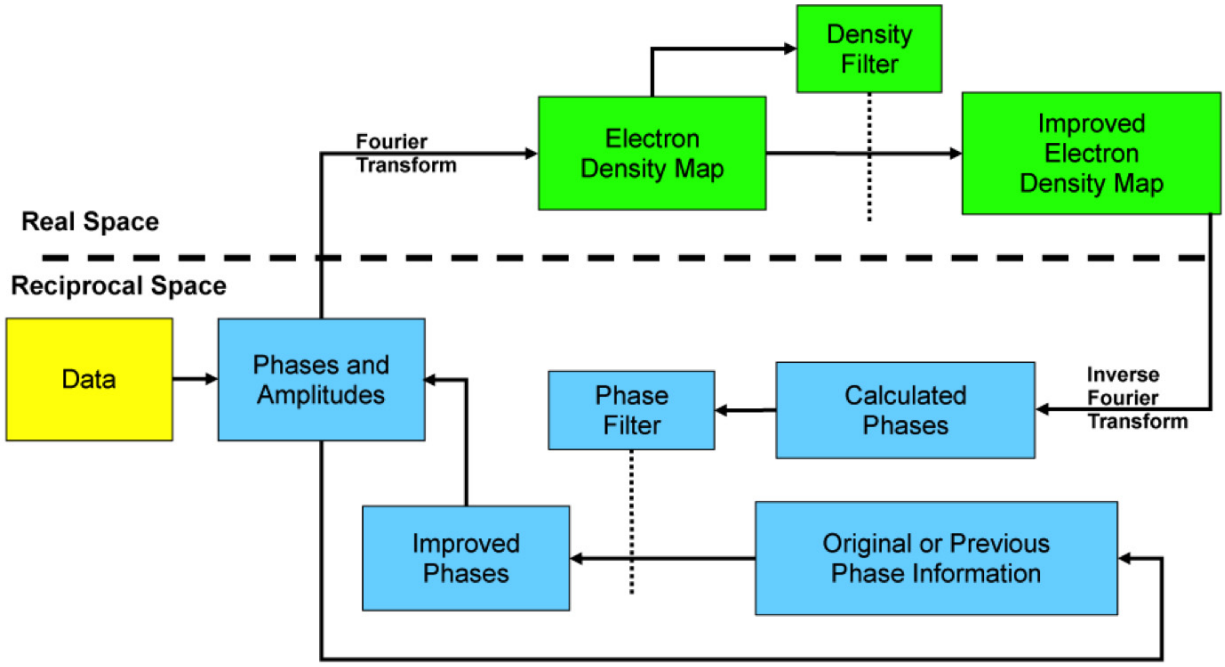


Figure 1.7 The ISAS flow-chart. Phases and amplitudes calculated from the original data are Fourier transformed to electron density. In real space, the electron density map is improved by a density filter and inverse Fourier transformed to calculated phases. A phase filter is used to average and combine the calculated and the original phases. The improved phases are used to calculate electron density for the next cycle. The iterative cycling is usually carried out four to eight times. The final phases are used to calculate the final electron density map for the protein. Adapted from: Wang, B. C. (1985). Methods Enzymol 115, 90-112.

technology and experimental strategy to reduce noise, use metalloproteins having Fe, Ni, etc, which are good anomalous scatterers, or introduce strong anomalous scatterers (e.g. SeMet) into protein/crystal. However, there is a limited number of metalloproteins and incorporation of SeMet may prove to be problematic for certain expression systems (e.g. eukaryotic systems) and may change the characteristic of the proteins. Today, with developments in detector technology and data collection and processing methodology the SAS method is being extended to the anomalous scattering from sulfur atoms (S-SAS phasing), which are found in the methionine and cystine residues that occur in most natural proteins.

1.7 History of Sulfur-SAS Phasing

The absorption edge of sulfur lies at 5.02 Å, which is not practically achievable at most synchrotron protein crystallography beamlines. In addition, a MAD experiment was carried out using ~5Å X-rays, if it were possible, would face problems such as a dramatic loss of beam intensity at this wavelength and severe air absorption effects. For this reason, single wavelength data collection carried out at a much shorter wavelength (less than 2.5 Å) are necessary for sulfur-SAS phasing and structure determination.

In 1981, Hendrickson and Teeter phased the structure of crambin, a 45-residue protein, using anomalous scattering signal from six sulfur atoms (Hendrickson *et al.*, 1981). Their approach is a statistics method assuming that the protein phases can be approximated as the phases of the anomalous scatterers. Thus it requires a significantly high anomalous sulfur scattering contribution of magnitude to the overall structure factor. This method cannot resolve a structure when it is applied to larger proteins that usually have much less sulfur content.

In 1985, B. C. Wang introduced the ISAS method, which allows us to determine protein phases using only one set of anomalous scattering data. Wang also showed through computer

simulation the structure of the 12 kDa protein Rhe could be successfully phased based on the anomalous scattering from a single disulfide bond. The simulation showed remarkably enhanced density map and predicted that structure could be phased by anomalous scattering signal from only two sulfur atoms ($\Delta f'' = 0.56\ e^-$) in the sequence using 1.54 Å X-rays (1 sulfur per 57 residues). Although theoretically applicable, experimental sulfur SAS phasing could be realized at that time since the instruments and techniques were not able to accurately measure the small difference in intensities.

In 2000, the development in X-ray and computer technology finally made sulfur-SAS phasing a viable choice when Liu *et al.* solved the *de novo* structure of the 22 kDa protein obelin using anomalous signal from 8 sulfur atoms and a 1.7 Å set of anomalous scattering data recorded using 1.74 Å synchrotron X-rays (Liu *et al.*, 2000). To date, 38 structures in PDB have been solved by sulfur-SAS phasing. Although it is not a routine phasing method sulfur-SAS phasing has proved its phasing power and should play an ever increasing role as technology and data collection methodology continue to improve.

1.8 Applications of Sulfur-SAS Phasing at UGA and SER-CAT

The relatively weak anomalous signal from sulfur makes the signal-to-noise a critical factor in determining the success of phasing. The data collection experiment must be carefully designed since both the sulfur anomalous scattering signal and X-ray absorption will increase with the increasing X-ray wavelength. The two major types of X-ray sources are in-house sources, which produce X-ray radiation by accelerating electrons at high voltage against a metal target; and synchrotron sources, which produce X-ray radiation when electrons change direction as they orbit in the synchrotron ring. Wavelengths of X-rays generated from a rotating anode generator (0.71069, 1.5418 and 2.2909 Å) or synchrotrons (normally in the tunable range

between 0.8-2.0 Å) are very much below the absorption edge of sulfur. Therefore, an optimal wavelength should be chosen as the compromise of the sulfur anomalous signal and the X-ray absorption effects.

At UGA, a modified X-ray source system has been installed, where chromium $K\alpha$ X-Rays (λ = 2.2909 Å) are generated by a RU-H3R generator (operating at 45 kV 90 mA) containing a chromium target and focused on the crystal through an Osmic CMF 15-50Cr8 optic. Reflections are recorded on a modified large-aperture R-AXIS IV detector (Rigaku) with helium-flushed beam path and modified beam stop (Yang *et al.*, 2003; Xu *et al.*, 2005). The helium path is used to minimize the absorption effects caused by air for the diffracted X-rays. Using Cr $K\alpha$ X-rays the anomalous scattering signal ($\Delta f''$) of sulfur is 1.14 e^-, almost double the value of 0.54 e^- using Cu $K\alpha$ X-rays (λ = 1.5418 Å). This modified home source is convenient to use and provides a very stable X-ray beam at longer wavelengths. However, a typical 360° data set takes 24 hours or more to collect depending on the crystal quality and X-ray exposure time. A total of 8 crystal structures have been solved by sulfur-SAS phasing using diffraction data collected on the UGA system. Among them, the structure of 150 residue protein Pfu-542154 which was phased from the anomalous scattering signal measured from 3 sulfur atoms from and approaches the limit for the 1985 Rhe protein of one sulfur per 57 residues used in Wang's 1985 simulations.

At a synchrotron source, X-rays are generated by changing the direction of electrons with almost the speed of light using bending magnets or insertion devices in the storage rings. The most important characteristics of synchrotron radiation are the substantial improvements to the quality and the brightness of the X-ray beams. Because of the tunable and extraordinarily bright X-ray beam they produce, synchrotron sources, which are widely used in MAD phasing, can also be used for sulfur-SAS phasing experiments.

The Advanced Photon Source (APS) at Argonne National Laboratory (ANL) is a third generation synchrotron light source, which generates the most brilliant X-ray beams in the United States of America. Electrons are accelerated inside a booster synchrotron to 7 GeV before they are injected into the storage ring. The storage ring is 1,104 meters in circumference and consists of 40 sectors. To date, 34 sectors have been built at APS focused on Chemistry, Environmental Science, GeoScience, Life Science, Materials Science, Physics, and Polymer Science.

The Southeast Regional Collaborative Access Team (SER-CAT) was founded in 1997. It was originally formed to benefit macro-molecular crystallographers and structural biologists in the southeastern region of USA and now consists of 25 member institutions. SER-CAT operates two beamlines at APS, ANL: the 22-ID insertion-device beamline and the 22-BM bending-magnet beamline (Table 1.1). SER-CAT supports research in all aspects of X-ray structural biology from new structures to drug design. The 22-ID beamline uses a Si (220) double crystal monochromator and has a full beam flux capacity at 7×10^{12} ph/s. It provides X-rays with energy range of 6–24 keV (λ = 2.06 – 0.52 Å). A Mar 300 CCD detector (a 300 mm × 300 mm 4X4 array of 16 CCD chips) is installed a Rosenbaum A-frame. The 22-BM beamline uses a Si (111) double crystal monochromator and has a full beam flux capacity at 2×10^{11} ph/s. Its X-ray energy range is 7–17.5 keV (λ = 1.77 – 0.71 Å). A Mar 225 CCD detector a 225 mm × 225 mm 3X3 array of 9 CCD chips) is installed on a Rosenbaum A-frame. Both beamlines include an improved ALS/Berkeley style crystal automounter, which is an essential component of the remote-control system. Using the remote-control system, users can screen and collect X-ray diffractions without traveling to the beamline. The automounter's Dewars allow for 230 (22ID)

Table 1.1 SER-CAT beamline parameters.

	22ID	**22BM**
Full Beam Flux (ph/s)	7×10^{12}	2×10^{11}
Energy bandwidth (eV)	0.5	3–4
Focused Beam Size (μm)	120 × 100	90 × 90*
Delivered Flux (ph/s)		
100 by 100 micron	4×10^{12}	1×10^{11}
50 by 50 micron	1×10^{12}	4×10^{10}
20 by 20 micron	3×10^{11}	1×10^{10}
Energy Range (keV)	6–24	7–17.5
Detector	Mar 300 CCD	Mar 225 CCD
Sample Changer	Yes	Yes
Dewar Capacity (samples)	230	96
Remote Access	Yes	Yes

and 96 (22BM) crystals to be placed in the hutch at one time significantly increasing data collection throughput.

Initial tests in 2003 showed that the quality of longer wavelength (> 1.5 Å) data collected using the 22-ID beamline using was inferior to the data collected at UGA using the chromium in-house source. Two problems with the 22ID design were identified: radiant heating within the monochromator and severe 2nd harmonic contamination from Si (220) first crystal. Improved shielding of 1st and 2nd crystals was applied to reduce radiant heat effects and an improved 2nd crystal design provided better heat transfer during experimentations. In addition, during this time a white beam adjustable aperture was added in front of the monochromator to protect the monochromator from increased heat load caused by a planned increase in APS ring current. It slits the system upstream of the monochromator so that the monochromator power load can be lowered. The adjustable aperture can also be used to solve the harmonic contamination problem. Normally, the upstream aperture's slits are set at 2mm in the vertical and 4mm in the horizontal, which allows the full beam to pass. For sulfur-SAS phasing data collection, the upstream slits are reduced to either 0.5 by 0.5 mm or 0.75 by 0.75 mm. After all these measures were employed, another round of test indicated significant improvements in both data quality and beam stability during low energy experiments.

The current design of X-ray data collection setup at 22-ID beamline (Figure 1.8) involves a helium beam path with Kapton window installed on the MAR 300 CCD area detector. The beam stop is installed on the center area of the window so that an automounter can be used to mount crystals during experiments. This restricts the minimum crystal-to-detector distance to 125 mm. Using 1.9 Å X-rays, 2.5 and 2.2 Å resolution data can be recorded at the edge and corner of detector, respectively, and is sufficient for automated map tracing.

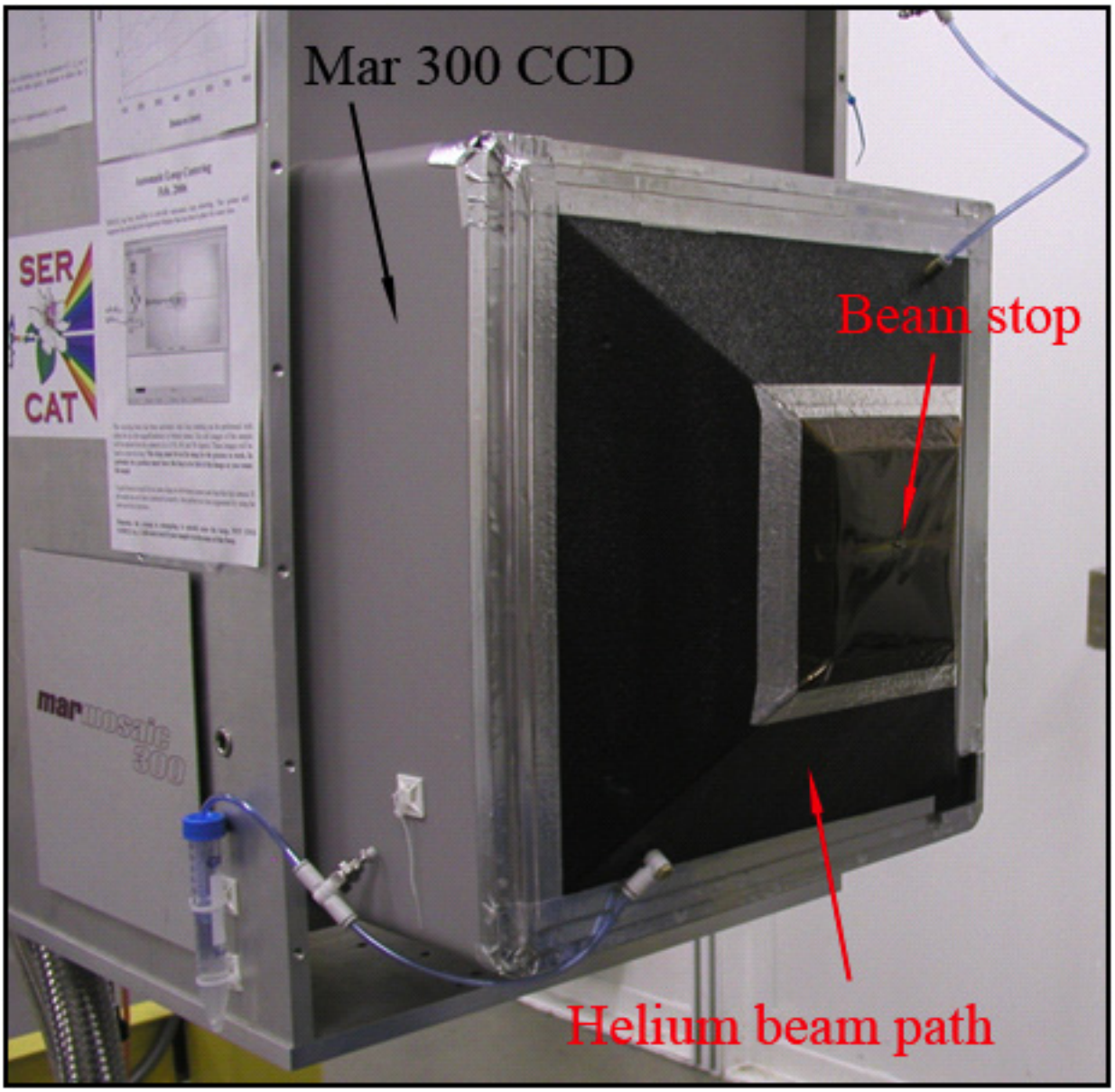


Figure 1.8 The current design of X-ray data collection setup at 22-ID beamline (SER-CAT) for sulfur-SAS phasing experiments.

To directly monitor anomalous scattering signal in diffraction data, a new statistical measure, R_{as} was put forth by Fu et al (Fu *et al.*, 2004). R_{as} is calculated by the ratio of Δa and Δc (Figure 1.9). Operation Δ, defined as the average ratio of Bijvoet differences in intensity and standard deviation of intensity, is a measure of average Bijvoet differences in diffraction data. Δa, calculated using acentric reflections, represents the level of signal and noise, while Δc, calculated using centric reflections, only represents noise level. Data with R_{as} less than 1 do not contain anomalous scattering signal and thus could not be used for phasing. Analysis of data from the structures previously solved by sulfur phasing revealed that a minimum R_{as} of 1.5 at 3 Å resolution should be expected for a successful structure determination by sulfur atoms. With this statistic, it would be possible to monitor anomalous signal while the crystal is still mounted and collecting, to see whether severe radiation damage or other error has outnumbered the anomalous signal. A reasonable data processing strategy could also be decided based on R_{as}.

The in-house chromium source at UGA is now generally used as a test bed for improving low energy data collection at SER-CAT. High quality data for both phasing and refinement are usually collected at SER-CAT to ensure the success of structure determination. As a result of experiments carried out using the UGA chromium X-ray test bed and SER-CAT's response to these tests that has significantly increased beam stability at lower energies the first S-SAS structure was determined at SET-CAT in early 2008 on a well diffracting crystal. This work presents the second structure recently solved at SER-CAT on a medium diffracting crystal and indicates that sulfur-SAS could become a routine phasing method at SER-CAT.

$$R_{as} = \frac{anomalous\ signal}{noise} = \frac{\Delta a}{\Delta c} = \frac{\left\langle \frac{|I_a|}{\sigma_I} \right\rangle}{\left\langle \frac{|I_c|}{\sigma_I} \right\rangle} = \frac{\frac{\sum_a |I_{+(a)} - I_{-(a)}|}{\sigma_{I(a)}}}{\frac{\sum_c |I_{+(c)} - I_{-(c)}|}{\sigma_{I(c)}}}$$

Figure 1.9 Definition of R_{as}. R_{as} is calculated by the ratio of Δa and Δc. Δa, calculated using acentric reflections, represents the level of signal and noise, while Δc, calculated using centric reflections, only represents noise level. Adapted from: Fu, Z. Q., J. P. Rose and B. C. Wang (2004). Acta Crystallogr D **60**: 499-506.

CHAPTER TWO

Methods and Materials

2.1 Target Selection

Six prokaryotic genomes were selected in our research, including *Aeropyrum pernix* K1, *Archaeoglobus fulgidus*, *Clostridium thermocellum*, *Pyrococcus furiosus*, *Pyrococcus horikoshii*, and *Thermus thermophilus*. Pfam 20.0 server was installed on an IBM cluster machine, which contains 64 nodes and 128 CPUs. The ORF sequences from those six genomes were downloaded from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, and candidate non-Pfam targets were selected by searching against Pfam 20.0. Using all PDB sequences as a database for a BLAST search, the sequences whose expectation values are less than 0.1 were excluded. Next, the predicted trans-membrane protein sequences were also excluded from remaining targets using TMHMM server 2.0 (Krogh *et al.*, 2001). The resulting ORFs (328 sequences from *Aeropyrum pernix*, 223 sequences from *Clostridium thermocellum,* 268 sequences from *Archaeoglobus fulgidus,* 205 sequences from *Pyrococcus horikoshii,* and 261 sequences from *Thermus thermophilus*) formed the target pool for our non-Pfam structural genomics studies.

2.2 Gene Cloning and Expression Screen

Genes of the previously selected targets were amplified by PCR using genomic DNA as the templates. After the PCR products were confirmed by agarose gel analysis, the genes of interest were cloned into appropriate expression vectors by Gateway® system (Invitrogen), a site-specific recombination-based cloning system due to its flexibility and simplicity.

A variety of N-terminal tags were incorporated in the vectors to improve expression and solubility of target proteins or to aid in the purification of the protein (Stevens, 2000; Terpe, 2003). For this work, a His_6 (e.g. HHHHHHENLYFQGGSG) purification tag was used to allow for metal affinity purification of the protein and an MBP (maltose binding protein) tag was used for to increase protein solubility during expression. To avoid altering the properties of the target proteins, these tags were generally removed during the purification process. Therefore, a TEV (Tobacco Etch virus) protease cleavage site (ENLYFQGGSG) was inserted between the tag and the protein sequence in the construct.

The expression vector containing the target gene was then transformed into the appropriate host cell line. Bacterial cells such as *E. coli* were chosen as a host for its high efficiency, low cost and well established protocols. All cloned products were validated by DNA sequencing (Sequencing and Synthesis Facility of the University of Georgia, http://www.ssf.uga.edu/).

In the expression screen, cells producing native proteins were grown in 5 mL Luria-Bertani (LB) media (Table 2.1) and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at OD_{600} ~0.5. Cells were collected after 3 hours following the IPTG induction. Alternatively, cells producing SeMet-labeled proteins were first grown in 5 mL PA0.5G media (Table 2.1) and transferred into 20 mL PASM5052 media (Table 2.1) in the presence of 125 μg/mL SeMet after 8 hours. Cells were harvested after 20 hours growth in PASM5052 media. The whole-cell expression results were analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

Table 2.1 Compositions of LB, PA0.5G and PASM5052 media.

LB	
Total Volume	1 L
Bacto Tryptone	10.0 g
Yeast Extract	5.0 g
NaCl	5.0 g

PA-0.5G		PASM-5052	
Total Volume	50 ml	Total Volume	1 L
H_2O	46.13 ml	H_2O	900 ml
$MgSO_4$ (1 M)	50 µl	$MgSO_4$ (1 M)	1.25 ml
1000X Metal Mix	5 µl	1000X Metal Mix	1.25 ml
40% Glucose	0.63 ml	50X 5052	20 ml
20X NPS	2.5 ml	20X NPS	50 ml
Methionine (25 mg/ml)	0.2 ml	vitamin B12 (100 µM)	1 ml
17aa (CYM)	0.5 ml	17aa (CYM)	20 ml
		Methionine 25mg/ml	400 µl
		Se-Met 25mg/ml	5 ml

20X NPS	
Total Volume	1 L
H_2O	900 ml
$(NH_4)_2SO_4$	66 g
KH_2PO_4	136 g
Na_2HPO_4	142 g

50X 5052	
Total Volume	1 L
H_2O	730 ml
Glycerol	250 g
Glucose	25 g
α -lactose	100 g

1000X Metal Mix	
Total Volume	100 ml
sterile H_2O	36 ml
$FeCl_3 \cdot 6H_2O$ (0.1 M)	50 ml
$CaCl_2$ (1 M)	2 ml
$MnCl_2 \cdot 4H_2O$ (1 M)	1 ml
$ZnSO_4 \cdot 7H_2O$ (1 M)	1 ml
$CoCl_2 \cdot 6H_2O$ (0.2 M)	1 ml
$CuCl_2 \cdot 2H_2O$ (0.1 M)	2 ml
$NiCl_2 \cdot 6H_2O$ (0.2 M)	1 ml
$Na_2MoO_4 \cdot 5H_2O$ (0.1 M)	2 ml
$Na_2SeO_3 \cdot 5H_2O$ (0.1 M)	2 ml
H_3BO_3 (0.1 M)	2 ml

17aa (CYM)	
Total Volume	100 ml
H_2O	36 ml
Na·Glutamic acid	1 g
Aspartic acid	1 g
Lysine	1 g
Arginine·HCl	1 g
Histindine·HCl	1 g
Alanine	1 g
Proline	1 g
Glycine	1 g
Threonine	1 g
Serine	1 g
Glutamine	1 g
Asparagine	1 g
Valine	1 g
Leucine	1 g
Isoleucine	1 g
Phenylalanine	1 g
Tryptophan	1 g

2.3 Protein Expression

Cell cultures containing target proteins confirmed to be expressed from the small-scale screens were scaled up for protein production. Two types of media were used for expression depending on whether native (LB medium) or SeMET labeled proteins (auto-inducing medium (Studier, 2005)) were being produced.

For expression in the LB media, cell culture was incubated in 50 mL of LB media with appropriate concentration of antibiotics at 37 ºC for 8 hours. The culture was then added to 1 L of LB media with appropriate concentration of antibiotics. After induction with 0.5 mM IPTG at OD_{600} ~ 0.5 the 1 L culture was incubated for and additional 3 hrs at 25 ºC, 20 ºC, or 15 ºC depending on the protein, since it has been shown that expression at lower temperatures results in better protein stability, solubility, and folding (Chesshyre *et al.*, 1989; Georgiou *et al.*, 1996).

For expression in the auto-inducing media, cell culture was incubated in 50 mL of PA0.5G media with appropriate concentration of antibiotics at 37 ºC for 8 hours. The culture was then added to 1 L of the PSAM5052 media in the presence of 125 μg/ml SeMet with appropriate concentration of antibiotics. After OD_{600} reaches 0.5, the 1 L culture was incubated at a lower temperature for a contined growth for 18 to 20 hours.

The cells from the large-scale expression were harvested by centrifugation at 6000 rpm for 15 minutes and stored at -80 ºC till next step of purification. Results of the large-scale protein expression were also analyzed by SDS-PAGE.

2.4 Protein Purification

Protein purity is one of the most important factors that affect protein crystallization. To achieve the highest level of purity and homogeneity from the protein purification, a three-stage

(metal affinity - ion-exchange - gel filtration) chromatography process (Liu *et al.*, 2005b) was designed and carried out on ÄKTAprime (GE Healthcare).

For purification, the cell pellets were re-suspended in 25 mL Ni affinity binding buffer containing 20 mM sodium phosphate, 0.2-1.2 M sodium chloride, pH 7.6, and lysed by ultrasound (6 × 30 s, on ice) in the presence of 5 mM β-mercaptoethanol and 1 mM phenylmethylsulphonyl fluoride (PMSF). Samples were centrifuged at 12,000 rpm for 30 minutes at 4 ºC to remove cell debris and the supernatant was recovered for further purification.

Based on its high specificity, metal affinity chromatography is one of the most common methods of protein purification and usually chosen as the first step in the purification process. Briefly, the Ni affinity column is charged with nickel ions that are immobilized by the stationary phase of the column. The solution containing the protein of interest (with a His_6 Ni affinity tag) is then applied to the column here the protein is then retained in the column by the interaction of the His_6 tag and the immobilized Ni ions and not tagged proteins eluted. An imidazole gradient is then used to elute the tagged protein. For these studies, the supernatant of the lysed sample was loaded onto a 5ml HiTrap™ Chelating HP Column (GE Healthcare, Ni-affinity) using a flow rate of 1 mL/min. Then the unbound contaminants were removed from the column by washing with 10 column volumes of Ni affinity binding buffer (see above) using a flow rate of 5 ml/min. Next, a 50-mL linear gradient (0 to 0.5 M imidazole, flow rate 1 mL/min) was used to elute the his-tagged protein. The purity of eluted peak fractions was analyzed by SDS-PAGE.

All targets from the genomes other than *Pyrococcus furiosus* also underwent TEV protease cleavage reaction after Ni-affinity purification in order to remove the fusion tags required for expression. Here, protein samples from the 1 L expressions were mixed with 1-1.5 mL TEV protease (unknown concentration, purified by H. Xu) and dialyzed against the Ni

affinity binding buffer. The reaction carried out for 24 h at room temperature or 48 h at 4 ºC, depending on the thermal stability of target proteins. All cleavage reactions were monitored by SDS-PAGE.

In addition, proteins from hyperthermophilic or thermophilic were incubated at 65 °C for 1 h after TEV cleavage to denature and remove additional contaminants including TEV protease. The samples were then centrifuged and filtered. A second round of Ni affinity chromatography separated the cleaved protein and tags with the target protein collected in the initial flow through. The purity of the flow through and the elution peak fractions was analyzed by SDS-PAGE.

Generally, the next step of the purification process involves ion-exchange chromatography, either cation or anion exchange depending on the target. Ion exchange separates proteins based on differences in their net surface charge. In the ion exchange experiment, proteins bound to oppositely charged groups on the column matrix at low ionic strength and eluted using a high ionic strength buffer, for example, sodium chloride up to 1 M in concentration. For improved results, the differences between buffer pH and the pI of target proteins should be greater than 1 pH unit. Due to the high ionic strength of the buffer used for Ni affinity chromatography, the protein samples from the previous purification step were dialyzed overnight against the ion-exchange loading buffer (25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 25 mM sodium chloride, 5 mM ethylenedinitrilotetraacetic acid (EDTA), 1 mM dithiothreitol (DTT)) with appropriate pH. The dialyzed samples were then loaded onto a 5 mL HiTrap Q- (anion-exchange) or SP-Sepharose (cation-exchange) column (GE Healthcare) using a flow rate of 1 mL/min. Next, the unbound contaminants were removed from the column by washing (5 mL/min)it with 10 column volumes of ion-exchange loading buffer. Finally, gradient elution at 1 mL/min with increasing sodium

chloride concentration (0 to 1 M) was used to elute the desired proteins. The purity of elution peak fractions was analyzed by SDS-PAGE.

In gel filtration (size exclusion) chromatography, smaller molecules are able enter into the porous gel matrix and thus are retarded as they move through the column. In contrast, larger molecules that cannot enter the porous gel matrix are not retarded (or retarded as much) so they quickly move through the column. By this mechanism, gel filtration chromatography separates proteins of different sizes and shapes, even different polymerization states of same protein. Generally, 50 to 100 mM sodium chloride is added to the buffer, to provide a suitable ionic strength to avoid hydrophobic interactions between the packing material and the protein molecules. In spite of the advantages, such as high-resolution separation and improved homogeneity of proteins, gel filtration chromatography is time-consuming and may result in sample loss. Therefore, for this work, it was only used as in cases where the purity of the samples from the initial purification required improvement.

For samples that required gel filtration, the proteins were first concentrated to a 3-5 mL volume. The concentrated samples were then loaded to a Superdex 75 or 200 column (GE Healthcare) with a flow rate of 1 mL/min. Proteins and contaminants were separated by (washing 0.5-1 mL/min) with a gel filtration buffer containing 25 mM HEPES, 100 mM NaCl, 5 mM EDTA, 1 mM DTT and 5% (v/v) glycerol. The separation range of Superdex 75 and 200 columns are 3,000-70,000 and 10,000-600,000 Da, respectively. The purity of all peak fractions was analyzed by SDS-PAGE.

Once high purity (> 95%) protein was achieved, the protein sample was concentrated to at least 10 mg/mL and used to set up the initial crystallization trials. Protein concentration was measured and calculated by their UV absorbance at 280 nm. In addition, samples containing

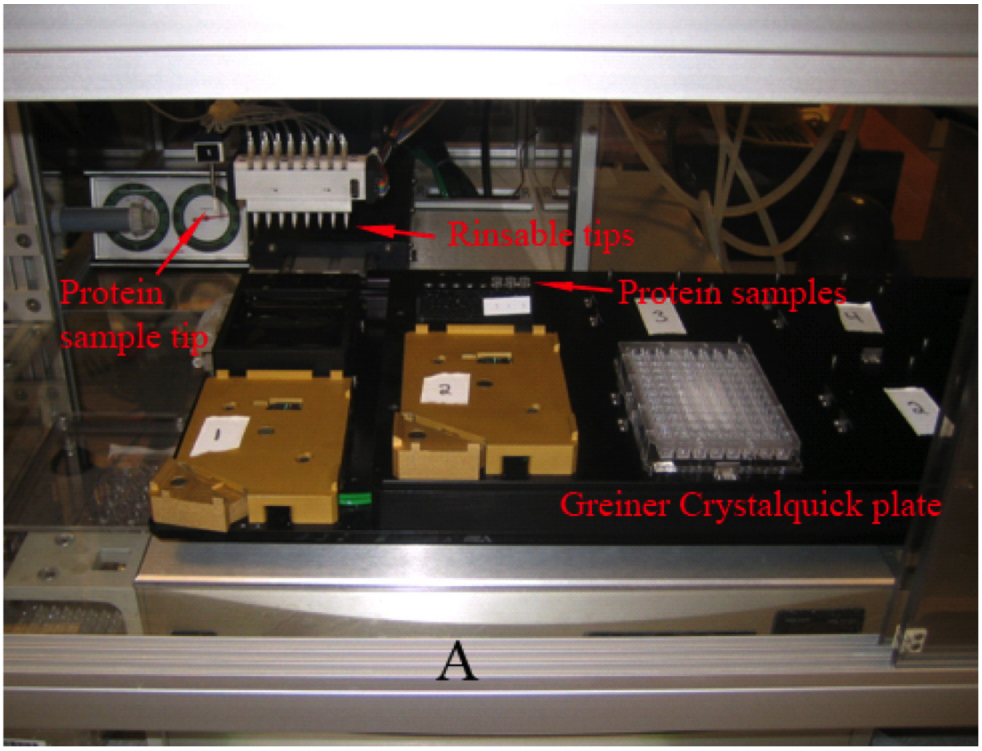
phosphate buffer were dialyzed against the gel filtration buffer (see above) before concentration since phosphate salts can easily crystallize and confuse the results of the crystallization screen.

2.5 Crystallization

For each purified protein sample, an initial crystallization screen was set up to identify conditions that produced crystals. Parameters analyzed in the initial screen including: pH, precipitant, salt, temperature and set-up technique. The initial screen consisted of 384 conditions from 7 commercial screening kits [Crystal Screen I, Crystal Screen II, MemFac, PEG/Ion, Crystal Screen Cryo (Hampton Research), Wizard I and II (Emerald Biosystem)] and one locally designed screen (Shah *et al.*, 2005). The initial screens were setup using either a Cartesian Honeybee crystallization robot (Genomic Solutions) or an ORYX6 crystallization robot (Douglas Instruments) as described below.

The Cartesian Honeybee was used for screening by the sitting drop vapor diffusion method using three well Crystalquick plates (Hampton Research) (Figure 2.1A). As shown in Figure 2.1B, in sitting drop vapor diffusion method, a 0.2 to 2 micro-liter drop containing a mixture of equal volumes of the protein solution (~10 mg/mL) and screening solution is placed in a sealed system together with a reservoir containing the screening solution. Since the concentration of screening solution in the drop is lower than the concentration in the reservoir (due to mixing) water vaporizes from the drop and condenses in the reservoir over time to achieve equilibrium. As water leaves the drop, the protein in the drop is concentrated to the point where the protein solution becomes supersaturated and crystallization may occur.

In practice, the Crystalquick plates were prepared from stock using a Genesis RSP robot (Tecan) where each of the 96 reservoirs in the plate were filled with 100 µl of screening solution (one solution per well). The four plates that comprised the screen were then loaded into the



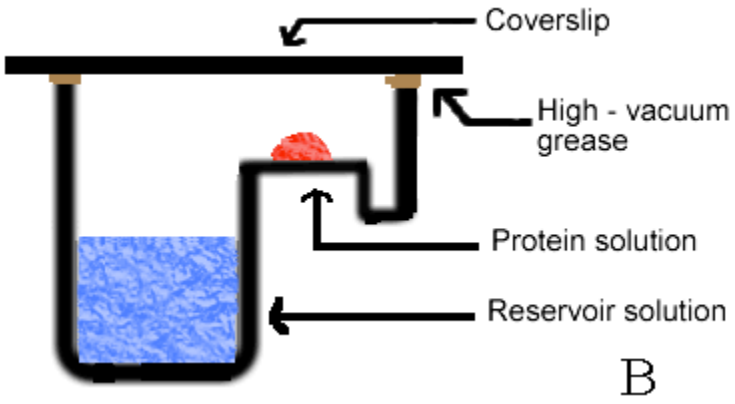


Figure 2.1 Sitting drop vapor diffusion method set up by the Cartesian Honeybee crystallization robot. A) The Cartesian Honeybee crystallization robot. B) Diagram of sitting drop vapor diffusion method. Adapted from http://www.dt.k.u-tokyo.ac.jp/research/PC_2.files/image002.gif.

Honeybee robot where 200 nano-liters of each screening solution were dispensed into their respective well. Finally, 200 nano-liters of the protein solution were added to each well forming a 400 nano-liters sitting drop. The design of the Crystalquick plate allows one to screen three independent samples or same sample with three different concentrations at one time. A complete 384 well screen can be setup in under an hour.

The ORYX6 robot was used to screen for conditions that produce crystals using the modified micro batch under oil method (Chayen *et al.*, 1990). In addition, the ORYX6 was used for crystal optimization and additive screening. As illustrated in Figure 2.2A, 1-2 microliter drops containing a (1:1) mixture of the protein and the screening solution are dispensed into the bottom of the well and is covered by a paraffin-silicone (80:20) oil mixture. The water in the aqueous drop slowly evaporates through the oil mixture increasing the concentration of protein in the drop as described previously. The modified microbatch method can produce relatively larger protein crystals and is a very good tool for optimization and additive screens.

In practice, a locally modified Douglas Instruments ORYX6 robot (Figure 2.2B) was used to set up crystallization drops consisting of 0.5 µL of the protein concentrate and 0.5 µL of the screening solution (one setup per well) on a 72-well Nunc plate (Nalge Nunc International) (Figure 2.2B). If necessary, the volume ratio of protein solution to screening solution could be adjusted via software to achieve the best diffracting crystals. To construct the drop, 0.5 µL of the protein solution was loaded into channel-1 of the robots 3-channel microdispenser. Next, 0.5 µL of the screening solution (from a stock plate), were loaded into channel-2 of the robots 3-channel microdispenser (see Figure 2.2B). The two solutions were dispensed into the well and mixed and the drop covered with a small amount of paraffin oil. Once all wells in the plate were setup 4 mL of a (80:20) paraffin-silicone oil mixture was layered on top of the wells. The oil

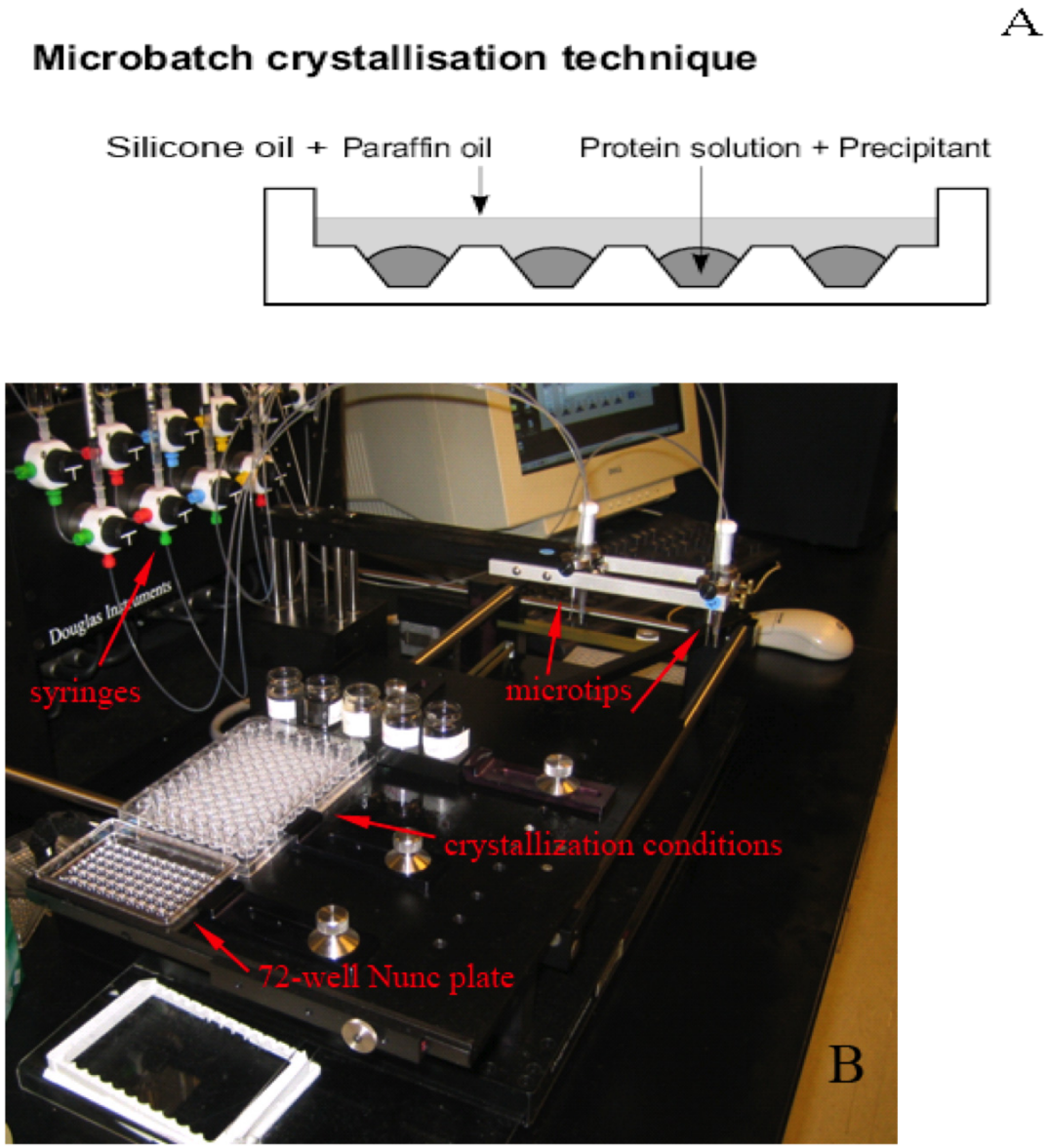


Figure 2.2 Modified microbatch under oil method set up by a locally modified Douglas Instruments ORYX6 robot. A) Schematic representation of modified microbatch under oil, adapted from http://www-cryst.bioc.cam.ac.uk/~dima/whitepapers/xtal-in-action/Timg11.gif. B) A locally modified Douglas Instruments ORYX6 robot.

mixture allows the evaporation rate of the drop to be controlled by the amount of silicone oil used. Using this technique, a complete 384-condition screen can be setup in less than 2 hours.

The initial screening plates, after setup were generally stored at 18 °C and each well observed through a microscope twice a week during the first two weeks and once every two days afterwards for crystallization hits. In cases where it appeared that the crystals grew too quickly (e.g. many small crystals) plates were stored at 4 °C.

Once an initial crystallization hit is confirmed the conditions that produced the crystal(s) are optimized using either a single- or double-grid screen centered on the lead conditions. In cases where the lead conditions that did not contain a salt component a 36 well single-grid screen (pH and precipitant grid) was carried out. Alternatively, for cases where the lead condition contained salt an additional 36 well single-grid screen (pH and salt grid) was also included in the optimization. Finally, in cases where the optimized crystals were unusable due to their diffraction quality additive screens (Additive Screen 1 and 2 (Hampton Research)) were employed.

For phasing heavy atom (Au, Hg, Pt, etc) incorporation into the crystal lattice was required for proteins that did not contain metal or seleno-methionine labels. Generally native crystals are soaked in solutions containing heavy atoms where the heavy atoms diffuse along the mother liquor channels of protein crystals and bind to the specific sites on the protein surface (Blundell *et al.*, 1976). For the work described here a variation of this approach is usedin which several grains of the heavy atom salt (from commercially available kits, Hampton Research) were added to the crystallization drops after crystals of harvestable size were observed. The crystal soaking time used ranged from 10 minutes to 2 hours or even several days (Rould, 1997).

The success of heavy atom incorporation was judged by the presence of a set of consistent significant peaks in the anomalous difference Patterson map.

2.6 Data Collection and Processing

For data collection, the crystal was harvested using a small CryoLoop (Hampton Research) and flash-cooled to 100 K by submersion liquid nitrogen (Teng, 1990). This technique has been shown to significantly improve the data quality during X-ray data collection. In some cases cryoprotectants (Teng, 1990), such as PEG, glycerol, and lithium sulfate, were required for the during the flash freezing process to prevent ice formation and reduce the increase in mosaicity that usually accompanies the freezing process. In practice, crystals were harvested from the crystallization drop and transferred to a drop containing the cryoprotectant and allowed to equilibrate for several seconds before they were recovered and flash-cooled to 100 K. PEG and glycerol were usually mixed with mother liquor from which crystals came and different concentrations (< 30%) of cryoprotectant tested to determine the optimal concentration. In cases where the crystals were grown from high salt solutions lithium sulfate (> 1 M) served as the cryoprotectant.

The diffraction quality of all crystals was tested using in-house copper rotating anode X-ray source. Well diffracting crystals were reserved and shipped to SER-CAT for future high resolution and/or phasing data collection. This in-house diffraction screening was also used to help design and improve crystallization experiments and search for heavy atom incorporation.

SER-CAT data were collected at 100 K using synchrotron radiation from 22-ID and 22-BM with diffraction images recorded using either a MAR 300 CCD (22ID) or MAR 225 CCD (22BM) detectors. X-ray wavelengths were optimized for the anomalous scatterers present in the protein/crystal. For data collected for sulfur-SAS phasing, 1.9 Å X-rays were chosen as a best

compromise between higher X-ray absorption at lower energies and the strength of the sulfur anomalous scattering signal. For crystals containing selenometionyl labeled protein data were collected at or near the selenium absorption edge (0.97Å).

Prior to data collection, a series of images were recorded with different oscillation steps (0.5° - 1.0°) and exposure times and used to decide the data collection parameters: exposure time, oscillation step, the crystal-to-detector distance, X-ray beam size and data collection strategy. Care must be used in choosing the exposure time since although longer exposure times improve weak data in high-resolution shell but long exposures time increases the background and radiation damage to the crystal, and can lead to saturation of low-resolution reflections. The oscillation step is set based on to beam divergence and the crystal's mosaicity to minimize reflection overlap. Likewise the crystal-to-detector distance, which determines the resolution of the experiment, is set to avoid any overlapped reflections. If possible, the X-ray beam size is chosen to match the size of the crystal to reduce background and improve signal-to-noise. Finally a data collection strategy (starting point and total crystal rotation) based on the crystal's Laue group and orientation is generated such that a complete data set can be recorded in the shortest possible time. In most cases this minimum rotation range is expanded so that redundant data can be added to the data set to increase the accuracy of the data and to increase the signal-to-noise of the data set.

In practice, for data collected on 22ID using ~1Å X-rays the exposure time was selected such that the full dynamic range of the detector was used and few (< 3) saturated reflections were observed on the image and corresponded to exposure times in the range of 1-2 sec. The X-ray beam size was set at 100 μm × 100 μm and a 0.5° or 1° or 0.5 oscillation step was used. The test images were taken with a 300 mm crystal-to-detector distance and the final distance was decided

based on the diffraction resolution observed on the test images. A complete data set usually contained 360-degree of data.

For the sulfur-SAS experiments, a longer 2-4 second exposure time was used since the intensity of the X-ray beam decreases at lower energies. The minimum crystal-to-detector distance was set to 125 mm accommodate restrictions imposed by the helium beam path and the beamline's crystal automounter. The data collection strategy used for these experiments was adjusted to produce a highly redundant data set.

Raw data were collected, then indexed, integrated and scaled using either HKL2000 (Otwinowski *et al.*, 1997) or d*TREK (Pflugrath, 1999). In combination with 3DSCALE (Fu *et al.*, 2004), d*TREK was used to calculate R_{as} for a single or merged data sets. The HKL2000 did not provide a set of unmerged, unscaled reflections required by 3DSCALE.

2.7 Structure Determination

The description of heavy atom or anomalous substructure needed for the phase calculation was carried out using SHELXD (Uson *et al.*, 1999) or SOLVE (Terwilliger *et al.*, 1999). Next an experimental electron density map is computed based on positions of the heavy atoms or anomalous scatterers using SOLVE or ISAS. The map was then improved and the protein sequence was fitted into it using RESOLVE (Terwilliger, 2000; Terwilliger, 2003). RESOVE improves electron density maps based on maximum-likelihood method, a statistical approach is used to directly maximize the total probability of the phases and combine experimental X-ray diffraction data with the expected characteristics of an electron density map of a macromolecule. If needed DM (Cowtan, 1994) was used to further modify and improve electron density map since it applies real space constraints to the phasing information obtained from experimental data with options including: solvent flattening, histogram mapping (Zhang *et*

al., 1990), NCS averaging (Schuller, 1996), etc. Finally, if the resolution permitted (> 2.3Å) ARP/wARP (Perrakis *et al.*, 1999) was used to fit the entire polypeptide chain (including side chains) to the electron density map.

Once an initial model has been fitted to the experimental electron density map it is refined against the X-ray diffraction data to improve the fit of the molecular model to the experimental data. The atomic positions in the model, their occupancies and temperature factors are adjusted to (1) increase the agreement between the observed (X-ray data) and calculated (model based) structure factors and (2) meet the stereochemical constraints that describe the geometry of the protein. The progress and success of the refinement can be monitored by: 1) a decrease in R_{work} and R_{free} (see Table 3.5 for explanation); 2) a decrease in the R.M.S.D from ideality for bond lengths and angles; and 3) a disappearance of residues from unfavorable regions of the Ramachandran plot that is used to analyzes the main-chain conformational angles in the polypeptide. Models were usually refined using REFMAC (Murshudov *et al.*, 1997) and manually adjusted (when necessary) by Coot (Emsley *et al.*, 2004). Generally several iterations of manual adjustment followed by refinement were needed to produce the final model. In addition, in the final stages of the fitting/refinement process solvent molecules (modeled as water) were identified based on their peak height in the electron density map and the formation of hydrogen bond to either protein or other solvent molecules.

The final model including solvent was then validated for stereochemical correctness using MOLPROBITY (Davis *et al.*, 2007) and PROCHECK (Laskowski *et al.*, 1993) and then deposited into Protein Data Bank (PDB) (Berman *et al.*, 2000). In addition, the MOLPROBITY analysis which adds hydrogen atoms to the to a protein model carries out an all-atom contact and

high-accuracy stereochemical analysis can be applied from the beginning of the refinement to improve the efficiency in refinement and accuracy of the final structure (Arendall *et al.*, 2005).

For these studies two high throughput structure determination pipelines developed by SECSG were used: SGXPro (Fu *et al.*, 2005) and the cluster based Sca2structure (Liu *et al.*, 2005a). Both were developed to integrate a variety of crystallographic structure determination programs using a parallel workflow engine to increase the efficiency of structure determination process by systematically searching both program and parameter space for optimal program parameterization.

SGXPro suite includes 3DSCALE, SHELXD, ISAS, SOLVE/RESOLVE, DM, SOLOMON, DMMULTI (Abrahams *et al.*, 1996), BLAST, AmoRe (Navaza, 1994), EPMR (Kissinger *et al.*, 1999), PHASER (Mccoy *et al.*, 2007), COOT, ARP/wARP and MAID (Levitt, 2001). In as illustrated in illustrated in the Figure 2.3 SGXPro flowchart SOLVE and SHELXD were run with a set of data resolution cutoffs to define the heavy/anomalous scatterer substructure. SGXPro then uses this information to determine the protein handedness and generate phases using different resolution cutoffs. RESOLVE is then used to improve the phases and for automated sequence fitting. SGXPro then lists the top five models including output files and phases.with the phases. The output listed all files including log files of those top solutions. SGXPro employs a user-friendly interface and displays the best solution obtained in COOT directly. SGXPro also can be used define the space group and to calculate (3DSCALE) the data sets R_{as} value.

The Sca2structure pipeline integrates SOLVE/RESOLVE, ISAS, and DM, ARP/wARP and REFMAC from of the CCP4 program suite (Bailey, 1994), see Figure 2.4A. A dictionary-driven web-based user interface (Figure 2.4B) is used to collect select the experimental data file

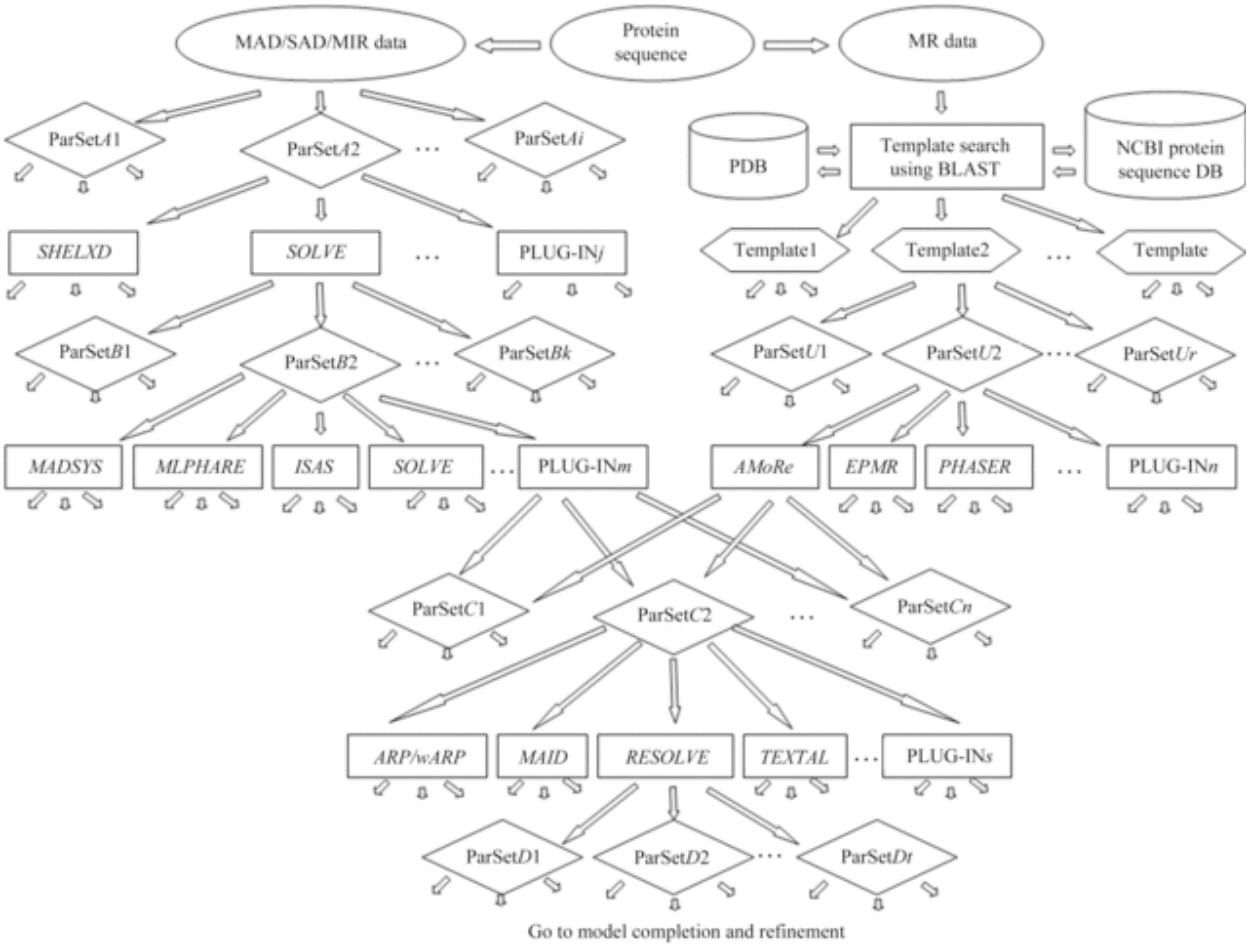


Figure 2.3 Flow charts of the SGXPro jobs to solve structures using MAD/SAS/MIR or MR methods. Adapted from: Fu, Z. Q., J. Rose and B. C. Wang (2005). Acta Crystallogr D Biol Crystallogr 61(Pt 7): 951-9.

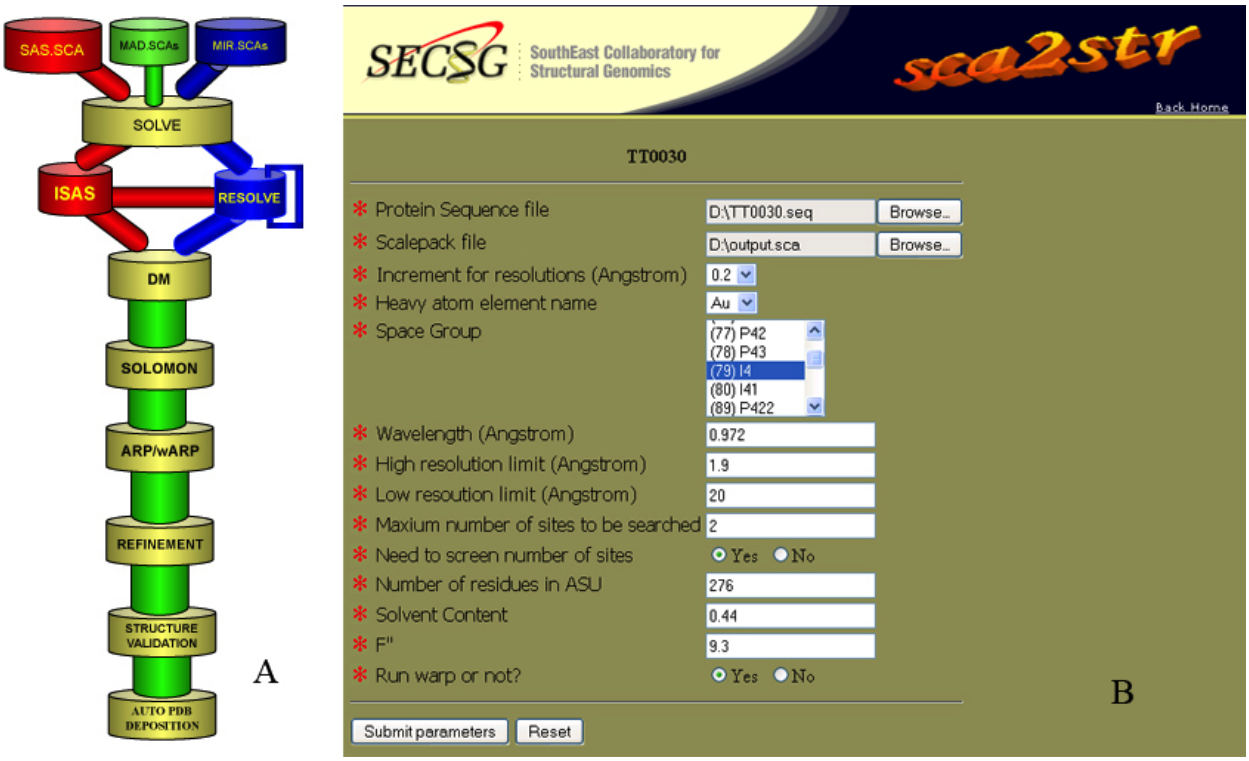


Figure 2.4 Sca2Structure pipeline at UGA. A) Flow chart of Sca2Structure pipeline. B) The dictionary-driven web-based user interface to input data.

and to input parameters such as resolution ranges, space group, heavy atom type, and possible number of heavy atoms. The workflow-management system then submits multiple structure-determination jobs to be run in parallel on the 128-processor cluster each with a slightly different set of program parameters. Using this approach program parameter space can be screened in under an hour to determine if a solution is possible for the given data set. Upon completion, the results are harvested, analyzed, sorted, and presented to the user as a web-based table with links to tar files that contained output files of various solutions.

2.8 Structural Analysis and Functional Prediction

Structural similarity analysis was carried out using either SSM (Second Structure Matching, http://www.ebi.ac.uk/msd-srv/ssm/) (Krissinel *et al.*, 2004) or DALI (Distance Matrix Alignment; http://www.ebi.ac.uk/dali/) (Holm *et al.*, 1996) to examine a protein structure for similarity within whole PDB or SCOP archive. Topology diagrams of structures were generated by PDBsum (http://www.ebi.ac.uk/pdbsum/) (Laskowski *et al.*, 1997). These results were used as the base to determine whether the new structure represented a new SCOP fold. Probable assemblies of the target proteins were predicted by the Protein Interfaces, Surfaces and Assemblies (PISA) service (Krissinel *et al.*, 2007) at European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html).

Although non-Pfam proteins have unique sequences, their functions still could be indicated by their structures, especially when compared to similar structures of known function. Methods used could in these analyses include: i) SUPERFAMILY server (http://supfam.org/SUPERFAMILY/index.html) (Gough *et al.*, 2001) to find remote relationships with structure-known proteins; ii) ProFunc server (http://www.ebi.ac.uk/thornton-srv/databases/profunc/) (Laskowski *et al.*, 2005) to identify the possible biochemical function by

identifying possible cation/anion-binding regions, Helix-turn-helix (HTH) DNA-binding motifs, possible enzyme active sites or ligand binding sites; and iii) operon analysis by MicrobesOnline Operon Predictions (http://www.microbesonline.org/operons/) (Price *et al.*, 2005) to further classify the target protein and predict possible functions. In addition, human interpretation of the results is required to assess the validity of the information espically where the analyses disagree.

CHAPTER THREE

Results and Discussion

3.1 Statistics of Non-Pfam Protein Production

The population of non-Pfam sequences accounts for more than 30% of the selected six genomes, and even reaches 62% in *Aeropyrum pernix* (Table 3.1). A total of 1,542 ORFs were selected for cloning and expression screen. 1,268 genes were cloned and 715 targets passed expression screen for large-scale production (Xu, unpublished data).

In this work, 51 out of 62 selected target proteins were expressed in large-scale media and 23 proteins were successfully purified by chromatography methods. Crystallization hits were observed in 8 proteins (Table 3.2). Crystal structures of AF1382 and TT0030 were determined and deposited into PDB.

The production of non-Pfam proteins accorded with our expectations that there would be no extraordinary difficulties in expression, purification, and crystallization of non-Pfam proteins as compared to Pfam proteins. In the PSI-1 stage for Pfam structure production, SECSG (http://targetdb.pdb.org/statistics/sites/SECSG.html#status) had cloned 14,876 ORFs and 6,410 were expressed. Within 837 purified proteins, 239 were crystallized and 96 crystal structures solved. In all PSI-1 centers, the average success rate in going from purified protein to crystal was 36% while the success rate in going from purified protein to structure was only 7% (Liu *et al.*, 2005b). Those two statistics were 35% and 8.7%, respectively, in the work described here.

During target selection, we did not purposely exclude targets that are predicted to have difficulties in crystallization. Thus these statistics are not biased and reveal the real information

Table 3.1 Target selection of non-Pfam proteins

Genome	Number of ORFs	Percentage of non-Pfam genes (%)	Number of the selected non-Pfam targets
Pyrococcus furiosus	2065	30.97	258
Pyrococcus horikoshii	2061	39.20	205
Aeropyrum pernix	2694	62.34	328
Archaeoglobus fulgidus	2407	30.78	268
Thermus thermophilus	1908	30.22	260
Clostridium thermocellum	3163	32.36	223
Total	14298	38.24	1542

Table 3.2. Summary of expression, purification and crystallization of the selected non-Pfam targets.

a: Number of crystallized targets / Number of purified targets × 100%

Genome	Number of targets	Number of expressed targets	Number of purified targets	Number of crystallized targets	Success rate, from purified protein to crystal [a] (%)
Pyrococcus furiosus	8	3	3	0	0
Pyrococcus horikoshii	12	12	6	2	33.3
Archaeoglobus fulgidus	20	16	5	2	40
Thermus thermophilus	22	20	9	4	44.4
Total	62	51	23	8	34.8

for non-Pfam sequences. Additional savaging methods (Liu *et al.*, 2005b) could be applied in each step to improve the success rate.

3.2 Structural and Functional Aspects of AF1382

3.2.1 Gene cloning of AF1382

The gene encoding *Archaeoglobus fulgidus* ORF 1382 (AF1382) was amplified from the genomic DNA of *Archaeoglobus fulgidus* DSM4304 by PCR and cloned into plasmid pDEST-527. A TEV cleavage site was constructed between the target gene and the N-terminal His_6 tag. It was then transformed into *E. coli* host strain BL21 DE3 RPX (Stratagene).

The protein is composed of 95 amino acids (11.14 kDa molecular weight) with pI of 5.2. Because it contains 4 methionines and 1 cystine, the protein could be a good target for sulfur-SAS phasing, as well as Se-SAS phasing. Native and SeMet AF1382 were expressed, purified, and crystallized in parallel.

3.2.2 Expression, Purification and Crystallization of Native AF1382

Cells carrying AF1382 were first grown in 50 mL LB medium with 100 μg/mL ampicillin at 37 ºC for 8 hours and transferred to 1 L LB medium. The expression was induced with 0.5 mM IPTG at 25 ºC after the culture reached $OD_{600} \sim 0.5$. Three hours after the IPTG induction, cells were harvested by centrifugation (6000 rpm × 15 min) and pellets were stored in -80 ºC. Expression of native AF1382 was confirmed by SDS-PAGE (Figure 3.1).

Purification started with 4.6 g biomass from 1 L *E. coli* culture. The cell pellets of native AF1382 were re-suspended in 25 mL Ni affinity binding buffer and lysed by ultrasound (6 × 30 s, on ice) in the presence of 5 mM β-ME and 1 mM PMSF. The sample was centrifuged at 12,000 rpm for 30 minutes at 4 ºC. The supernatant was saved and filtered through 0.45 μm filters. The protein solution was loaded to a 5 mL HiTrap™ Chelating HP Column (GE Healthcare, Ni-

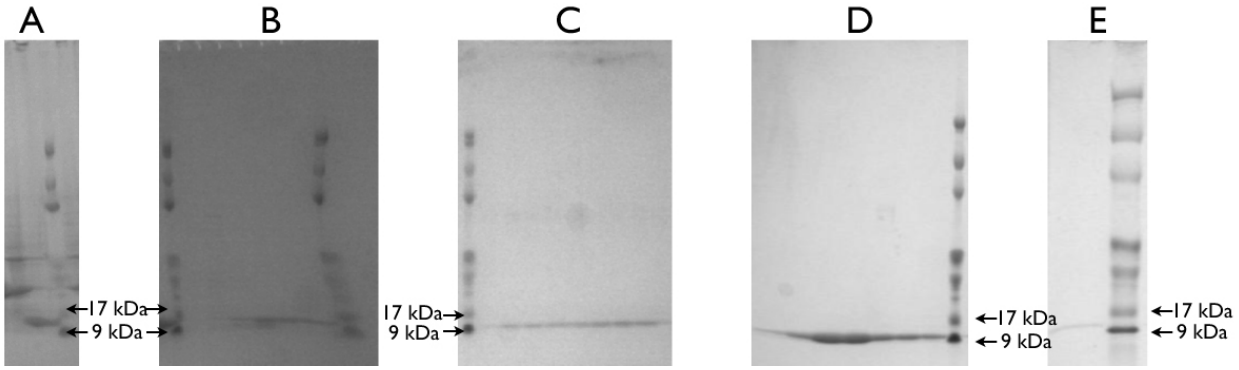


Figure 3.1 Purification of the native AF1382. A) Un-induced cell culture as a control (left) and induced expression (right); B) fractions of elution peak from Ni affinity chromatography; C) fractions of flow-through from 2nd Ni chromatography run after TEV cleavage and heat treatment; D) fractions of elution peak from ion-exchange chromatography; E) final concentrated native AF1382 sample (11.2 mg/ml).

affinity) and eluted with a 0-500 mM imidazole gradient at 4 ºC. The eluted fractions were pooled and subject to a 24-hour TEV cleavage at room temperature and were dialyzed against the Ni affinity binding buffer at the same time. After TEV treatment was complete, the mixture was heated to 65 ºC for 1 h, followed by centrifugation and filtering. The supernatant was re-applied to a Ni-affinity chromatography column, and the flow-through fractions were pooled and dialyzed overnight against 2 L of the ion-exchange loading buffer pH 7.6. The protein sample was then applied to a 5 mL HiTrap Q-Sepharose column (GE Healthcare, anion-exchange) and eluted by a 25-1000 mM sodium chloride gradient at 4 ºC. Pooled protein fractions were concentrated to 3 mL and applied to a Superdex 200 column (GE Healthcare, size exclusion). Purity of peak or flow-through fractions in each chromatography purification step was analyzed by SDS-PAGE (Figure 3.1). The final protein sample was concentrated to 1200 μL at 11.2 mg/mL.

Initial crystallization screening trials were set up in microbatch mode against 384 conditions as described in 2.5 and incubate at 18°C. Crystals of native AF1382 (Figure 3.2) were observed within 2-3 days from several conditions in initial screenings: Crystal Screen-35, Crystal Screen II-13, Wizard I-44, MemFac-1, MemFac-8, MemFac-11, MP1-5, and MP1-6 (Table 3.3). All initial protein crystals had poor diffraction quality, possibly because crystal lattices grew too fast. It was thought to be reasonable to reduce the protein concentration. Optimization plates were then set up with proteins at original concentration and re-screening plates were set up with 2-fold and 4-fold diluted proteins. The crystal used for phasing was obtained from optimization condition 21 of MP1-5. The 1.0 μL drop contained equal volume of 11.2 mg/mL protein concentrate and precipitant solution containing 0.1 M sodium citrate/citric acid pH 5.1, 0.1 M sodium chloride, 0.1 M lithium sulphate, and 25% v/v PEG 400. The crystal

Table 3.3 Initial crystallization conditions for the native AF1382.

Condition name	Salt	Buffer	Precipitant
Crystal Screen-35		0.1 M sodium acetate pH 4.6	8% w/v PEG 4000
Crystal Screen II-13	0.2 M ammonium sulfate	0.1 M sodium acetate pH 4.6	30% w/v PEG Monomethyl Ether 2000
Wizard I-44	0.2 M calcium acetate	0.1 M acetate pH 4.5	30% (v/v) PEG-400
MemFac-1	0.1 M sodium chloride	0.1 M sodium acetate pH 4.6	12% v/v 2-Methyl-2,4-pentanediol
MemFac-8	0.1 M magnesium chloride	0.1 M sodium acetate pH 4.6	18% v/v PEG 400
MemFac-11	0.1 M magnesium chloride	0.1 M sodium acetate pH 4.6	12% w/v PEG 6000
MP1-5	0.1 M sodium chloride, and 0.1 M lithium sulphate	0.1 M Sodium citrate/citric acid pH 5.5	30% v/v PEG 400
MP1-6	0.1 M sodium chloride, and 0.1 M magnesium chloride	0.1 M sodium citrate/citric acid pH 5.5	30% v/v PEG 400

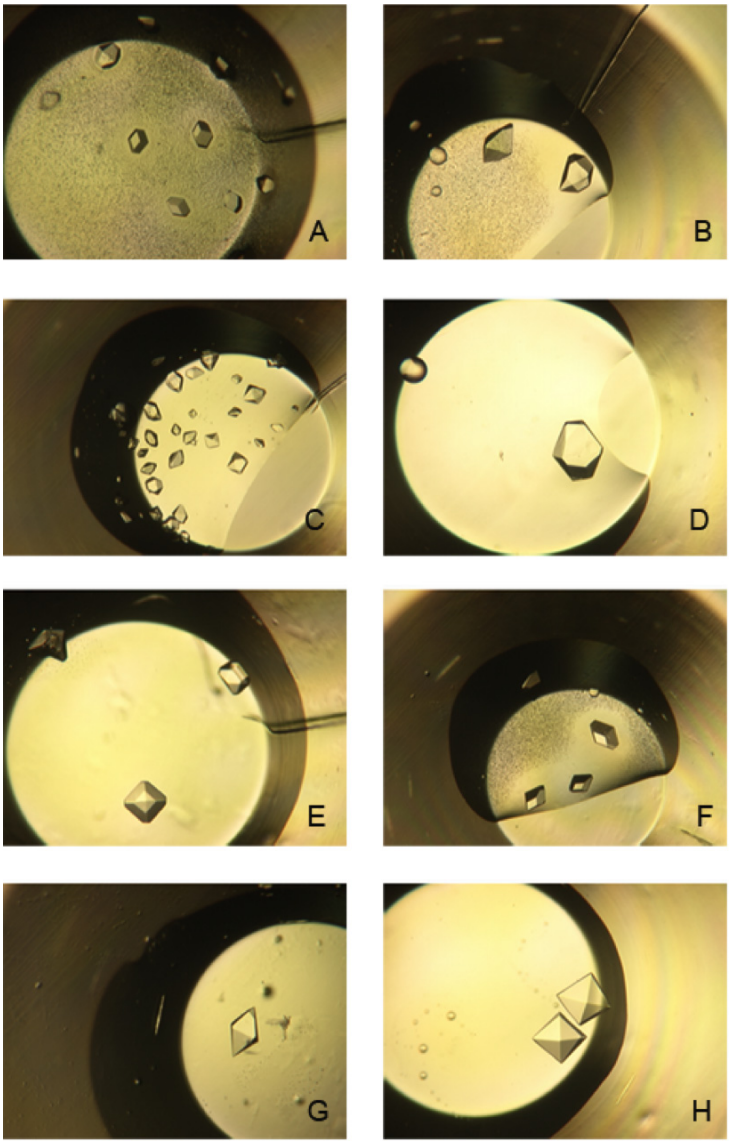


Figure 3.2 The native AF1382 crystals observed in different conditions. A) crystals observed in Crystal Screen 35; B) crystals observed in MemFac-8; C) crystals observed in MemFac-11; D) crystals observed in MemFac-1; E) crystals observed in MP1-5; F) crystals observed in Wizard I-44; G) a crystal observed in Optimization of MP1-5; H) crystals observed in the re-screen of Crystal Screen II-13.

used to collect the refinement data was obtained from the re-screened Crystal Screen II-13 (0.2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6, 30% w/v PEG Monomethyl Ether 2000). The 1.0 μL drop contained equal volume of 2.8 mg/mL protein concentrate and precipitant solution.

3.2.3 Expression, Purification and Crystallization of SeMet AF1382

Cells carrying AF1382 were grown in 50 mL PA0.5G medium with 100 μg/mL ampicillin at 37 ºC for 8 hours and transferred to 1 L of PASM5052 medium in the presence of 125 μg/mL SeMet. When OD_{600} reached 0.5, cell cultures were transferred to 25 ºC for another 20-hour growth. Cells were harvested by centrifugation (6000 rpm × 15 min) and pellets were stored in -80ºC. Expression of SeMet AF1382 was confirmed by SDS-PAGE (data not shown).

The purification of SeMet AF1382 protein started with 4.0 g biomass from 1 L *E. coli* culture. Purification procedures of the SeMet AF1382 protein were the same as those used for the native protein (data not shown). The final protein sample was concentrated to 500 μL at 15.6 mg/mL.

Initial crystallization screening trials were set up in microbatch mode against 384 conditions as described in 2.5 and incubate at 18°C. Crystals of SeMet AF1382 protein (Figure 3.3) were observed in Crystal Screen-10 (0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.6, and 30% w/v PEG 4000). Optimization was carried out after the protein crystals were confirmed. An Additive Screen 2 (Hampton Research) based on optimization 9 (0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.9, and 15% w/v PEG 4000) was used to improve diffraction quality. The best diffracting quality crystal was obtained from 1.0 μL drop containing 40% volume of 15.6 mg/mL protein concentrate, 40% volume of optimized precipitant and 20% volume of 10% v/v ethyl acetate.

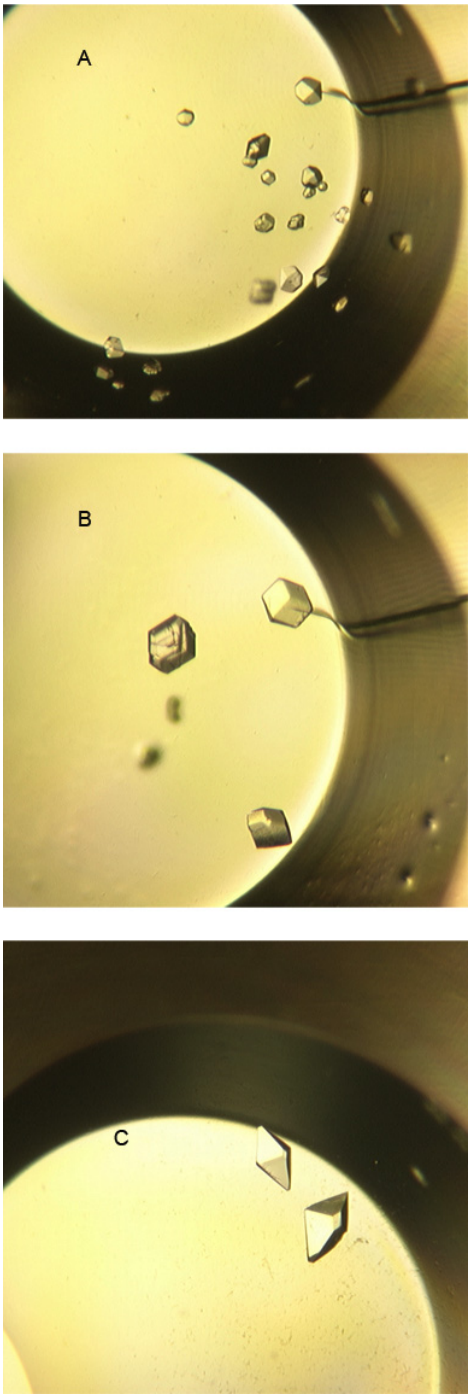


Figure 3.3 The SeMet-labeled AF1382 crystals observed in different conditions. A) Crystals observed in Crystal Screen-10; B) crystals observed in Optimization of Crystal Screen-10; C) crystals observed in Additive Screen 2 based on optimization of Crystal Screen-10.

3.2.4 Data Collection and Processing of Native and SeMetAF1382

All the crystals were mounted by CryoLoop (Hampton Research) and flash-cooled to 100 K. Cryoprotection was applied to crystals that showed ice ring during X-ray diffraction. Their diffraction quality was screened using the in-house copper X-ray source. Crystals that diffracted above 3 Å resolution were shipped to SER-CAT at APS for data collection. None of the SeMet crystals diffracted higher than 3 Å resolution using the in-house copper source. The best diffracting SeMet crystal was then shipped to APS for data collection.

All the diffraction data were collected on 22-ID, SER-CAT. X-rays at 1.9 Å wavelength were chosen for sulfur-SAS phasing on native crystals. Two 360-degree data sets were collected (1º oscillation steps) on the same crystal. The exposure time of first and second data set was 3 and 2 seconds, respectively. The crystal-to-detector distance was 125 mm.

0.9724 Å wavelength X-rays were used for refinement data. One set of data was collected (1º oscillation steps). The exposure time was 1 second. The crystal-to-detector distance was 230 mm.

For SeMet crystals, wavelength of X-rays was set to be 0.9724 Å because it is close to the absorption edge of selenium. One set of data was collected with 1º oscillation angle. The exposure time was 1 second. The crystal-to-detector distance was set at 230 mm.

All the native and SeMet data were processed by HKL2000 for structure determination. To evaluate R_{as}, the individual and the merged phasing data were first processed by d*TREK. The integration files generated by d*TREK were then scaled by 3DSCALE to calculate R_{as} at different resolution ranges.

Statistics of data processing for native crystals are presented in Table 3.4. The resolutions of the merged and two single phasing data sets are 2.3 Å. The native crystal

Table 3.4 Statistics of data processing for the native AF1382.

a, $R_{sym} = \Sigma|I - \langle I \rangle|/\Sigma I$, where I is the observed intensity of reflections. R_{merge} is calculated from the merged data.

Statistics	Phasing data 1	Phasing data 2	Merged data	Refinement data
X-ray wavelength (Å)	1.9	1.9		0.9724
Space group	$P4_2$	$P4_2$	$P4_2$	$P4_2$
Unit cell dimensions (Å)	a = b= 53.54, c = 41.25	a = b= 53.52, c = 41.24	a = b= 53.54, c = 41.25	a = b= 53.03, c = 40.97
Resolution range (highest resolution shell)	50.00-2.30 (2.38-2.30)	50.00-2.30 (2.38-2.30)	50.00-2.30 (2.38-2.30)	50.00-1.85 (1.92-1.85)
Completeness (%)	95.7 (66.8)	99.8 (99.2)	99.9 (99.2)	99.9 (99.8)
Redundancy	12.0 (3.9)	13.6 (10.7)	25 (13.4)	13.2 (12.5)
R_{sym} (%)[a]	4.3 (32.3)	4.1 (18.5)		5.4 (20.4)
R_{merge} (%)[a]			4.5 (25.4)	
I/sigI	72.3 (3.34)	93.7 (17.58)	124.7 (17.38)	60.66 (11.96)
Mosaicity (°)	0.54	0.53	0.58	0.56
Unique Reflections	5155	5313	5317	9863
Reflections measured	61502	72075	132985	130989

belonged to space group $P4_2$. The unit cell dimensions of the phasing crystal were a = b = 53.54 Å, c = 41.25 Å. The solvent content and Matthews coefficient (Vm) of native crystal used in the phase calculations was estimated to be 54.4% and 2.7 $Å^3$/Da. There was one protein molecule in each asymmetric unit.

The resolution of refinement data collection was 1.85 Å. The unit cell dimensions of refinement crystal were a = b = 53.03 Å, c = 40.97 Å.

Statistics of data processing for SeMet crystals are presented in Table 3.5. The resolution of SeMet data collection was 2.8 Å. The SeMet crystal belonged to space group $P4_2$, and unit cell dimensions were a = b = 53.96 Å, c = 41.30 Å.

Processed data from the first phasing data set was loaded into SGXPro and failed to produce an interpretable electron density map. The same native crystal was re-mounted and another data set was collected with smaller exposure time to minimize radiation damage to the crystal. Then two data sets were manually merged with anomalous flag on. However, no improvement was observed.

A different strategy in data processing was then applied by choosing a larger spot size of 0.9 for indexing. The re-processed data sets were manually merged with anomalous flag on. Neither single re-processed data set could lead to a successful structure determination.

3.2.5 Improved Data Quality and Anomalous Signal after Data Merging

The enlarged indexing spot size and data merging are the keys to the success of this structure while the latter is more critical.

Several improvements in the quality of phasing data could be easily observed in Table 3.4. R_{sym} is an indicator routinely used to judge the quality of the diffraction data set. A lower value of R_{sym} indicates better data quality. Usually we would expect an increased R_{sym} after

Table 3.5 Statistics of data processing for the SeMet-labeled AF1382.

a, $R_{sym} = \Sigma|I - \langle I \rangle|/\Sigma I$, where I is the observed intensity of reflections.

Statistics	Phasing data
X-ray wavelength (Å)	0.9724
Space group	$P4_2$
Unit cell dimensions (Å)	a = b = 53.96, c = 41.30
Resolution range (highest resolution shell)	50.0-2.80 (2.90-2.80)
Completeness (%)	98.1 (98.8)
Redundancy	12.4 (7.6)
R_{sym} (%)[a]	6.5 (33.6)
I/sigI	42.3 (4.89)
Mosaicity (°)	N/A
Unique Reflections	3296
Reflections measured	41033

several data sets are merged together. The R_{merge} for the merged data in this case was only slightly higher than the R_{sym} observed in both single data and was still in the range expected for synchrotron X-ray diffraction data (< 6%). I/σI in merged data was 72.5% and 33% higher than that in single data set, respectively. The sharply increased redundancy in merged data was almost the sum of redundancies in two individual data sets. Higher redundancy could effectively reduce the random noise thus enhance anomalous signal. According to previous analysis (Fu *et al.*, 2004), the doubled redundancy is the most important improvement in data quality after data merging.

R_{as} value in different resolution ranges from different data sets were plotted in Figure 3.4. The individual data sets have better R_{as} in certain resolution range, comparing to each other. The merged R_{as} is much higher than that in any single data set. At 3 Å resolution, R_{as} in both single data sets is 1.1 and merged R_{as} is 1.3. Although its R_{as} is lower than predicted standard of 1.5, the merged data set is still able to provide us reasonable phases. At this point, it is not sure AF1382 is an exception in sulfur-SAS phasing or its R_{as} could be the future minimal standard. One important factor should not be ignored that it has been four years since the introduction and analysis of R_{as} in sulfur-SAS phasing. With accumulated advances in technology, especially the improvements made to optics at SER-CAT, it is very possible to solve a structure by sulfur-SAS with weaker anomalous signal level.

Another interesting phenomenon is that the quality and R_{as} of the second data set are much better than the first set in resolution ranges lower than 3 Å, although a complete data set has been already taken on that crystal. Longer exposure time is usually adopted to improve weak data in high-resolution range. As SHELXD locates the anomalous scatterers and calculate phases only by the data to a resolution 3.5 Å, weak data in the higher resolution range are not

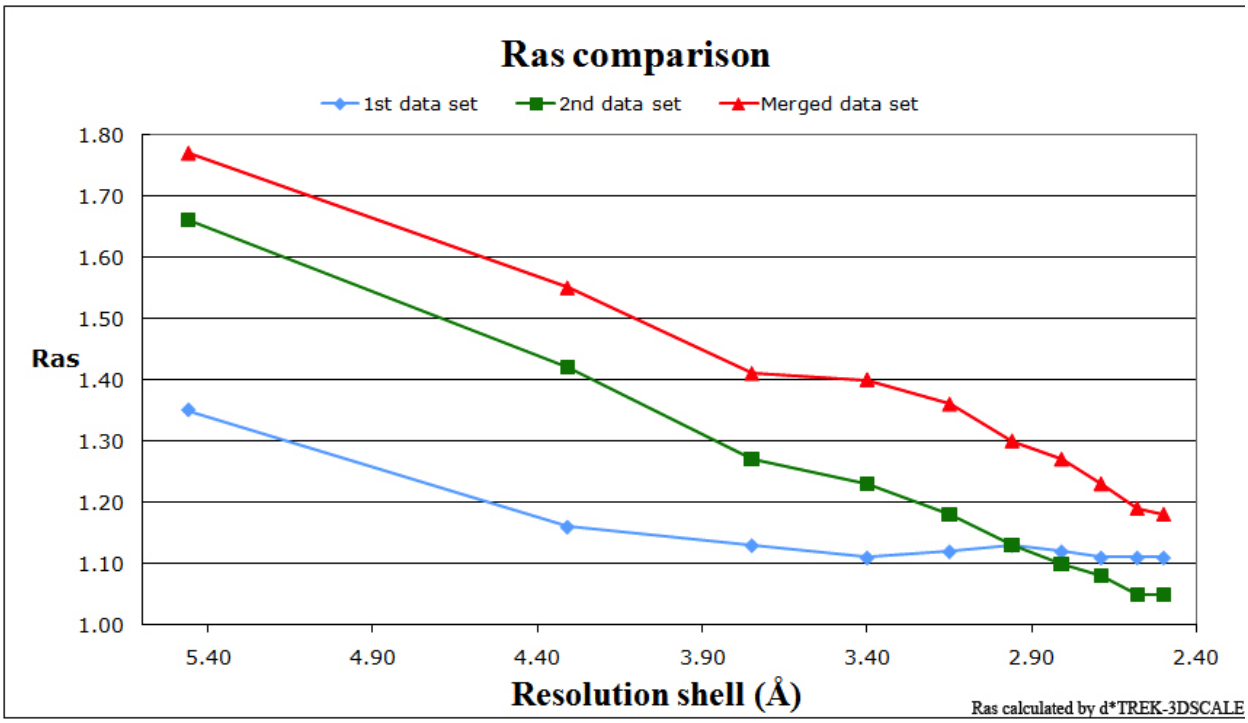


Figure 3.4 The comparison of R_{as} in the individual and the merged phasing data of native AF1382. The R_{as} in the merged data is higher than that in the individual data sets. At 3 Å resolution, the merged R_{as} reaches 1.3, while the individual R_{as} is only 1.1.

necessary for the phasing. In this work, the shorter exposure time provides a lower background and improved data quality in the low-resolution range.

3.2.6 Structure Determination of AF1382

SeMet AF1382 data were loaded into SCA2Structure pipeline. Less than half residues were traced in the best solution and many residues were not correctly placed. The resolution is lower than 2.8 Å and the electron densities of most side chains in α-helices were not observed in the map. As native crystals diffracted much better than SeMet crystals, the structure determination by Se-SAS was halted.

Initial phases were generated by using the merged reprocessed native data with SGXPRO. Except for the sulfur atom in Met1 which is often disordered in crystals, the remaining four sulfur atom sites were located by SHELXD using 3.5 Å merged data. Phases calculated by sulfur phasing were then extended to 2.6 Å and improved phases and auto-tracing were generated by SOLVE/RESOLVE using this map. A total of 70 out of 95 residues were fitted into the experimental electron density. The resulting phases were then extended to refinement data set and an initial model was built automatically using ARP/wARP. The model was manually adjusted and refined using Coot and REFMAC, respectively. When the free R-factor was reduced below 30%, molecules with peak heights above 3.2σ and good hydrogen bonding geometries were identified by using ARP/wARP and included in the model. Validation was carried out using MOLPROBITY and PROCHECK before deposition of the model coordinates and structure-factor amplitudes to the PDB.

3.2.7 Structure of AF1382 Belongs to a Winged-Helix Fold

The final model includes 87 of 95 amino acids, 714 protein atoms, and 39 water molecules. A total of 8 residues, including Met1, Glu2, Asp3 and Glu4 at N-terminus and Glu92,

Asn93, Asp 94 and Thr 95 at C-terminus, are disordered in crystal lattice, so their electron densities cannot be observed. Statistics of refinement were presented in Table 3.6. The R_{work} factor is 23.9% and R_{free} factor is 28.0% using 5% of total reflections as a test data set. The model has good stereochemistry with RMSDs in bond length and angles of 0.010 Å and 1.12°, respectively. Analysis by Molprobity shows no residues in disallowed regions of Ramachandran plot and very low clashscore of 6.43. The Molprobity score is 1.36 and in the 98th percentile. Atomic coordinates have been deposited with the PDB accession code 2QVO.

Monomer of AF1382 (Figure 3.5A) is a mixed alpha plus beta (α+β) structure with five helices and two β-strands (Figure 3.5B). The helices H2, H3, and H4 comprise a bundle of three α-helices, followed by the two β-strands that form an anti-parallel β-sheet through a reverse turn. The helix H5 is connected to the wing at C-terminus.

PISA prediction shows that AF1382 is most likely to form a dimer in solution. The dimer was stabilized by hydrophobic interactions at C- and N-termini and three pairs of hydrogen bonds that are formed between Arg5 and Asp76; Lys7 and Lys11'; Gln83 and Gln 83' in the related molecules, respectively. Formation of the dimer buries 2,250 Å^2, or 19.4%, of the total monomer surface area.

A SSM search (Table 3.7) confirmed a 93-residue protein with homologous structure to AF1382: F93 in reading frame F of the *Sulfolobus* turreted icosahedral virus (STIV) genome (PDB accession code 2OBP, (Larson *et al.*, 2007)). The SSM Z score and RMSD between two structures are 5.1 and 1.54 Å, respectively. Their structural similarities could be observed in the superposition of two structures (Figure 3.6).

The topology (Figure 3.7) and structural similarity of AF1382 have revealed that it belongs to a winged-helix or winged-helix-turn-helix (wHTH) fold, a subclass of the HTH

Table 3.6 Quality of the present AF1382 model.

a, $R_{work} = \Sigma|F_{obs} - F_{calc}|/F_{obs}$.

b, R_{free} is as for R_{work} but calculated using a 5% test set of reflections excluded from the refinement

c, 100th percentile is the best.

Refinement	
Number of refined atoms	753
Number of water molecules	39
R_{work} factor[a]	23.9
R_{free} factor[b]	28.0
R.m.s.d. from ideal	
Bonds (Å)	0.010
Angles (°)	1.12
Mean *B* value	23.01
Atom clash score	6.46
Ramachandran favored	84 / 85 (no outlier)
Molprobity score[c]	1.36 (98th percentile)

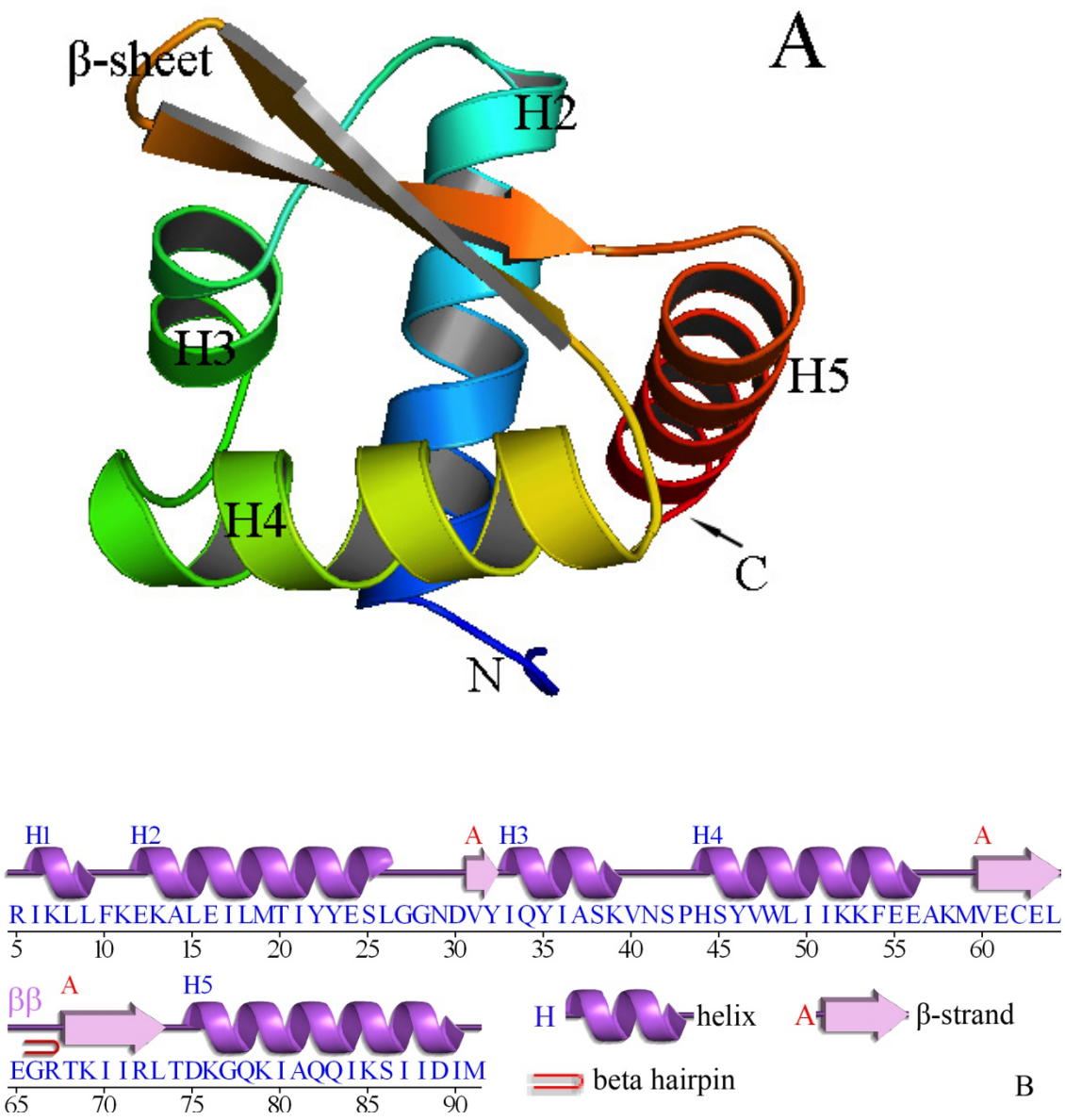


Figure 3.5 The overall structure of AF1382. A) A cartoon representation of the AF1382 structure. Blue area: N-terminus; red area: C-terminus. The helices H2, H3, and H4 comprise a bundle of three α-helices, followed by the two β-strands that form an anti-parallel β-sheet through a reverse turn, drawn by PyMOL. B) The primary sequence of AF1382 annotated with secondary structural elements, generated by PDBsum.

Table 3.7 The SSM search results for structures homologous to AF1382. Q-score represents the quality function of Ca-alignment. It reaches 1 only in the case of identical structures, and drops down with increasing RMSD or decreasing alignment length. Z-score measures the statistical significance of a match in terms of Gaussian statistics. RMSD stands for the Root Mean Square Deviation, calculated between Ca-atoms of matched residues at best 3D superposition of the query and target structures. N_{algn} is the number of matched residues. N_g is a quality characteristics of Ca-alignment. %seq is calculated as a fraction of pairs of identical residues among all aligned. %sse tells what fraction of Secondary Structure of query chain was identified in the target protein.

Scoring			Rmsd (Å)	N_{algn}	N_g	%seq	Query	Target (PDB entry)			
Q	P	Z					%sse	Match	%sse	N_{res}	Title
0.63	4.9	6.4	1.47	74	2	18	86	2obp: A	100	81	YP_298295.1 from Ralstonia eutropha JMP134 (putative DNA-binding)
0.59	3	5.1	1.54	77	2	16	86	2co5:A	86	92	F93 from STIV
0.58	4.1	5.9	1.87	80	3	21	86	2pg4: A	86	92	NP_147569.1 from Aeropyrum pernix
0.54	5	6.9	1.68	76	1	14	86	1r1u:B	86	93	Metal-sensing transcriptional repressor CzrA from Staphylococcus aureus
0.54	4.3	6	1.67	66	2	15	71	2jt1:A	100	71	Pefi protein from Salmonella typhimurium (NMR)
0.54	3	5.3	1.55	61	2	8	71	2acj:D	100	63	B/Z junction containing DNA bound to Z-DNA binding proteins
0.53	3.8	5.9	1.29	56	2	13	71	1j75:A	100	57	DNA-binding domain Zalpha of DLM-1 bound to Z-DNA

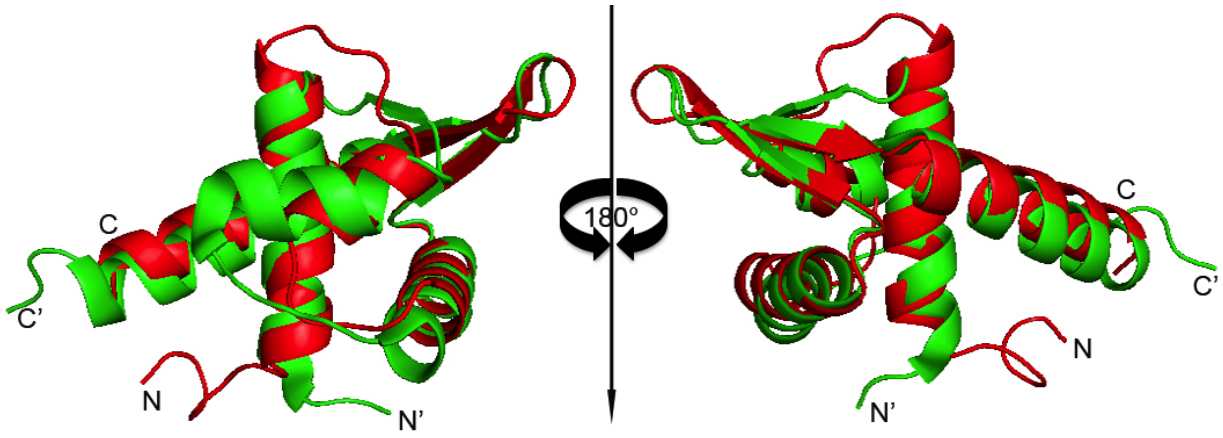


Figure 3.6 Superposition of AF1382 and STIV F93 monomer. AF1382 is structurally homologous to STIV F93 (SSM Z score = 5.1, RMSD 1.54 Å). Red: AF1382 (C and N); green STIV F93 (C' and N'); drawn by PyMOL.

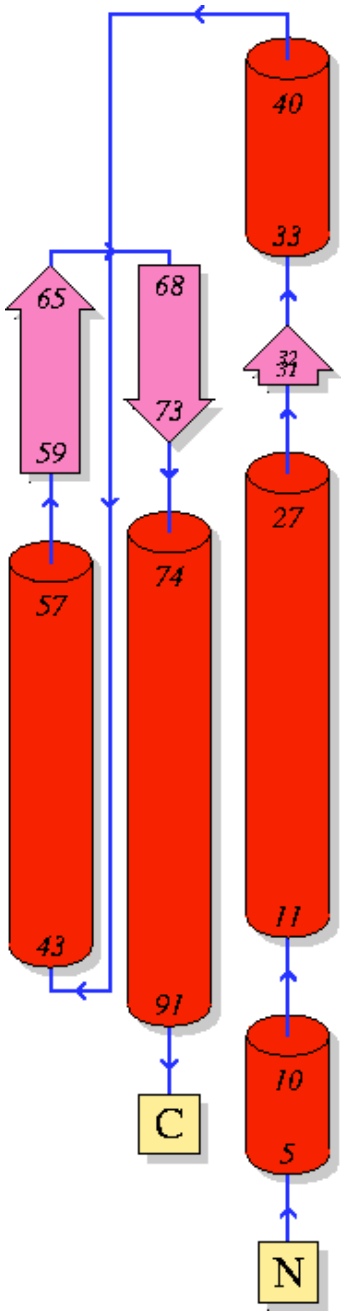


Figure 3.7 The topology of AF1382. Cylinder: α-helix; arrow: β-strand; generated by PDBsum. It suggests that AF1382 belongs to a winged-helix fold.

protein superfamily (Aravind *et al.*, 2005). The bundle of three α-helices in AF1382, which is right-handed, coincides with the basic feature of the HTH motif that they form a triangular outline when the third helix is placed in the front and in horizontal orientation (Figure 3.5). The double stranded β-sheet represents the wing of this fold motif. Variations on this fold in AF1382 include N- and C-terminal α-helices (H1 and H5). C-terminal helix (H5) extends away from the wing, parallel to helix H2. C-terminal extensions to the core wHTH domain are usually involved in dimerization (Aravind *et al.*, 2005).

3.2.8 Putative Interactions with DNA

Sequence search against Pfam 20.0 shows that AF1382 belongs to Pfam-B 160313, automatically generated from an alignment taken from Prodom 2005.1 (PD029452). Superfamily search indicates that AF1382 belongs to the "Winged helix" DNA-binding domain superfamily. The sequential alignment generated by ESPript (Gouet *et al.*, 1999) shows only 13% identity between the AF1382 and the STIV F93 (Figure 3.8).

The third helix in HTH motif usually serves as the recognition helix that could be inserted into the major groove of the DNA. The wing would interact with the minor groove. The two- and three-stranded versions of the wHTH are discovered in DNA-binding domains of some of the largest families of prokaryotic transcription factors, as well as several eukaryotic DNA-binding domains.

The STIV F93 has been confirmed to be capable of binding a 44 base pair synthetic DNA and a 201 base pair DNA of the STIV genome (bases 1201-1368) (Larson *et al.*, 2007). The recognition helix in its structure is inserted into adjacent major grooves of the DNA, while the positive charged wing and the N-terminus interact with the ribose-phosphate backbone. The spatial arrangements of residues in the recognition helices of two proteins are very similar to

Alignment for AF1382 and F93

Figure 3.8 The sequential alignments between AF1382 and STIV F93 sequences showing 13% identity.

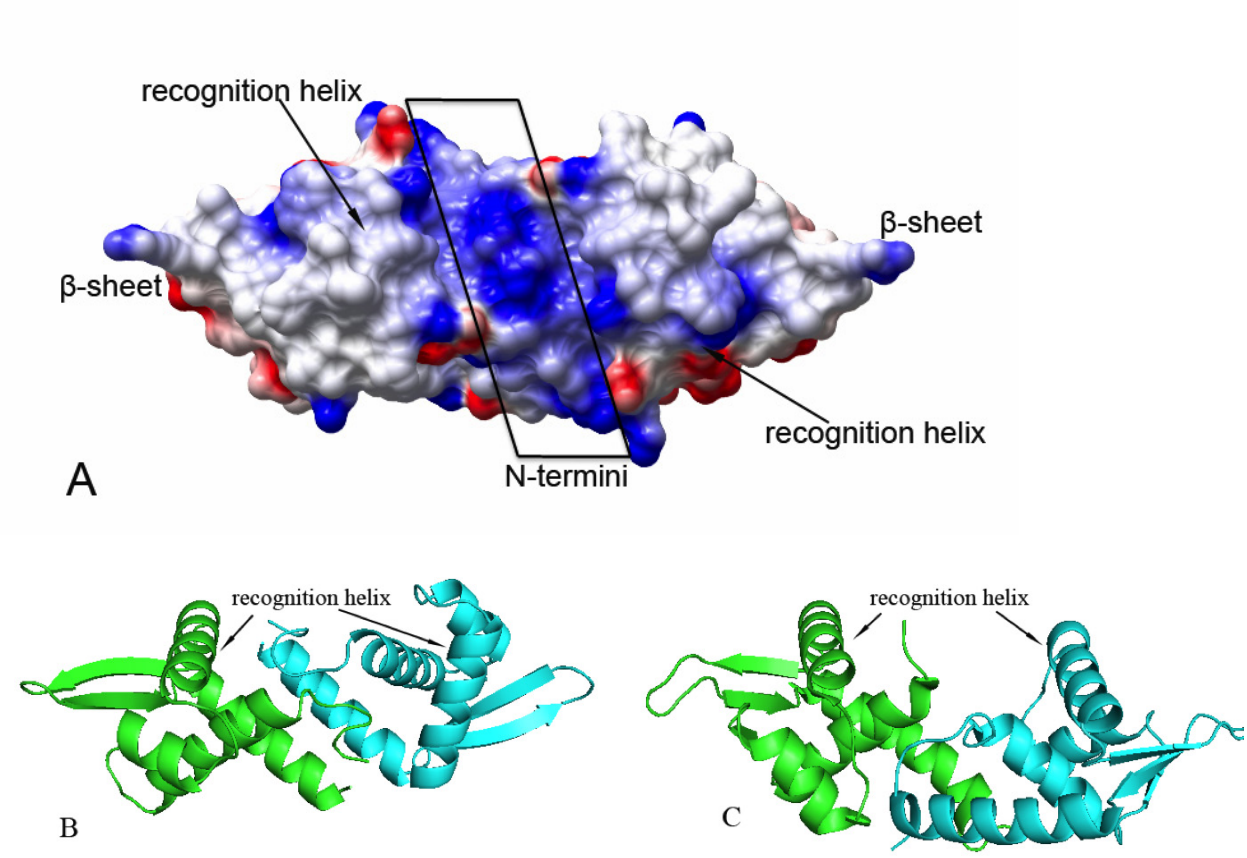


Figure 3.9 Putative interactions with DNA in AF1382. Panel A and B represent the same orientation. A) A surface electrostatic potential representation of the AF1382 dimer, predicted by PISA. The positive potentials are colored blue, and the negative potentials are colored red. The N-termini (in the box) and β-sheet areas show positive charges on them, drawn by CCP4MG. B) A cartoon diagram of the crystallographic dimer of AF1382. C) A cartoon diagram of the dimer of STIV F93. The chain with green color in the STIV F93 structure has been superposed to the chain with same color in the AF1382 structure. The orientation of two recognition helices in AF1382 dimer is slightly different from that in the F93 dimer. It may require a slight rotation in DNA to fit in the putative DNA binding region in the AF1382 structure. Panels B and C are drawn by PyMOL.

each other, with RMSD of 0.69 Å. In the surface electrostatic potential representations (Figure 3.9A, generated by CCP4MG (Potterton *et al.*, 2004)), the β-sheet and the N-terminus in AF1382 show positive charges on them. The orientation of two recognition helices in AF1382 dimer is slightly different from that in the F93 dimer (Figure 3.9). It may require a slight rotation in DNA to fit in the putative DNA binding region in the AF1382 dimer structure. Therefore, AF1382 is very likely to have a positive DNA binding ability, which could be confirmed by electrophoretic mobility shift assays in the future.

3.2.9 Phasing Power of Sulfur-SAS

A comparison of electron density map and traced residues calculated by SGXPro between native and SeMet-labeled AF1382 is displayed in Figure 3.10. The orientations and the contour (1σ) are same in both maps. The sulfur-SAS data provided better initial phasing results of the native AF1382 than the results of the SeMet-labeled protein by Se-SAS. Most side chain densities cannot be observed in the SeMet map and the only helix that is correctly traced in the SeMet map is H2. The density of the loop that connects H2 and H3 in the SeMet map is missing while all but two residues have been traced in the same area of the native map. The residues traced in the helix H4 area (not seen from this view) of the SeMet map cannot fit into the geometry of a helix. The density of the two β-stands in the SeMet map cannot be separated and is shorter than that in the native map. The backbone of the β-sheet in the native map is very clear and in a proper geometry. Density of helix α5 area in the SeMet map is connected to the density of a molecule in an adjacent unit cell. Only 4 residues have been traced in the SeMet map while 15 residues have been correctly traced to form a helix in the same area of the native map.

Comparisons above have demonstrated the phasing power of sulfur-SAS comparing to Se-SAS. When targeting average quality crystals, the sulfur-SAS is able to provide the same

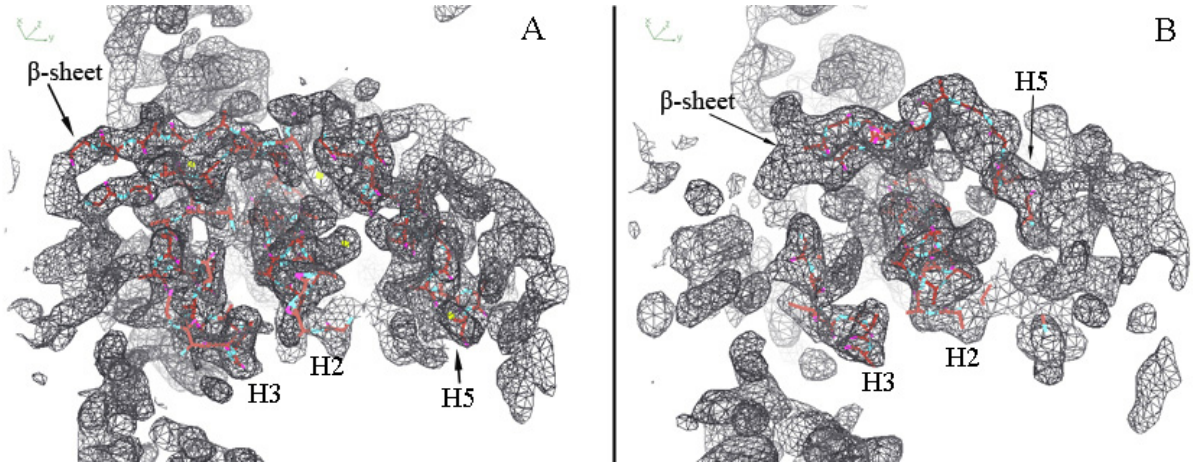


Figure 3.10 Comparison of the initial electron density maps and traced residues of the native and the SeMet-labeled AF1382. Both results are calculated by SGXPro. The maps are contoured at 1 σ. The native AF1382 has provided better phasing results by sulfur-SAS than the SeMet-labeled protein by Se-SAS. A) The initial electron density maps and traced residues of the native AF1382. Two β-strands and most α-helices have been docked in the electron density map. B) The initial electron density maps and traced residues of the SeMet-labeled AF1382. Only helix H2 has been correctly docked. Density of helix H5 is connected to the density of residues from another unit cell. Density of two β-strands cannot be separated and only a small part of the β-strands have been docked into the electron density map.

Table 3.8 The structures solved by sulfur-SAS using the synchrotron X-rays. All the structures are sorted by the years of their deposition.

PDB ID	Space group	Resolution of phasing data (Å)	Year of deposition	Reference
1EL4	$P6_2$	2.5	2000	(Sekar et al., 2004)
1GP0	$P4_32_12$	2.5	2001	(Brown et al., 2002)
1I4U	$P2_12_12_1$	2.5	2001	(Gordon et al., 2001)
1O81	$P4_22_12$	1.5	2002	(Micossi et al., 2002)
1O6J	$P4_12_12$	2.35	2002	(Micossi et al., 2002)
1LPL	$P6_122$	2.5	2002	(Li et al., 2002)
1P5S	$P2_1$	2.2	2003	(Wang et al., 2004)
1P65	$P3_221$	2.6	2003	(Doan and Dokland, 2003)
1VKQ	$P3_121$	1.6	2004	(Sekar et al., 2004)
1YAV	$P2_12_12_1$	2.1	2004	To be published
1YNB*	C2	1.76	2005	To be published
1YOC*	$C222_1$	1.7	2005	To be published
2HZG	$P2_12_12_1$	2	2006	To be published
2CL2	$P2_12_12_1$	2	2006	(Vasur et al., 2006)
2CG6	$P4_12_12$	2.1	2006	To be published
2GNN	$P4_122$	3	2006	(Pieren et al., 2006)
2OQN	$P4_12_12$	1.9	2007	(Kim et al., 2007)
2QVO	$P4_2$	2.3	2007	To be published

level of accuracy as Se-SAS or even higher (Sekar *et al.*, 2004), with proper strategies in the structure determination. High-resolution data with minimal noise are usually expected for the success of sulfur-SAS phasing experiments. Since 2003, most structures that are solved by sulfur-SAS using synchrotron X-rays have about 2 Å or higher resolutions (Table 3.8) with the advances in technology. However, resolution is not an indicator of anomalous scattering signal and only the lower resolution range is needed to determine locations of the anomalous scatterers and to calculate initial phases. For the crystals that have average diffraction quality, such as medium resolution, it is more efficient and reasonable to merge and average multiple data sets from single or multiple crystals to achieve higher R_{as}, rather than to push the efforts of producing crystals of better diffraction quality.

3.3 Structural and Functional Aspects of TT0030

3.3.1 Gene cloning of TT0030

The gene encoding *Thermus thermophilus* ORF 0030 (TT0030) was amplified from the genomic DNA of *Thermus thermophilus* by PCR and cloned into plasmid pDest-527 with N-terminal His_6 tag and TEV cleavage site in between. It was then transformed into *E. coli* host strain BL21 DE3 RPX (Stratagene).

The protein is composed of 138 amino acids with pI of 5.65. The only sulfur atom in the sequence is from Met1. Therefore, it is not suitable for sulfur-SAS or Se-SAS phasing. Thus, heavy metal incorporation to native protein crystal is required for phasing.

3.3.2 Expression, Purification and Crystallization of Native TT0030

Cells carrying TT0030 were first grown in 50 mL LB medium with 100 µg/mL ampicillin at 37 ºC for 8 hours and transferred to 1 L LB medium. The expression was induced with 0.5 mM IPTG at OD_{600} ~ 0.5. Three hours after the IPTG induction, cells were harvested

by centrifugation (6000 rpm × 15 min) and pellets were stored in -80 ºC. Expression of native TT0030 was confirmed by SDS-PAGE (Figure 3.11).

Purification started with 5.0 g biomass from 1 L *E. coli* culture. The cell pellets of native TT0030 were re-suspended in 25 mL Ni affinity binding buffer and lysed by ultrasound (6 × 30 s, on ice) in the presence of 5 mM β-ME and 1 mM PMSF. The sample was centrifuged at 12,000 rpm for 30 minutes at 4 ºC. The supernatant was then filtered through 0.45 μm filters and loaded to a 5 mL HiTrap™ Chelating HP Column (GE Healthcare, Ni-affinity). The protein was eluted with a 0-500 mM imidazole gradient at 4 ºC. The eluted fractions were pooled and subject to 24-hour TEV cleavage at room temperature while they were dialyzed against lysis buffer. After TEV treatment was complete, the mixture was heated to 65 ºC for 1 h, followed by centrifugation and filtering. The supernatant was applied to a Ni-affinity chromatography column and flow-through fractions were collected. No ion exchange chromatography was required. The flow-through fractions from the 2nd Ni-affinity chromatography were pooled and dialyzed against gel filtration buffer pH 7.6. Purity of protein peak fractions in each chromatography purification step was examined by SDS-PAGE (Figure 3.11). The final protein sample was concentrated to 1500 μL at 28.3 mg/mL.

Initial crystallization screening trials were set up in microbatch mode against 384 conditions as described in 2.5 and incubate at 18°C. Crystals of the native TT0030 were observed in many conditions from initial screenings. A limited numbers of optimizations were carried out based on the crystals' diffraction quality. The best-diffracting crystals (Figure 3.12) were obtained from optimization conditions of Crystal Screen Cryo-38 (0.09 M Na HEPES pH 7.5, 1.26 M sodium citrate, and 10% glycerol).

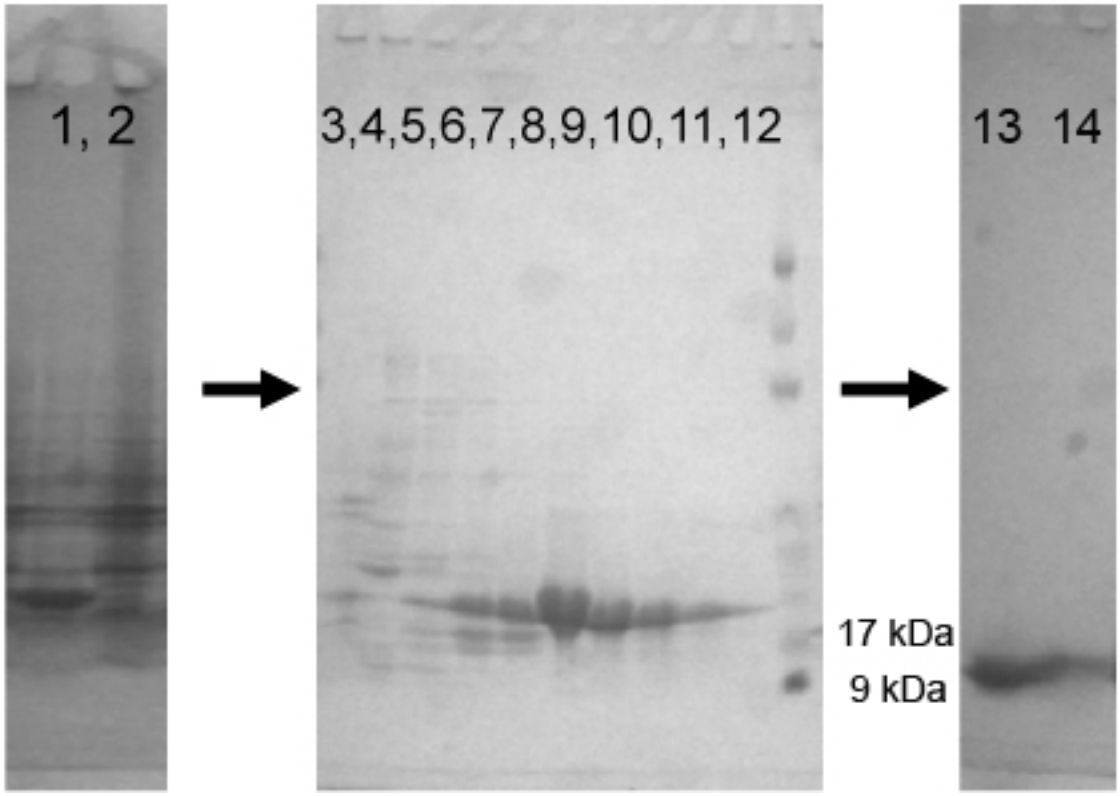


Figure 3.11 Purification of the native TT0030. Lane 1: induced expression; lane 2: un-induced cell culture as a control; lane 3-12: fractions of elution peak from Ni affinity chromatography; lane 13-14: fractions of flow-through from 2nd Ni chromatography run after TEV cleavage and heat treatment.

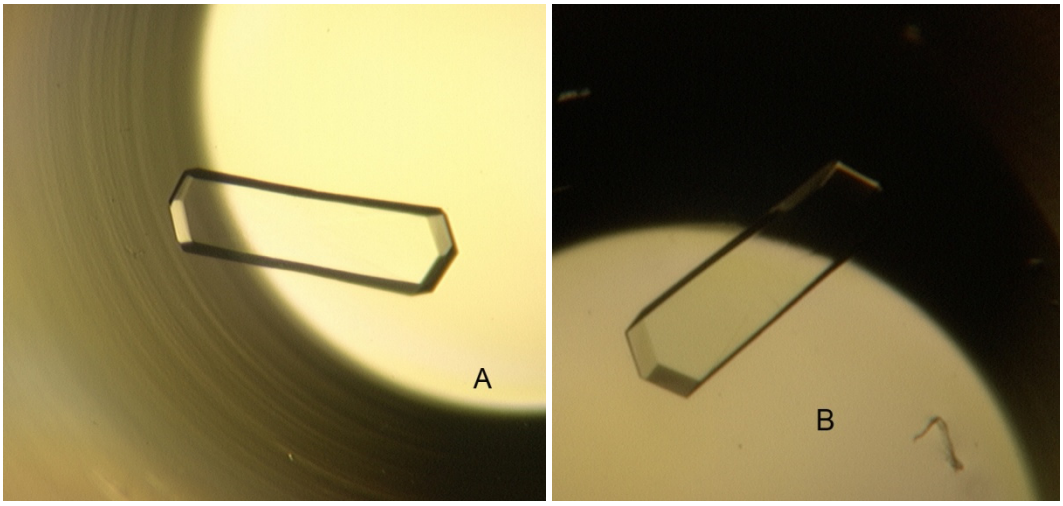


Figure 3.12 The best diffraction crystals of the native TT0030. A) A crystal observed from optimization condition 34 of Crystal Screen Cryo-38; B) a crystal observed from optimization condition 41 of Crystal Screen Cryo-38.

Table 3.9 The heavy atom soaking conditions for native TT0030 crystals.

Condition number	Compositions of the heavy atom soaking condition
1	K_2PtCl_6
2	KI, K_2PtCl_6
3	$C_{13}H_{17}HgNO_6$, $HgCl_2$
4	$HgCl_2$, K_2OsO_4
5	$Hg(OOCCH_3)_2$, $C_9H_9HgNaO_2S$
6	$Hg(OOCCH_3)_2$
7	$C_9H_9HgNaO_2S$
8	$KAu(CN)_2$, $KAuCl_4 \cdot XH_2O$
9	$KAuCl_4 \cdot XH_2O$, $NaAuCl_4 \cdot 2H_2O$
10	$NaAuCl_4 \cdot 2H_2O$, $HAuCl_4$
11	$HAuCl_4$
12	$KAuBr_4 \cdot 2H_2O$, $AgNO_3$
13	$KAuBr_4 \cdot 2H_2O$, $TlCl_3 \cdot XH_2O$
14	$Pb(CH_3CO_2)_2 \cdot 3H_2O$, $AgNO_3$
15	$TlCl_3 \cdot XH_2O$, $Pb(NO_3)_2$, $AgNO_3$
16	$TlCl_3 \cdot XH_2O$
17	CdI_2, K_2IrCl_6
18	$KAuBr_4 \cdot 2H_2O$, CdI_2, K_2IrCl_6

To produce isomorphous derivative crystals, 18 combinations (Table 3.9) of heavy atom salt powders from Heavy Atom Screen kit HR2-442, HR2-444, HR2-446, and HR2-448 (Hampton Research) were added to optimization drops after mountable crystals were observed. The soaking time was 30 minutes.

3.3.3 Data Collection and Processing of TT0030

46 crystals were mounted from 18 heavy atom soaking conditions and 2 native crystals were kept for refinement. All the crystals were mounted by CryoLoop (Hampton Research) and flashed cool to 100 K. Cryoprotection was applied to crystals that showed ice ring during X-ray diffraction. Their diffraction quality was screened by in-house copper X-ray source. The best diffraction crystals from individual heavy atom soaking conditions were shipped to SER-CAT at APS for data collection.

Diffraction data were collected on 22-ID, SER-CAT. X-rays at 0.984 Å wavelength were chosen because anomalous signal of heavy atoms are very strong at this wavelength. One 360-degree data set was collected with 1º oscillation steps and 1 second exposure time on each derivative crystal. The crystal-to-detector distance was set at 190 mm. One set of refinement data was collected with the same settings on a single native crystal.

All the data were index, integrated and scaled by HKL2000. Patterson maps and the ratios of anomalous signal to noise from different isomorphous derivative data were calculated by XPREP (Schneider *et al.*, 2002) to identify heavy atom incorporation in the crystals.

Statistics of data processing were presented in Table 3.10. The resolution of the phasing data was 1.9 Å. The isomorphous derivative crystal belonged to space group I4. The unit cell dimensions were a = b = 89.50 Å, c = 66.72 Å. The solvent content and Matthews coefficient

Table 3.10 Statistics of data processing for TT0030.

a, $R_{sym} = \Sigma|I - \langle I \rangle|/\Sigma I$, where I is the observed intensity of reflections.

Statistics	Isomorphous derivative	Refinement
X-ray wavelength (Å)	0.984	0.984
Space group	I4	I4
Unit cell dimensions (Å)	a = b = 90.21, c = 66.95	a = b = 89.58, c = 66.69
Resolution range (highest resolution shell)	50.00-1.9 (1.97-1.9)	50.00-1.5 (1.55-1.5)
Completeness (%)	99.9 (100.0)	71.8 (4.0)
Redundancy	14.5 (12.5)	12.2 (1.7)
R_{sym} (%)[a]	4.0 (36.1)	4.6 (32.8)
I/sigI	59.8 (6.54)	57.1 (1.96)
Mosaicity (°)	0.542	0.50
Unique Reflections	21161	30385
Reflections measured	307485	371031

(Vm) of isomorphous derivative crystal was estimated to be 42.9% and 2.15 $Å^3$/Da. There were two protein molecules in each asymmetric unit.

The resolution of the refinement data was 1.85 Å. The native crystal belonged to space group I4. The unit cell dimensions of native crystal were a = b = 89.58 Å, c = 66.68 Å.

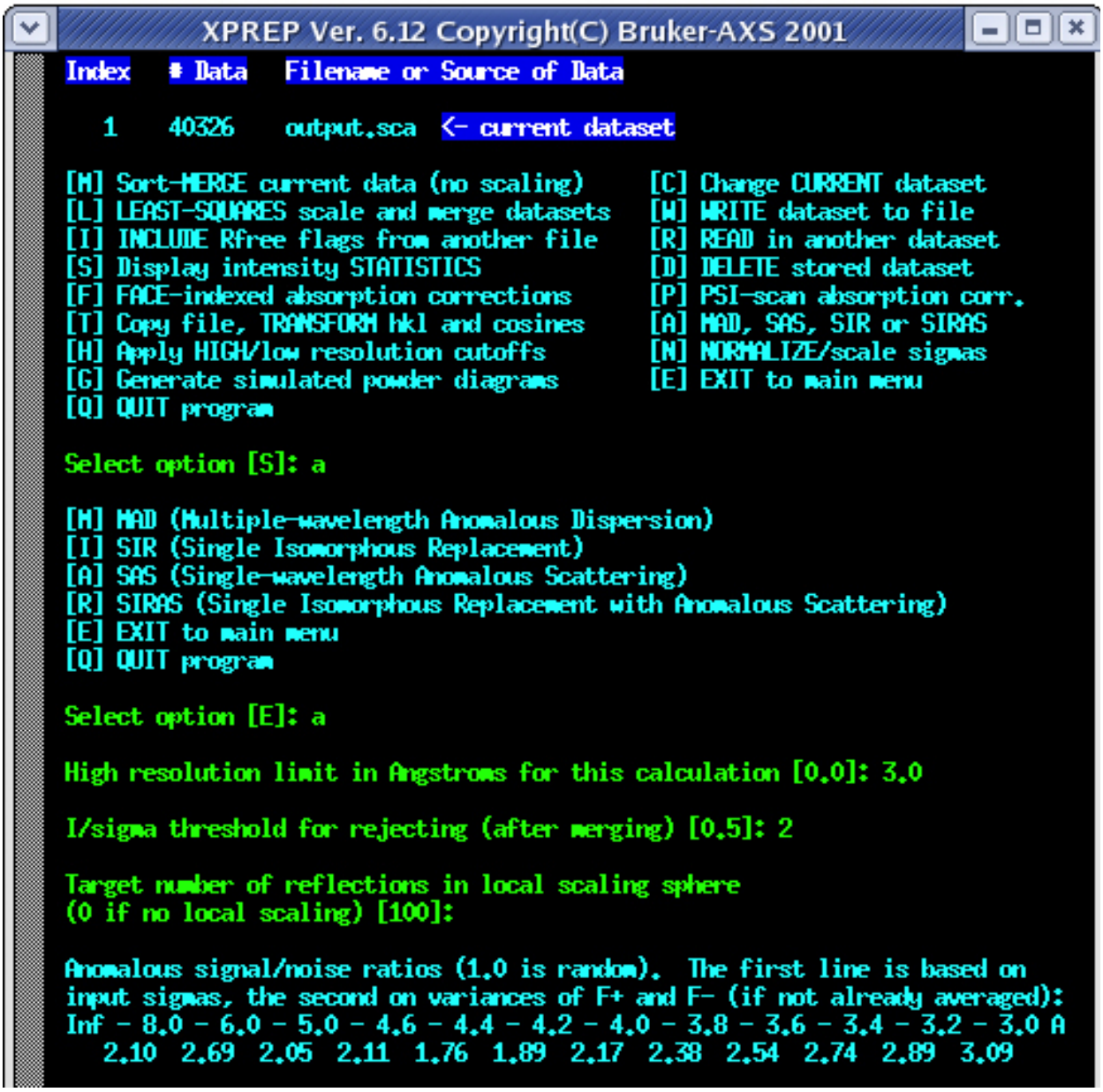
A quick calculation of the ratios of anomalous signal to noise in different resolution ranges (Figure 3.13A) indicated a successful isomorphous heavy atoms incorporation using $KAu(CN)_2$ and $KAuCl_4$ for 30 minutes. The incorporation of Au atoms were further confirmed by the sharp peaks (contoured at 3σ) in Patterson map at z = 0 (Figure 3.13B).

3.3.4 Structure Determination of TT0030

An isomorphous derivative data set was submitted to the Sca2structure high-throughput structure determination pipeline. In the Sca2structure results, two Au sites were found from the SOLVE analysis and a fitted sequence (64.5% complete) was automatically generated by RESOLVE fitting. A two-fold NCS axis was located between two molecules within an asymmetric unit and was used with DM to average electron densities of two chains. Initial refinement was carried out by REFMAC. Then an electron density map and a model were provided to ARP/wARP for automatically model building and sequence fitting. The resulting model was 94.2% complete and was manually adjusted and refined using Coot and REFMAC, respectively.

When phases of refined isomorphous derivative model were extended to native data, the rigid body refinement could not be performed as the free R-factor kept increasing. Thus, using refined isomorphous derivative model as a search model, molecular replacement solution was calculated based on the refinement native data. Using the 1.5 Å data, the structure was refined and adjusted against data between 20.0 and 1.60 Å by REFMAC and COOT, respectively. NCS

A


XPREP Ver. 6.12 Copyright(C) Bruker-AXS 2001
Index # Data Filename or Source of Data
1 40326 output.sca <- current dataset
[M] Sort-MERGE current data (no scaling)
[L] LEAST-SQUARES scale and merge datasets
[I] INCLUDE Rfree flags from another file
[S] Display intensity STATISTICS
[F] FACE-indexed absorption corrections
[T] Copy file, TRANSFORM hkl and cosines
[H] Apply HIGH/low resolution cutoffs
[G] Generate simulated powder diagrams
[Q] QUIT program
[C] Change CURRENT dataset
[W] WRITE dataset to file
[R] READ in another dataset
[D] DELETE stored dataset
[P] PSI-scan absorption corr.
[A] MAD, SAS, SIR or SIRAS
[N] NORMALIZE/scale sigmas
[E] EXIT to main menu
Select option [S]: a
[M] MAD (Multiple-wavelength Anomalous Dispersion)
[I] SIR (Single Isomorphous Replacement)
[A] SAS (Single-wavelength Anomalous Scattering)
[R] SIRAS (Single Isomorphous Replacement with Anomalous Scattering)
[E] EXIT to main menu
[Q] QUIT program
Select option [E]: a
High resolution limit in Angstroms for this calculation [0.0]: 3.0
I/sigma threshold for rejecting (after merging) [0.5]: 2
Target number of reflections in local scaling sphere
(0 if no local scaling) [100]:
Anomalous signal/noise ratios (1.0 is random). The first line is based on
input sigmas, the second on variances of F+ and F- (if not already averaged):
Inf - 8.0 - 6.0 - 5.0 - 4.6 - 4.4 - 4.2 - 4.0 - 3.8 - 3.6 - 3.4 - 3.2 - 3.0 A
2.10 2.69 2.05 2.11 1.76 1.89 2.17 2.38 2.54 2.74 2.89 3.09

Patterson section Z = 0.0000 for SAS delta(F) -> foo.hkl
Space group: I4
Cell: 89.496 89.496 66.719 90.000 90.000 90.000
+X down, +Y across, 256 x 256 grid, contour interval = 3.0 sigma
Super-sharpened, origin removed

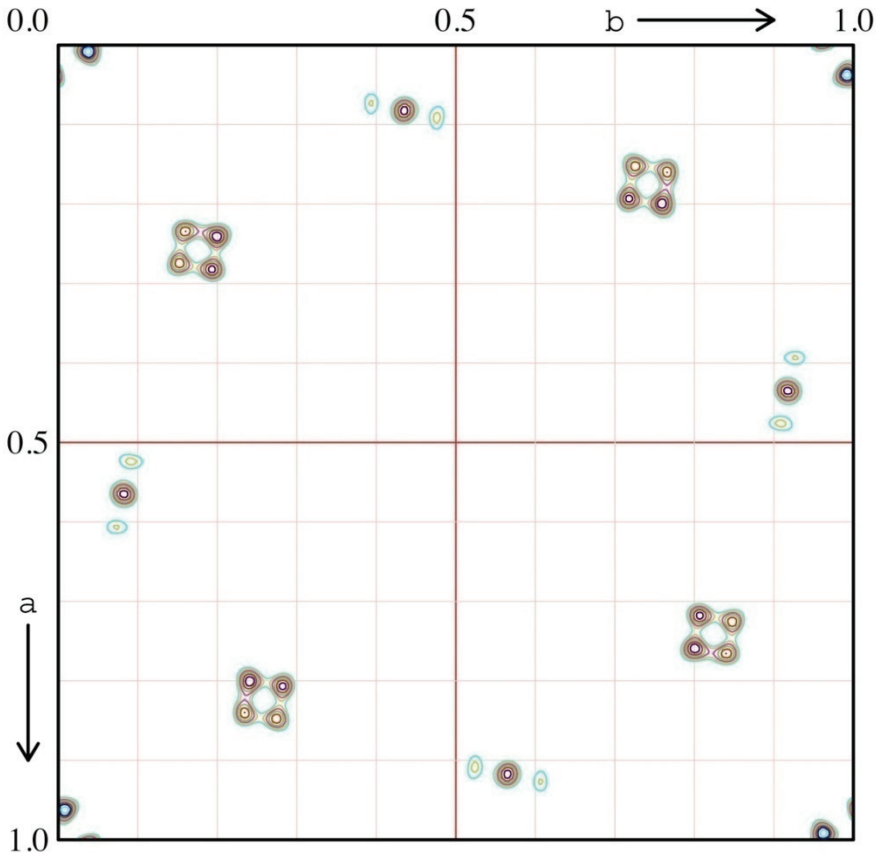


Figure 3.13 Confirmation of heavy atoms incorporation into the native TT0030 crystal. A) The calculated ratios of anomalous signal to noise shows strong anomalous signal in all the resolution ranges. B) Strong peaks at 3σ in the Patterson map confirm the successful incorporation of heavy atoms into the native crystal.

restraints were applied and gradually loosened during the refinement. When the free R-factor was reduced below 30%, water molecules with peak heights above 3.2 and good hydrogen bonding geometries were identified by using ARP/wARP and added to the model. Validation of the final protein model was carried out by MOLPROBITY and PROCHECK before deposition of the model coordinates and structure-factor amplitudes to PDB.

3.3.5 Structure of TT0030 is Similar to a Rossmann Fold

The final model includes 2 chains, 259 of 276 amino acids, 1,995 protein atoms, and 173 water molecules. Electron density of 17 residues, including Met1, Glu48, Glu49, Gln136, Ala137, Ala138, Met1', Ala2', Leu25', Ala26', Gln27', Asp28', Pro29', Glu30', Glu48', Ala137', Ala138', cannot be observed in the map. Statistics of refinement were presented in Table 3.11. The R_{work} factor is 26.8% and R_{free} factor is 27.9% using 5% of total reflections as a test data set. The model has good stereochemistry with RMSDs in bond length and angles of 0.009 Å and 1.16°, respectively. Analysis by Molprobity showed 97.6% residues in flavored regions of Ramachandran plot and very low clashscore of 5.99. The Molprobity score is 1.41 and in the 93rd percentile. Atomic coordinates and structure factors have been deposited with the PDB accession code 2IEL.

The TT0030 monomer is a mixed alpha and beta (α/β) structure with 5 helices and a five-stranded parallel β-sheet (Figure 3.14A). All β-strands are connected to adjacent helices via loops. The helices H1 and H2 are on the same side of β-sheet and the remaining helices are on the other side (Figure 3.14B). Interactions between the two protein molecules in an asymmetric unit are dominated by hydrophobic interactions of the residues in helices and loops as well as 18 hydrogen bonds (Figure 3.15).

Table 3.11 Quality of the present TT0030 model.

a, $R_{work} = \Sigma|F_{obs} - F_{calc}|/F_{obs}$.

b, R_{free} is as for R_{work} but calculated using a 5% test set of reflections excluded from the refinement.

c, 100th percentile is the best.

Refinement	
Number of refined atoms	2168
Number of water molecules	173
R_{work} factor[a]	25.6
R_{free} factor[b]	27.9
R.m.s.d. from ideal	
Bonds (Å)	0.009
Angles (°)	1.156
Mean *B* value	25.84
Atom clash score	5.99
Ramachandran favored	243/249 (no outlier)
Molprobity score	1.41 (93rd percentile)

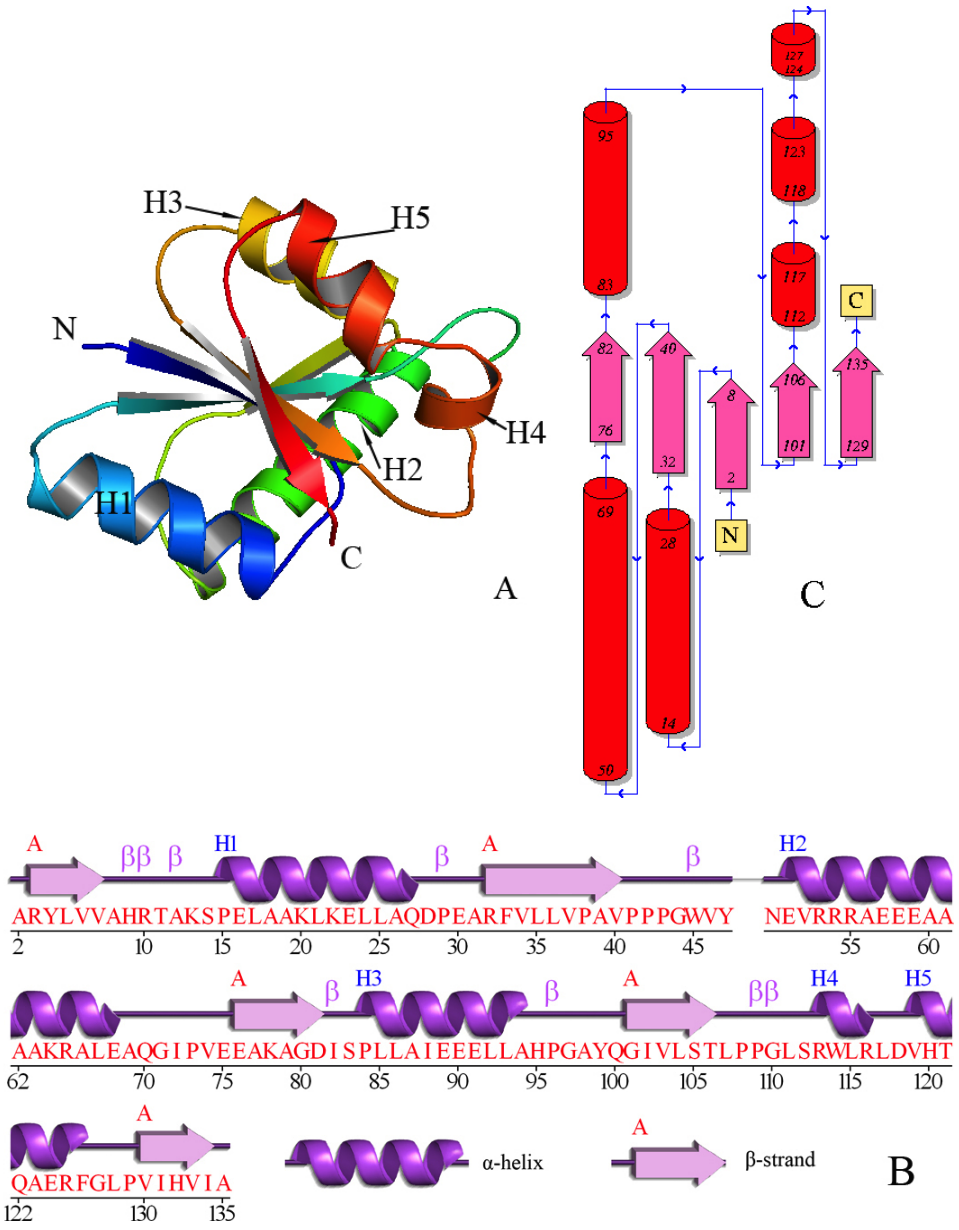


Figure 3.14 Overall structure of TT0030. A) A cartoon representation of the TT0030 monomer structure, drawn by PyMOL. Blue, N-terminus; red, C-terminus. B) The primary sequence of TT0030 annotated with secondary structural elements, generated by PDBsum. C) The topology of TT0030. Cylinder: α-helix; arrow: β-strand; generated by PDBsum. It suggests TT0030 is similar to a Rossmann fold.

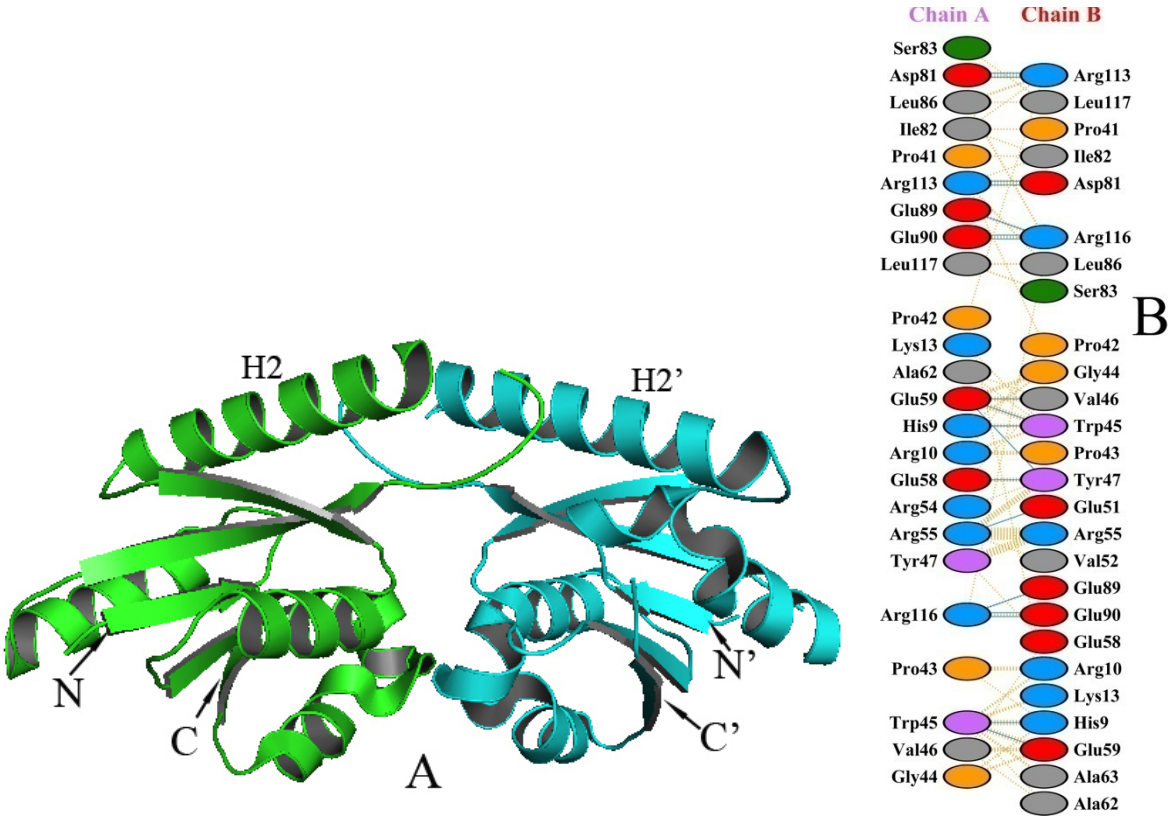


Figure 3.15 Interactions between the two TT0030 molecules in an asymmetric unit. A) The cartoon diagram of two molecules in an asymmetric unit, colored by protein chains, drawn by PyMOL. B) Residue interactions across the interface between two TT0030 molecules in an asymmetric unit. Orange: non-bonded contacts; blue: hydrogen bonds.

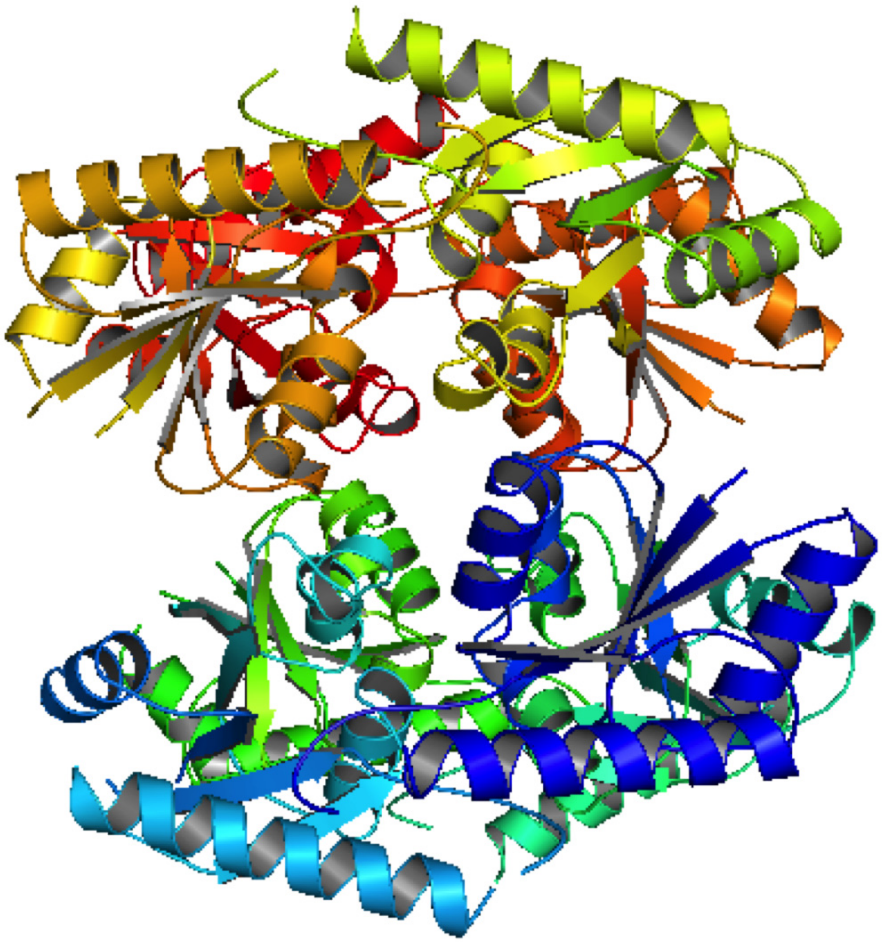


Figure 3.16 The PISA predicted octamer assembly of TT0030 in solution. The cartoon diagram is drawn by PyMOL. Different protein chains are displayed in different colors.

PISA prediction shows that TT0030 is most likely to form an octamer in solutions (Figure 3.16). Formation of the octamer buries 16,370 $Å^2$, or 27.9%, of the total monomer surface area.

A Dali search (Table 3.12) confirmed a 162-residue protein with homologous structure to TT0030: MJ0577 from *M. jannaschii* genome (PDB accession code 1MJH, (Zarembinski *et al.*, 1998)). The Dali Z score and RMSD between two structures are 10.0 and 3.4 Å, respectively. Their structural similarities could be observed in the superposition of two structures (Figure 3.17).

The structure similarities of TT0030 have revealed that it does not represent a new fold. The fold of TT0030 is very similar to the Rossmann fold. All the β-strands in TT0030 are linked by α-helices in the topological order beta-alpha-beta-alpha-beta-alpha-beta-alpha-beta.

3.3.6 Is There any Putative Interaction with ATP?

Sequence searches in Pfam 22.0 and Superfamily server failed to relate TT0030 to any family or superfamily.

A protein chain that belongs to Rossmann fold usually can bind one nucleotide. A protein molecule must have two paired Rossmann folds to bind dinucleotides such as nicotinamide adenine dinucleotide (NAD). According to the fold of TT0030, one can easily predict that it may have putative interactions with nucleotides.

MJ0577 has been confirmed to be a DNA-binding protein (Zarembinski *et al.*, 1998). The sequential alignment generated by ESPript (Figure 3.18) shows only 13% sequence identity between TT0030 and MJ0577. A total of 20 residues (Figure 3.18) are involved in the nucleotide-binding pocket of MJ0577. However, only two identical residues are found in the corresponding positions of TT0030. The surface electrostatic potential representations (Figure

Table 3.12 The Dali search results for structures homologous to TT0030. LALI, number of structurally equivalent residues; %IDE, percentage of identical amino acids over all structurally equivalent residues; Rmsd, root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.

Source	Protein	PDB ID	%IDE	LALI	Rmsd (Å)	Z-score
Aquifex aeolicus	Putative Universal Stress Protein	1Q77	12	117	2.9	11.0
Haemophilus influenzae	universal stress protein	1JMV	16	114	2.9	10.7
Methanococcus jannaschii	MJ0577	1MJH	13	122	3.3	10.4
Arabidopsis thaliana	Universal Stress Protein	2GM3	13	114	4.4	8.4
Pyrococcus horikoshii Ot3	PH1313	1VBK	9	101	3.5	7.4
	T4 polynucleotide kinase	1LTQ	12	101	2.7	7.4
	Glycinamide ribonucleotide synthetase	1GSO	11	90	2.8	7.2

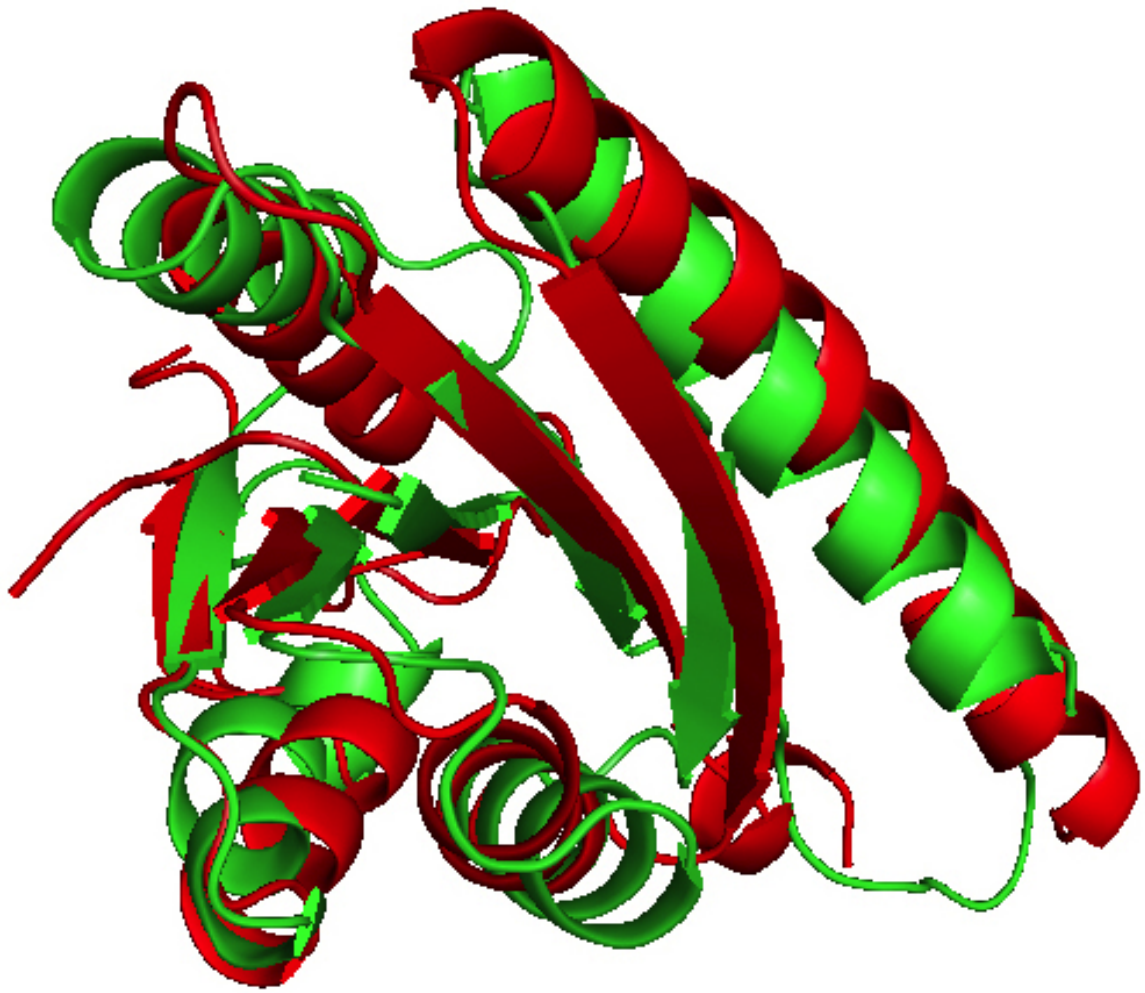


Figure 3.17 Superposition of the TT0030 and the MJ0577 structures. TT0030 is structurally homologous to MJ0577 (PDB accession code 1MJH) (Dali Z score = 10.0, RMSD 3.4 Å). Green: TT0030; red: MJ0577. The ribbon diagram is drawn by PyMOL.

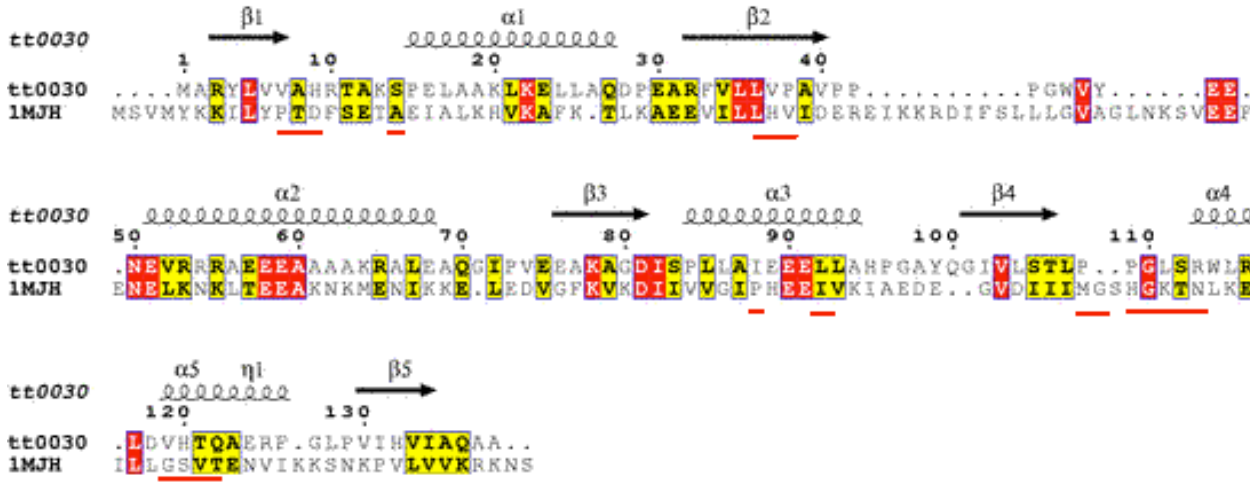


Figure 3.18 The sequential alignments between the TT0030 and the MJ0577 sequences showing 13% identity. The ATP binding residues in MJ0577 are underlined in red.

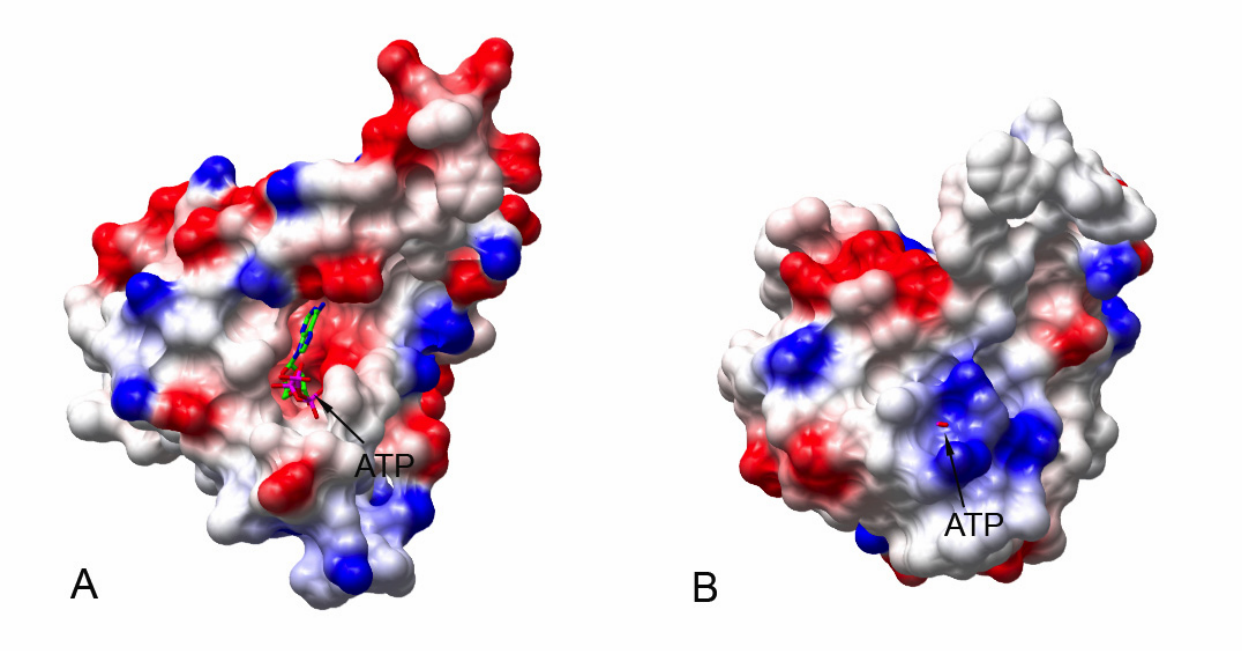


Figure 3.19 The surface electrostatic potential representations of the MJ0577 and the TT0030 monomer structures. The orientations of two monomers are identical to each other. Positive potentials are colored blue, and negative potentials are colored red. Maps are drawn by CCP4MG. A) An ATP molecule has been docked inside the binding pocket of MJ0577. B) The ATP binding pocket within the MJ0577 monomer is buried inside the surface of the TT0030 monomer.

3.19, generated by CCP4MG) of two structures show that the ATP-binding pocket in MJ0577 is buried inside the TT0030 molecule. The oligomerization of MJ0577 is important for the ATP-binding function (Zarembinski *et al.*, 1998). However, the two protein dimmers have different assemblies.

Although the fold of TT0030 and its homologous structure suggest putative interactions with ATP, the structural characteristics cannot provide enough evidence to support this hypothesis. No template for enzyme active site, ligand-binding, or DNA-binding could be found in the TT0030 structure from the 3D functional template searches by ProFunc server. Therefore, TT0030 is probably less likely to function as an ATP-binding protein. Clues about its binding ability to nucleotides might be revealed in the future with more homologous structures deposited in PDB.

3.4 Crystal Structures of Other Non-Pfam Proteins at UGA

In addition to AF1382 and TT0030, four crystal structures of non-Pfam proteins were determined during the pilot stage by other members in Wang's lab. The structure of PH1580 from *Pyrococcus horikoshii* was determined by Y. Li, *et al*. The structure of AF2093 from *Archaeoglobus fulgidus* was determined by J.C. Chang, *et al.* The structure of PF1176 from *Pyrococcus furiosus* was determined by L.Q. Chen, *et al*. The structure of AF0160 from *Archaeoglobus fulgidus* was determined by M. Zhao, *et al*. With permissions from authors, the structures of four non-Pfam proteins were included to explain the structural and biological meanings of this project.

The 152-residue PH1580 monomer (PDB accession code 2HQ4) is a mixed alpha and beta (α/β) protein of 7 α-helices and 9 β-stands (Figure 3.20). The crystallographic data are outlined in Table 3.13. There are two molecules in the asymmetric unit (Figure 3.20), stabilized

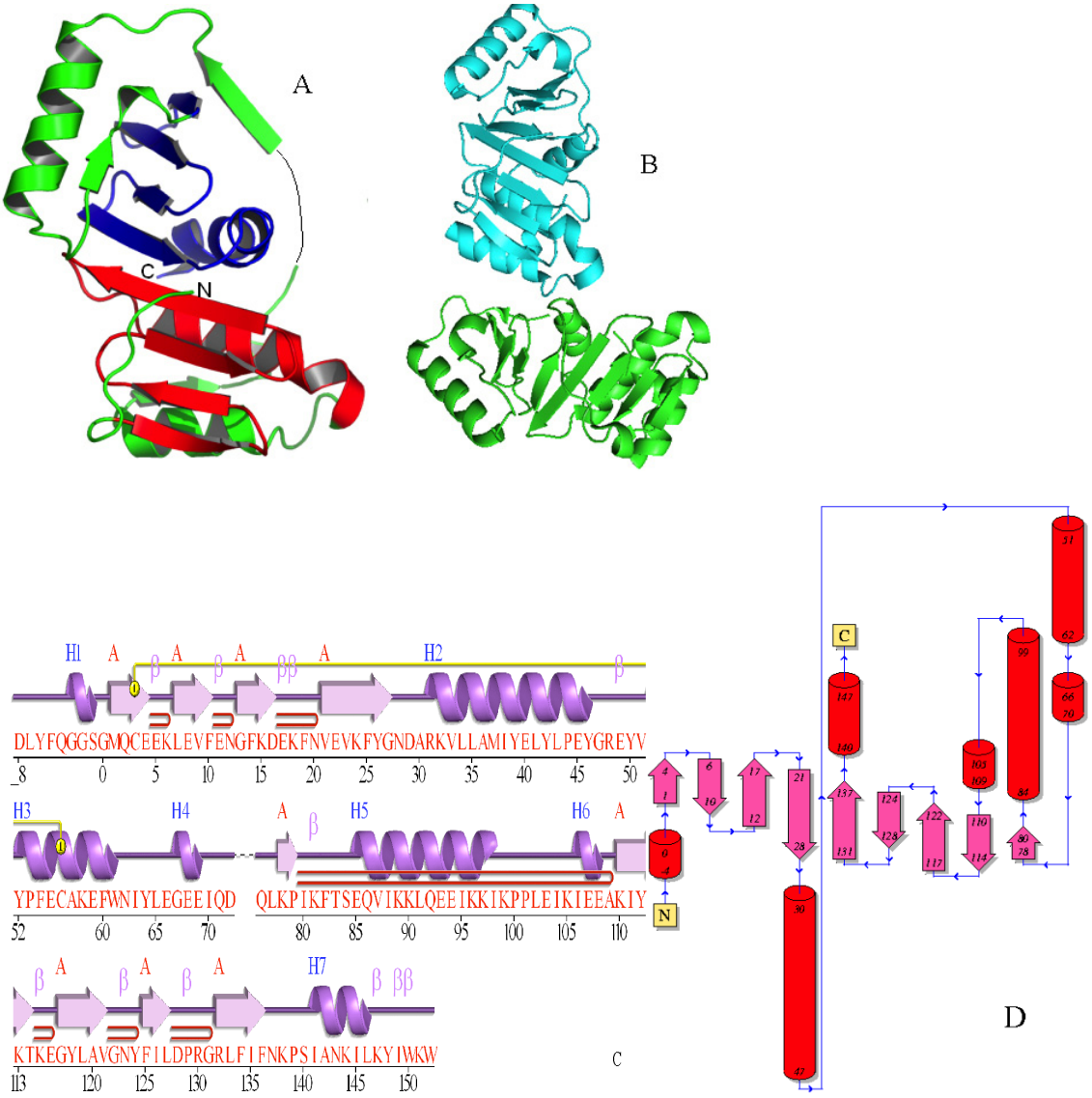


Figure 3.20 Overall structure of PH1580. A) A ribbon diagram of the PH1580 monomer by PyMOL. PH1580 has been recently classified into a new Pfam-A family PH1570 and a new PH1570-like SCOP fold (pre-SCOP, Alexey G. Murzin). B) A ribbon diagram of the two protein molecules in an asymmetric unit drawn by PyMOL, colored by different chains. C) Primary sequence of PH1580 annotated with secondary structural elements, generated by PDBsum. The residues with the number in yellow circles are connected by a disulphide bond. D) Topology of PH1580. Cylinder: α-helix; arrow: β-strand; generated by PDBsum.

Table 3.13 Crystallographic statistics of PH1580.

PDB Entry ID	**2HQ4**
Data used in refinement	
Software	REFMAC 5.2.0019
Resolution range (highest resolution shell) (Å)	20.0-1.99 (2.04-1.99)
Completeness (%)	97.4 (89.48)
Number of reflections	18779 (1825)
R_{work} factor[a]	22.8 (22.0)
R_{free} factor[b]	27.9 (27.8)
Free R value test set size (%)	5.10
R.m.s.d. from ideal values	
Bonds (Å)	0.019
Angles (°)	1.70
Experimental details	
Temperature (K)	100.0
pH	5.20
Number of crystals used	1
Radiation source	APS
Beamline	22-ID
Wavelength (Å)	0.979
Detector type	MAR CCD 300
Intensity-integration software	HKL-2000
Data scaling software	SCALEPACK
Number of unique reflections	28050
Resolution range (highest resolution shell) (Å)	50.0-1.99 (2.04-1.99)
Rejection criteria (Sigma(I))	2.00
Completeness (%)	99.1 (92.3)
Redundancy	6.90 (4.60)
R_{sym} (%)	7.50 (29.3)
<I/Sigma(I)>	11.6 (null)
Method used to determine the structure	Se-SAS
Software used	Sca2Structure
Crystallization condition	Hanging drop vapor diffusion, protein sample (12 mg/mL), 0.2 M ammonium dihydrogen phosphate, 20% w/v PEG 3350, PH 7.5, temperature 291K
Symmetry	C2

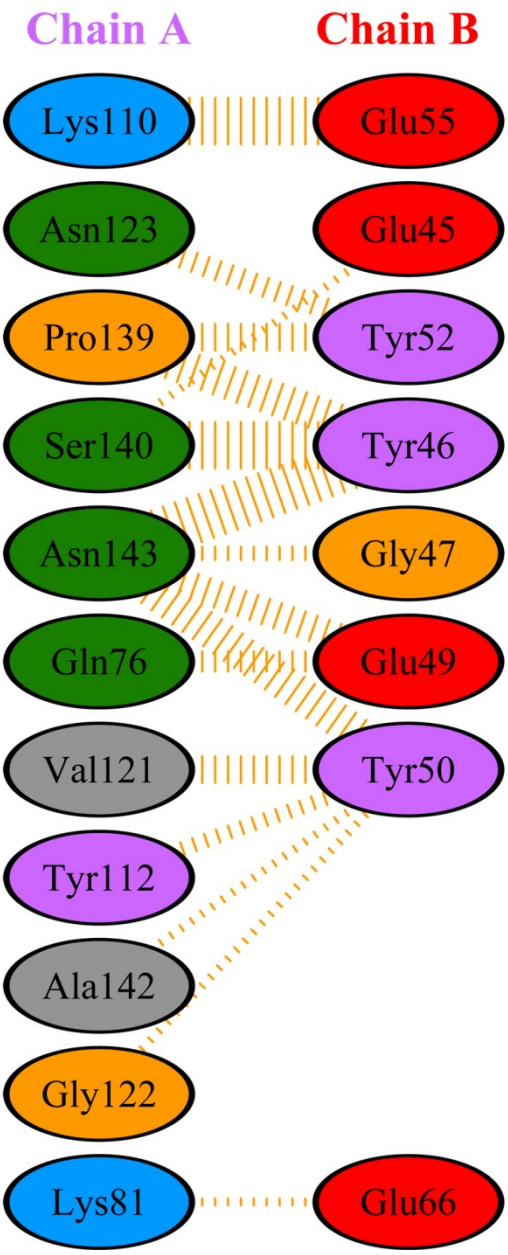


Figure 3.21 Residue interactions across the interface between two PH1580 molecules in an asymmetric unit. Orange: non-bonded contacts.

by hydrophobic interactions only (Figure 3.21). The PISA prediction shows that PH1580 is most likely to form a monomer in solution. PH1580 shows no similarity with any classified structure in the SCOP database. It has been recently classified into a new Pfam-A family PH1570 (accession code PF09638) and a new PH1570-like SCOP fold (pre-SCOP, Alexey G. Murzin). The new fold is described as two beta(4)-alpha structural repeats that are related by pseudo two-fold symmetry, with two extra helices in the linker region. The single 8-standed anti-parallel beta-sheet is formed in the order of 1-2-3-4-8-7-6-5. The structural repeats suggest the possibilities of a gene duplication event in the evolution of this protein family and of the extant non-duplicated relatives.

The 245-residue AF2093 monomer (PDB accession code 2PH7) consists of 9 α-helices and 8 β-stands (Figure 3.22). The crystallographic data are outlined in Table 3.14. 7 residues in a loop and 19 residues at the C-termini are disordered in the crystal lattice. There are two molecules in the asymmetric unit (Figure 3.22), stabilized by hydrophobic interactions and two hydrogen bonds (Figure 3.23). The PISA prediction shows that it is most likely to form a monomer in solution. AF2093 shows no similarity with any classified structure in the SCOP and the PDB database, thus is very possible to represent as a new SCOP fold.

A sequence search shows that the 97-residue PF1176 (PDB accession code 2HJM) belongs to Pfam-B PB173367 that is automatically generated from an alignment taken from Prodom 2005.1 (PD693109). The monomer of PF1176 is an all alpha protein consisting of an up-and-down four-helical bundle (Figure 3.24). The crystallographic data are outlined in Table 3.15. A total of 12 residues, including Lys86, Pro87, Arg88, Pro89, Pro90, Leu91, Leu92, Val93, Asp94, Asn95, Asp96, Leu97, at C-terminus are disordered in the crystal lattice while His_6 tag is directly connected to the protein at N-terminus. There are four molecules in the asymmetric unit

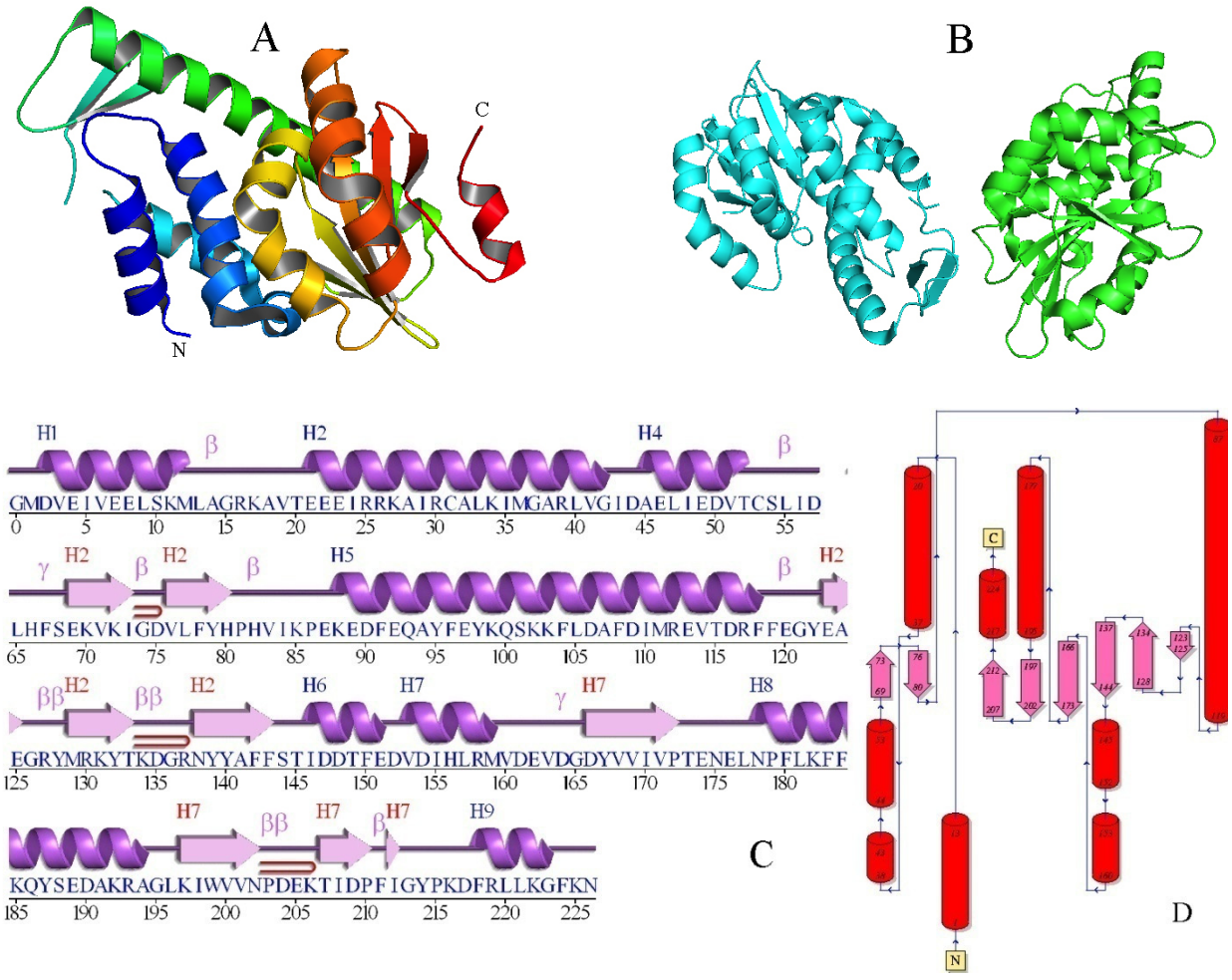


Figure 3.22 Overall structure of AF2093. A) A ribbon diagram of the AF2093 monomer by PyMOL. AF2093 shows no similarity with any classified structure in the SCOP and the PDB database. B) A ribbon diagram of the two protein molecules in an asymmetric unit drawn by PyMOL, colored by different chains. C) Primary sequence of AF2093 annotated with secondary structural elements, generated by PDBsum. D) Topology of AF2093, generated by PDBsum. Cylinder: α-helix; arrow: β-strand; generated by PDBsum.

Table 3.14 Crystallographic statistics of AF2093.

PDB Entry ID	**2PH7**
Data used in refinement	
Software	REFMAC 5.2.0019
Resolution range (highest resolution shell) (Å)	47.4-2.40 (2.46-2.40)
Completeness (%)	99.9 (99.66)
Number of reflections	27283 (1984)
R_{work} factor[a]	23.0 (33.5)
R_{free} factor[b]	27.1 (48.5)
Free R value test set size (%)	5.00
R.m.s.d. from ideal values	
Bonds (Å)	0.010
Angles (°)	1.34
Experimental details	
Temperature (K)	100.0
pH	4.90
Number of crystals used	1
Radiation source	APS
Beamline	22-ID
Wavelength (Å)	0.9724
Detector type	MAR CCD 300
Intensity-integration software	HKL-2000
Data scaling software	SCALEPACK
Number of unique reflections	28734
Resolution range (highest resolution shell) (Å)	47.4-2.40 (2.49-2.40)
Rejection criteria (Sigma(I))	Null
Completeness (%)	99.1 (99.8)
Redundancy	6.70 (5.20)
R_{sym} (%)	6.60 (53.5)
<I/Sigma(I)>	39.3 (2.08)
Method used to determine the structure	Se-SAS
Software used	Sca2Structure
Crystallization condition	Modified microbatch, protein sample (12 mg/mL), 20% PEG 4000, 0.05 M NaCl, 0.05 M Li_2SO_4, 0.1 M NaOAc, PH 4.9, temperature 291K
Symmetry	F23

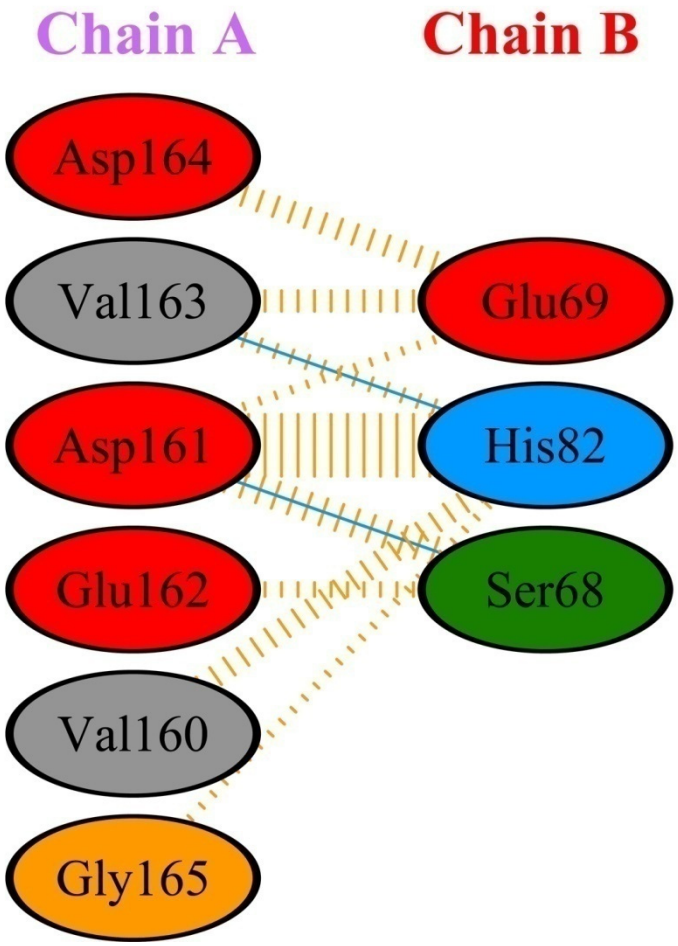


Figure 3.23 Residue interactions across the interface between two AF2093 molecules in an asymmetric unit. Orange: non-bonded contacts; blue: hydrogen bonds.

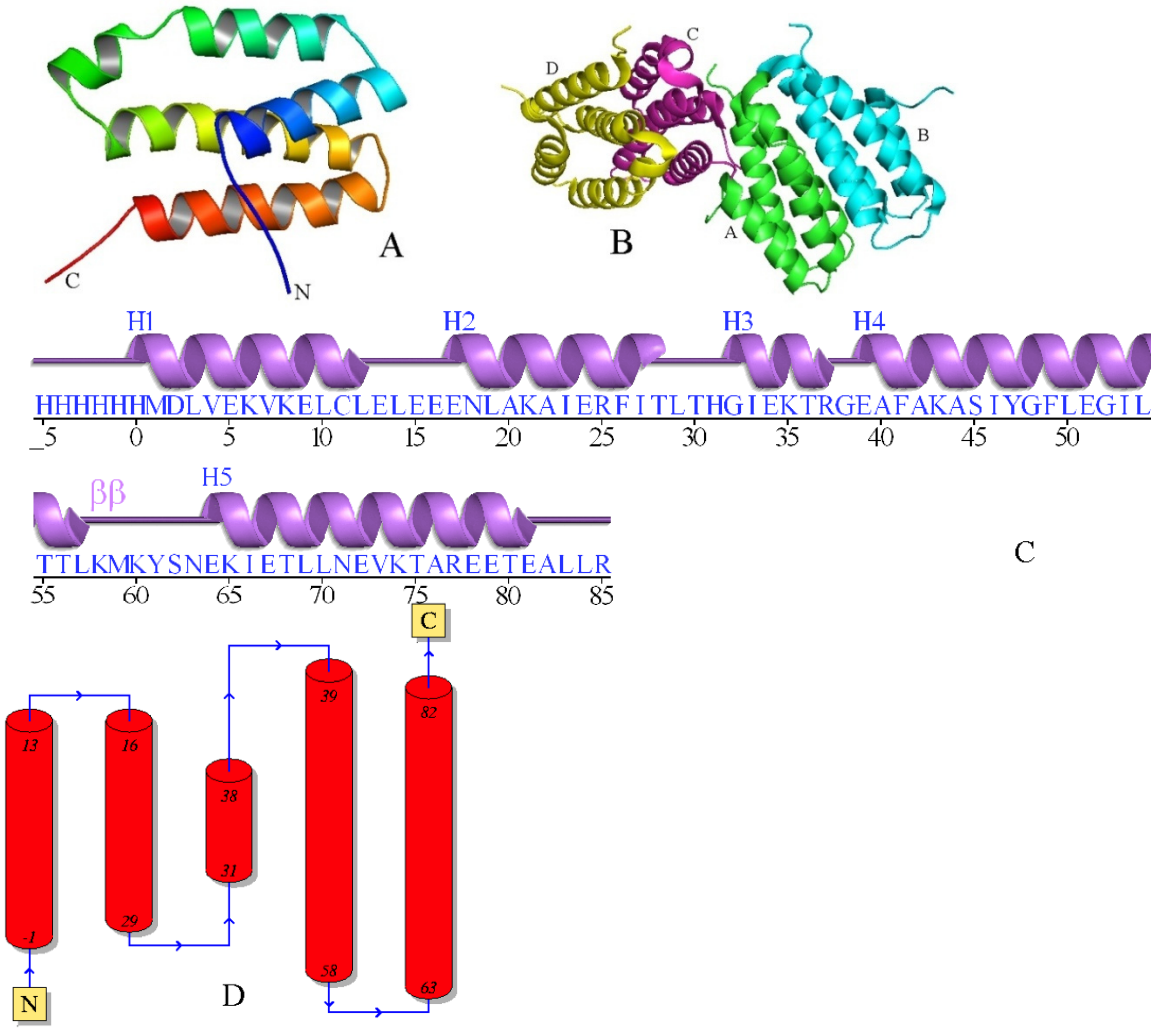


Figure 3.24 Overall structure of PF1176. A) A ribbon diagram of the PF1176 monomer by PyMOL. The monomer of PF1176 is an all alpha protein consisting of an up-and-down four-helical bundle. B) A ribbon diagram of the four protein molecules in an asymmetric unit drawn by PyMOL, colored by different chains. C) Primary sequence of PF1176 annotated with secondary structural elements, generated by PDBsum. D) Topology of PF1176. Cylinder: α-helix; arrow: β-strand; generated by PDBsum.

Table 3.15 Crystallographic statistics of PF1176.

PDB Entry ID	**2HJM**
Data used in refinement	
Software	CNS
Resolution range (highest resolution shell) (Å)	30.0-2.90
Completeness (%)	97.4
Number of reflections	18779
R_{work} factor[a]	22.8
R_{free} factor[b]	27.9
Free R value test set size (%)	5.00
R.m.s.d. from ideal values	
Bonds (Å)	0.007
Angles (°)	1.16
Experimental details	
Temperature (K)	100.0
pH	5.20
Number of crystals used	1
Radiation source	APS
Beamline	22-ID
Wavelength (Å)	0.9724
Detector type	MAR CCD 300
Intensity-integration software	HKL-2000
Data scaling software	SCALEPACK
Number of unique reflections	12787
Resolution range (highest resolution shell) (Å)	50.0-2.70 (2.80-2.70)
Rejection criteria (Sigma(I))	0.0
Completeness (%)	99.9 (98.9)
Redundancy	14.0 (11.5)
R_{sym} (%)	8.20 (60.0)
<I/Sigma(I)>	12.1 (3.50)
Method used to determine the structure	Se-SAS
Software used	Sca2Structure
Crystallization condition	Vapor diffusion, 100 mM sodium citrate, 30% PEG400, PH 5.2, temperature 298K
Symmetry	$P2_12_12_1$

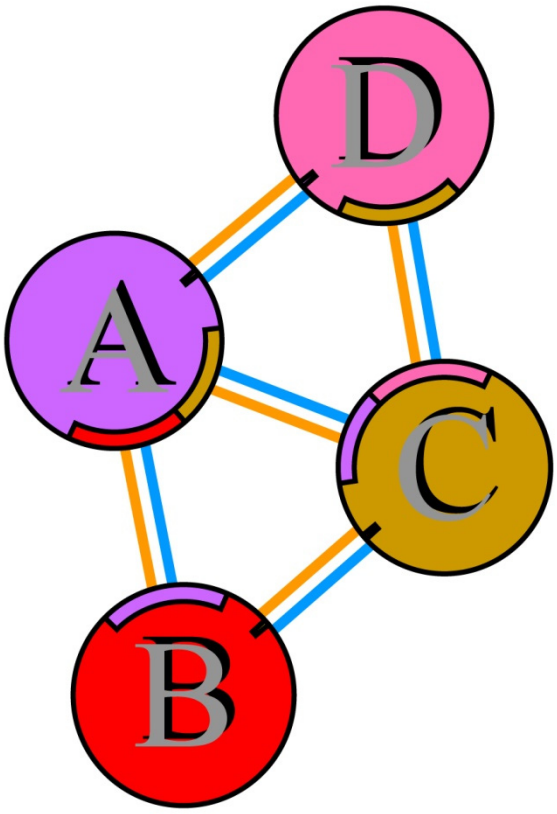


Figure 3.25 Schematic diagram of interactions between four PF1176 molecules in an asymmetric unit. Orange: non-bonded contacts; blue: hydrogen bonds.

Table 3.16 The Dali search results for structures homologous to PF1176. LALI, number of structurally equivalent residues; %IDE, percentage of identical amino acids over all structurally equivalent residues; Rmsd, root-mean-square deviation of Cα atoms in the least-squares superimposition of the structurally equivalent Cα atoms.

Source	Protein	PDB ID	%IDE	LALI	Rmsd (Å)	Z-score
E. coli	Toxin hemolysin E	1QOY	11	85	2.7	7.4
Yeast	ribosomal 80S-eEF2-sordarin complex	1S1H	3	61	1.9	6.8
Tabacco	Cell wall invertase inhibitor	1RJ1	9	87	3.1	6.6
Streptococcus pyogenes	SPy2152	2FU2	7	67	2.5	6.5
Squash *(Cucurbita moschata)*	glycerol-3-phosphate (1)-acyltransferase	1K30	9	65	2.8	6.5
Tabacco	Cell wall invertase inhibitor	2CJ8	11	85	3.0	6.4
Human papillomavirus type 16	Transactivation domain of E2	1DTO	5	66	3.5	6.3
Arabidopsis thaliana	pectin methylesterase inhibitor	1X8Z	5	79	2.8	5.8
synthesized	Due Ferro 1 (DF1)	1EC5	11	46	0.9	5.6

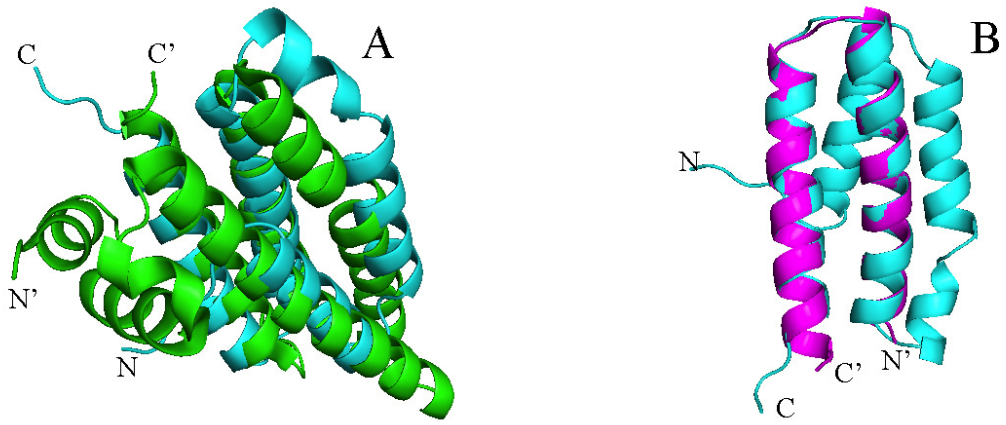


Figure 3.26 Superpositions of the PF1176 structure and its homologous structures. A) Superposition of the structures of PF1176 and a cell wall invertase inhibitor from tobacco (PDB accession code 1RJ1) (Dali Z-score = 6.6, and RMSD = 3.1 Å). Cyan: PF1176 (N, C); Green: 1RJ1 (N', C'). B) Superposition of the PF1176 and the Due Ferro 1 (DF1) (PDB accession code 1EC5) structures (Dali Z-score = 5.6, and RMSD = 0.9 Å). Cyan: PF1176 (N, C); Magenta: 1EC5 (N', C').

(Figure 3.24), connected by hydrophobic interactions and hydrogen bonds (Figure 3.25). The PISA prediction shows that it is most likely to form a dimer in solution. The Dali search (Table 3.16) reveals that PF1176 has similar fold to a cell wall invertase inhibitor from tobacco (PDB accession code 1RJ1) (Hothorn *et al.*, 2004) (Figure 3.26). The Dali Z-score and RMSD between two structures are 6.6 and 3.1 Å, respectively. An interesting observation is that the structure of PF1176 is also very similar to a chemically synthesized model protein Due Ferro 1 (DF1) (PDB accession code 1EC5) (Lombardi *et al.*, 2000) (Figure 3.26). The Dali Z-score and RMSD between two structures are 5.6 and 0.9 Å, respectively. DF1 was designed to be a minimal model for diiron proteins. The structural similarity between PF1176 and DF1 may be used to discover the evolutionary information of PF1176.

The 174-residue AF0160 monomer (PDB accession code 2IDG) is an all-alpha protein consisting of a bundle of 10 helices (Figure 3.27). The crystallographic data are outlined in Table 3.17. A total of 14 residues, including SER161, SER162, LEU163, VAL164, GLY165, GLU166, LYS167, ASN168, GLU169, GLY170, ALA171, ASP172, ASN173, ASN174, at C-terminus are disordered in the crystal lattice. There are three molecules in the asymmetric unit (Figure 3.27), stabilized by hydrophobic interactions, hydrogen bonds and salt bridge (Figure 3.28). The PISA prediction shows that it is most likely to form a dimer in solution. AF0160 is classified into TorD-like superfamily by the SUPERFAMILY web server. Superpositions have shown that it is structurally homologous to several TorD-like proteins: a putative anaerobic dehydrogenase component (PDB accession code 1S9U), the dimeric TorD chaperone from *Shewanella massilia* (PDB accession code 1N1C) (Tranier *et al.*, 2003), and a putative redox enzyme maturation protein from *Archaeoglobus fulgidus* (PDB accession code 2O9X) (Kirillova *et al.*, 2007) (Figure 3.29).

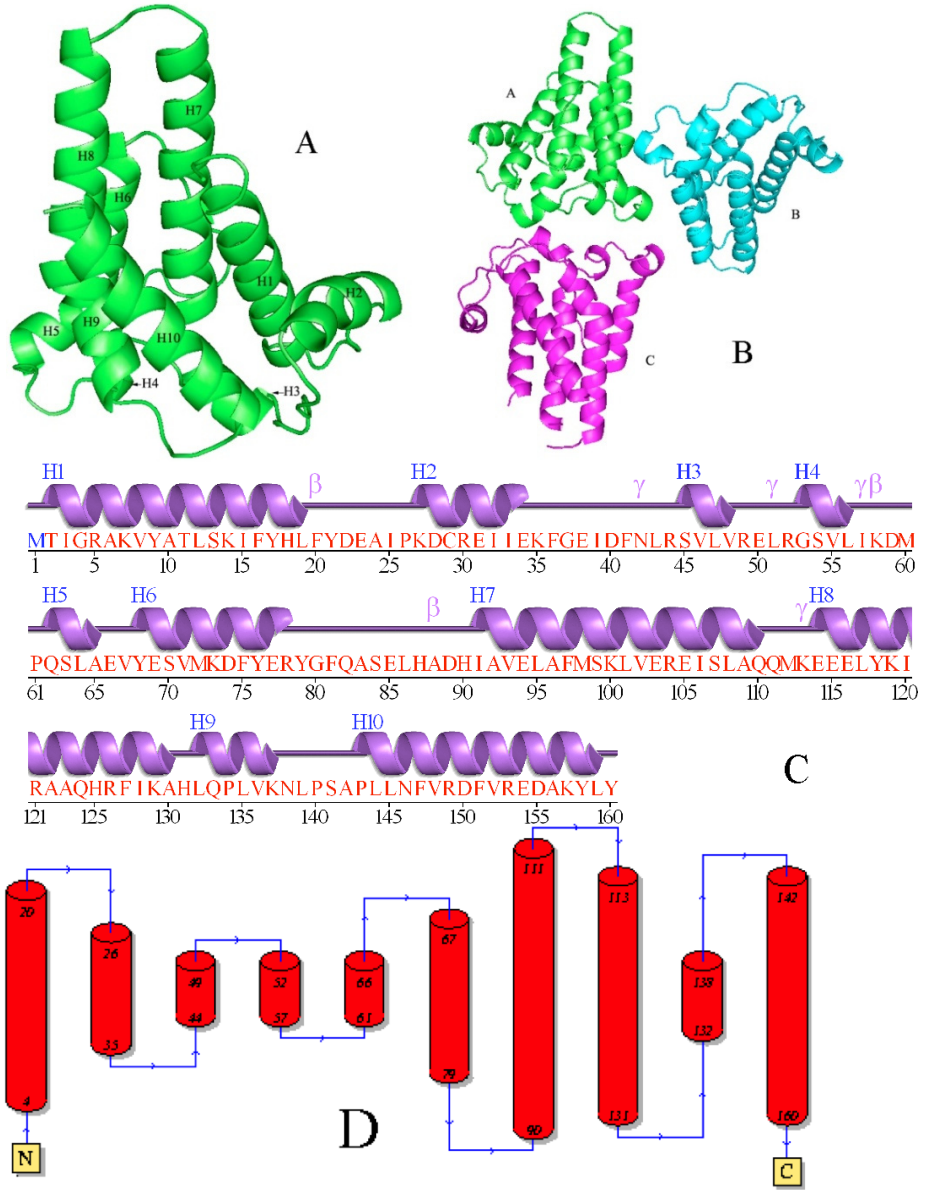


Figure 3.27 Overall structure of AF0160. A) A ribbon diagram of the AF0160 monomer by PyMOL. AF0160 is classified into TorD-like superfamily by SUPERFAMILY web server. B) A ribbon diagram of the three protein molecules in an asymmetric unit drawn by PyMOL, colored by different chains. C) Primary sequence of AF0160 annotated with secondary structural elements, generated by PDBsum. D) Topology of AF0160. Cylinder: α-helix; arrow: β-strand; generated by PDBsum.

Table 3.17 Crystallographic statistics of AF0160.

PDB Entry ID	**2IDG**
Data used in refinement	
Software	REFMAC 5.2.0019
Resolution range (highest resolution shell) (Å)	36.8-2.69 (2.77-2.69)
Completeness (%)	100.0 (100.0)
Number of reflections	16116 (1031)
R_{work} factor[a]	24.4 (35.8)
R_{free} factor[b]	29.7 (53.1)
Free R value test set size (%)	5.90
R.m.s.d. from ideal values	
Bonds (Å)	0.012
Angles (°)	1.36
Experimental details	
Temperature (K)	100.0
pH	6.90
Number of crystals used	1
Radiation source	APS
Beamline	22-ID
Wavelength (Å)	0.979
Detector type	MAR CCD 300
Intensity-integration software	HKL-2000
Data scaling software	SCALEPACK
Number of unique reflections	18724
Resolution range (highest resolution shell) (Å)	50.0-2.69 (2.80-2.69)
Rejection criteria (Sigma(I))	2.00
Completeness (%)	97.0 (89.0)
Redundancy	11.90 (6.30)
R_{sym} (%)	10.60 (41.0)
<I/Sigma(I)>	28.8 (4.12)
Method used to determine the structure	Se-SAS
Software used	Sca2Structure
Crystallization condition	Sitting drop vapor diffusion, protein sample (40 mg/mL), 30% PEG 3350, 0.15 M NaSCN, 0.01 M Spermine-HCl, PH 6.9, temperature 291K
Symmetry	$P2_12_12_1$

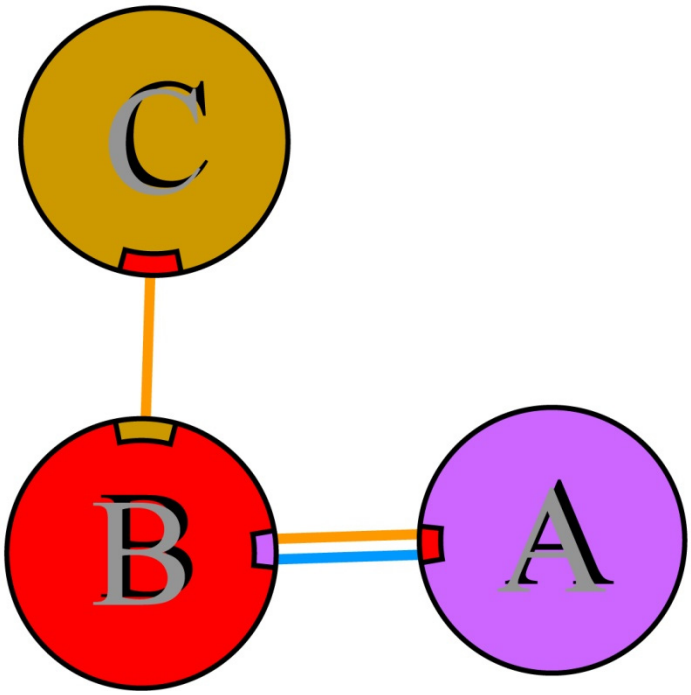


Figure 3.28 Schematic diagram of interactions between three AF0160 molecules in an asymmetric unit. Orange: non-bonded contacts; blue: hydrogen bonds.

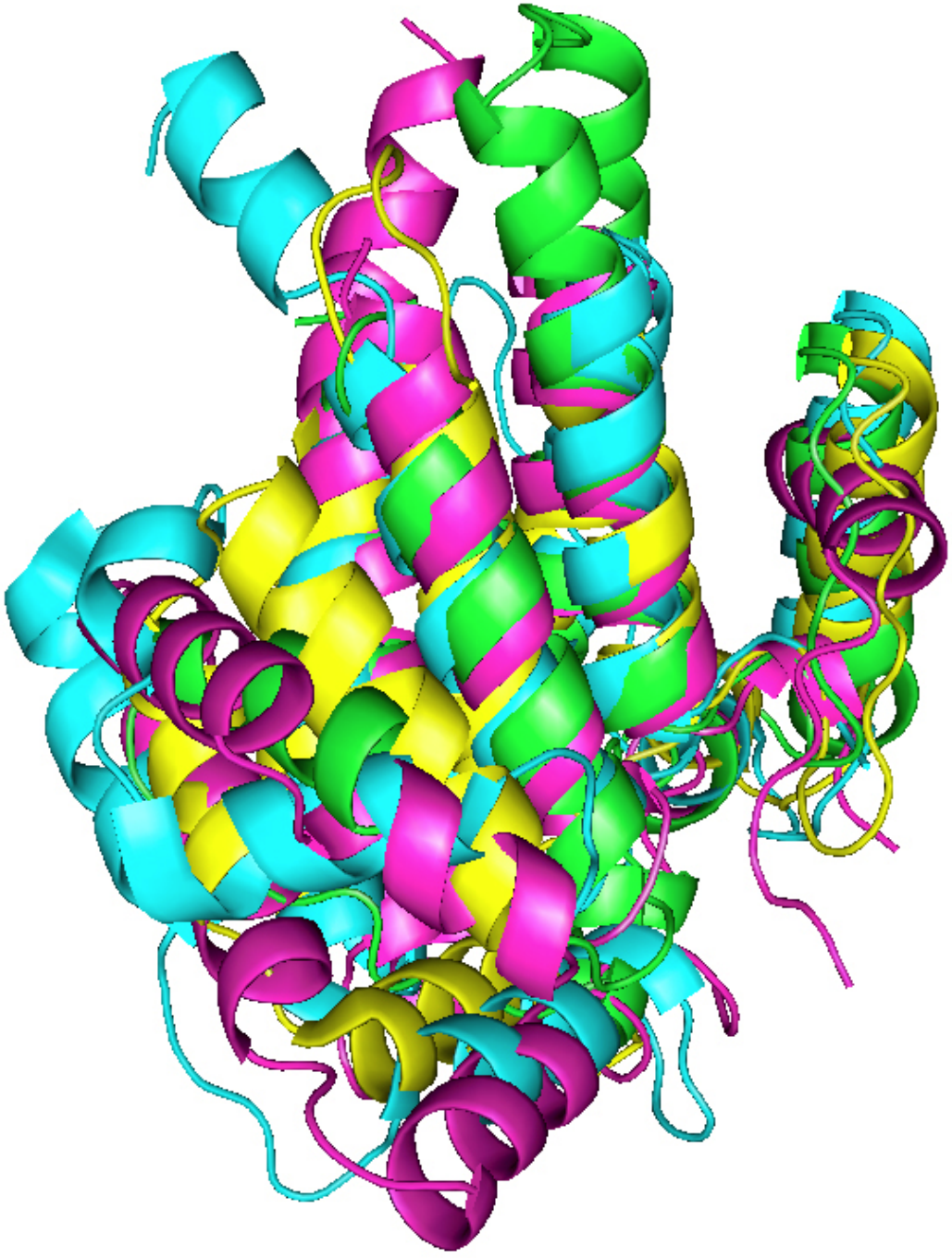


Figure 3.29 The superposition of the structures of AF0160 and several TorD-like proteins. Green: AF0160; Cyan: 1S9U; Magenta: 1N1C; Yellow: 2O9X.

CHAPTER FOUR

Conclusions

Many questions arise regarding the evolutionary information of non-Pfam proteins, which can be indicated by their structural and functional aspects. Due to the poor sequence homology, the unknown functional aspects can only be discovered through the structures of non-Pfam proteins. However, they have been underemphasized in most structural genomics projects from the beginning which mainly focused on the Pfam proteins. The research presented here investigated the structural and functional aspects of non-Pfam proteins from six genomes by means of the high throughput methods employed in the structural genomics projects. In addition, studies of methodology in macromolecule structure determination are conducted.

Protein production of the selected non-Pfam targets in the six genomes has suggested that majority of the non-Pfam sequences are not sequencing errors or fake genes. The cloning and expression success rates for non-Pfam genes are similar to that in Pfam genes at SECSG. Without exclusion of targets having potential difficulties in crystallization, the success rates of purification and crystallization for 62 non-Pfam targets are at the same level of those for Pfam proteins selected by the nine PSI-1 centers. The molecular weight of selected targets ranges from 7-78 kDa, which indicates that both the large and small non-Pfam targets correspond to expressed proteins.

Out of the six non-Pfam structures determined by X-ray crystallography at UGA, one structure (PH1580) has been confirmed as a new SCOP fold. Another structure (AF2093) is very likely to represent a new SCOP fold. Although statistical conclusion cannot be drawn due to the

limited number of the solved non-Pfam structures in the pilot stage, a significant number of new folds could be expected from the remaining targets. An important aim of structural genomics is to identify the protein universe in terms of the possible 3D protein folds. Giving the total number of about one million non-Pfam proteins in Pfam version 22.0, they should not be ignored in the future structural genomics projects. Otherwise, we may not be able to find the "true" protein fold space.

Proteins with homologous structures to the remaining four non-Pfam targets (AF1382, TT0030, PF1176 and AF0160) carry significant biological functions, such as DNA binding, ATP binding, and cell wall invertase inhibition. Although sequence identities between the non-Pfam targets and their structural homologous proteins are limited, the similar folds and structures could suggest putative functions of these non-Pfam targets. The structure similarity between PF1176 and DF1 and the structural repeats within PH1580 could be used to study their evolutionary information. Therefore, majority of the non-Pfam proteins should not be considered as rapidly evolving proteins with nonessential functions. On the contrary, some of them are predicted to play important roles for the uniqueness of these organisms.

The structure of AF1382 is the first one to be solved by sulfur-SAS using multiple synchrotron diffraction data sets. It proves that sulfur-SAS is a viable phasing method when dealing with medium-resolution data. Rather than resolution, the redundancy of the diffraction data is more critical in deciding the success of structure determination by sulfur-SAS. The main reason is that a higher redundancy indicates a lower noise level in the data. A constant monitoring of R_{as} in multiple data sets is useful to design a strategy of data merging. By extending the current limitations, we can see the advantages of sulfur-SAS that only native

crystals are needed. Thus we can avoid the troubles of altering the contents of the unit cell or changing the protein to a non-wild type form.

The unique sequences of non-Pfam proteins have brought us many puzzles to solve. The research presented here could be a good starting point to study the structural and functional aspects of non-Pfam proteins with high throughput methods. With more efforts carried out on them, all the truths behind non-Pfam proteins will hopefully be revealed in the future.

REFERENCES

Abrahams, J. P. and A. G. W. Leslie (1996). "Methods used in the structure determination of bovine mitochondrial F-1 ATPase." Acta Crystallogr D **52**: 30-42.

Adams, M. W. W., H. A. Dailey, L. J. Delucas, M. Luo*, et al.* (2003). "The Southeast Collaboratory for Structural Genomics: A high-throughput gene to structure factory." Accounts of Chemical Research **36**(3): 191-198.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang*, et al.* (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Andreeva, A., D. Howorth, S. E. Brenner, T. J. P. Hubbard*, et al.* (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." Nucleic Acids Research **32**: D226-D229.

Andreeva, A., D. Howorth, J. M. Chandonia, S. E. Brenner*, et al.* (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res **36**: D419-D425.

Aravind, L., V. Anantharaman, S. Balaji, M. M. Babu*, et al.* (2005). "The many faces of the helix-turn-helix domain: Transcription regulation and beyond." Fems Microbiology Reviews **29**(2): 231-262.

Arendall, W. B., 3rd, W. Tempel, J. S. Richardson, W. Zhou*, et al.* (2005). "A test of enhancing model accuracy in high-throughput crystallography." J Struct Funct Genomics **6**(1): 1-11.

Bailey, S. (1994). "The Ccp4 Suite - Programs for Protein Crystallography." Acta Crystallogr D **50**: 760-763.

Bairoch, A., L. Bougueleret, S. Altairac, V. Amendolia*, et al.* (2007). "The universal protein resource (UniProt)." Nucleic Acids Res **35**: D193-D197.

Bannen, R. M., C. A. Bingman and G. N. Phillips, Jr. (2007). "Effect of low-complexity regions on protein structure determination." J Struct Funct Genomics **8**(4): 217-26.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy*, et al.* (2000). "The Pfam protein families database." Nucleic Acids Res **28**(1): 263-6.

Bateman, A. and R. D. Finn (2007). "SCOOP: a simple method for identification of novel protein superfamily relationships." Bioinformatics **23**(7): 809-814.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland*, et al.* (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Bijvoet, J. M. (1954). "STRUCTURE OF OPTICALLY ACTIVE COMPOUNDS IN THE SOLID STATE." Nature **173**(4411): 888-891.

Blundell, T. L. and L. N. Johnson (1976). Protein crystallography. New York, Academic Press.

Brenner, S. E. (2000). "Target selection for structural genomics." Nat Struct Biol **7 Suppl**: 967-9.

Brenner, S. E. and M. Levitt (2000). "Expectations from structural genomics." Protein Sci **9**(1): 197-200.

Chandonia, J. M. and S. E. Brenner (2006). "The impact of structural genomics: Expectations and outcomes." Science **311**(5759): 347-351.

Chayen, N. E., P. D. S. Stewart, D. L. Maeder and D. M. Blow (1990). "An Automated-System for Microbatch Protein Crystallization and Screening." J Appl Crystallogr **23**: 297-302.

Chesshyre, J. A. and A. R. Hipkiss (1989). "Low-Temperatures Stabilize Interferon-Alpha-2 against Proteolysis in Methylophilus-Methylotrophus and Escherichia-Coli." Applied Microbiology and Biotechnology **31**(2): 158-162.

Chothia, C., J. Gough, C. Vogel and S. A. Teichmann (2003). "Evolution of the protein repertoire." Science **300**(5626): 1701-1703.

Corpet, F., J. Gouzy and D. Kahn (1999). "Recent improvements of the ProDom database of protein domain families." Nucleic Acids Research **27**(1): 263-267.

Coulson, A. F. W. and J. Moult (2002). "A unifold, mesofold, and superfold model of protein fold use." Proteins **46**(1): 61-71.

Cowtan, K. D. (1994). "DM: an automated procedure for phase improvement by density modification." Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography **31**: 34-38.

Davis, I. W., A. Leaver-Fay, V. B. Chen, J. N. Block*, et al.* (2007). "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids." Nucleic Acids Res **35**(Web Server issue): W375-83.

Dayhoff, M. O. (1976). "The origin and evolution of protein superfamilies." Fed Proc **35**(10): 2132-8.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-32.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones*, et al.* (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-51.

Friedel, G. (1913). Comptes Rendus Acad Sci Paris **157**: 1533-1536.

Fu, Z. Q., J. Rose and B. C. Wang (2005). "SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination." Acta Crystallogr D Biol Crystallogr **61**(Pt 7): 951-9.

Fu, Z. Q., J. P. Rose and B. C. Wang (2004). "Monitoring the anomalous scattering signal and noise levels in X-ray diffraction of crystals." Acta Crystallogr D **60**: 499-506.

Georgiou, G. and P. Valax (1996). "Expression of correctly folded proteins in Escherichia coli." Current Opinion in Biotechnology **7**(2): 190-197.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis*, et al.* (1996). "Life with 6000 genes." Science **274**(5287): 546-&.

Gouet, P., E. Courcelle, D. I. Stuart and F. Metoz (1999). "ESPript: analysis of multiple sequence alignments in PostScript." Bioinformatics **15**(4): 305-308.

Gough, J., K. Karplus, R. Hughey and C. Chothia (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." Journal of Molecular Biology **313**(4): 903-919.

Grant, A., D. Lee and C. Orengo (2004). "Progress towards mapping the universe of protein folds." Genome Biol **5**(5): -.

Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang*, et al.* (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." Nucleic Acids Res **29**(1): 41-43.

Hendrickson, W. A., J. R. Horton and D. M. Lemaster (1990). "Selenomethionyl Proteins Produced for Analysis by Multiwavelength Anomalous Diffraction (Mad) - a Vehicle for Direct Determination of 3-Dimensional Structure." Embo Journal **9**(5): 1665-1672.

Hendrickson, W. a. and M. M. Teeter (1981). "Structure of the Hydrophobic Protein Crambin Determined Directly from the Anomalous Scattering of Sulfur." Nature **290**(5802): 107-113.

Holm, L. and C. Sander (1993). "Protein-Structure Comparison by Alignment of Distance Matrices." J Mol Biol **233**(1): 123-138.

Holm, L. and C. Sander (1996). "Mapping the protein universe." Science **273**(5275): 595-602.

Hothorn, M., I. D'Angelo, J. A. Marquez, S. Greiner*, et al.* (2004). "The invertase inhibitor Nt-CIF from tobacco: A highly thermostable four-helix bundle with an unusual N-terminal extension." Journal of Molecular Biology **335**(4): 987-995.

Hubbard, T. J. P., B. Ailey, S. E. Brenner, A. G. Murzin*, et al.* (1999). "SCOP: a structural classification of proteins database." Nucleic Acids Res **27**(1): 254-256.

Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti*, et al.* (2008). "The 20 years of PROSITE." Nucleic Acids Res **36**: D245-D249.

Kirillova, O., M. Chruszcz, I. A. Shumilin, T. Skarina*, et al.* (2007). "An extremely SAD case: structure of a putative redox-enzyme maturation protein from Archaeoglobus fulgidus at 3.4 angstrom resolution." Acta Crystallographica Section D-Biological Crystallography **63**: 348-356.

Kissinger, C. R., D. K. Gehlhaar and D. B. Fogel (1999). "Rapid automated molecular replacement by evolutionary search." Acta Crystallogr D **55**: 484-491.

Koonin, E. V., Y. I. Wolf and G. P. Karev (2002). "The structure of the protein universe and genome evolution." Nature **420**(6912): 218-223.

Krissinel, E. and K. Henrick (2004). "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." Acta Crystallogr D **60**: 2256-2268.

Krissinel, E. and K. Henrick (2007). "Inference of macromolecular assemblies from crystalline state." Journal of Molecular Biology **372**(3): 774-797.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-80.

Larson, E. T., B. Eilers, S. Menon, D. Reiter*, et al.* (2007). "A winged-helix protein from suffiblobus turreted icosahedral virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus." Virology **368**(2): 249-261.

Laskowski, R. A., E. G. Hutchinson, A. D. Michie, A. C. Wallace*, et al.* (1997). "PDBsum: a Web-based database of summaries and analyses of all PDB structures." Trends Biochem Sci **22**(12): 488-90.

Laskowski, R. a., M. W. Macarthur, D. S. Moss and J. M. Thornton (1993). "Procheck - a Program to Check the Stereochemical Quality of Protein Structures." J Appl Crystallogr **26**: 283-291.

Laskowski, R. A., J. D. Watson and J. M. Thornton (2005). "ProFunc: a server for predicting protein function from 3D structure." Nucleic Acids Research **33**: W89-W93.

Leonov, H., J. S. B. Mitchell and I. T. Arkin (2003). "Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions." Proteins **51**(3): 352-359.

Levitt, D. G. (2001). "A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps." Acta Crystallogr D **57**: 1013-1019.

Liu, Z. J., D. W. Lin, W. Tempel, J. L. Praissman, *et al.* (2005a). "Parameter-space screening: a powerful tool for high-throughput crystal structure determination (vol D61, pg 520, 2005)." Acta Crystallogr D **61**: 1311-1311.

Liu, Z. J., A. K. Shah, J. E. Habel, J. D. Ng, *et al.* (2005b). "Salvaging Pyrococcus furiosus protein targets at SECSG." J Struct Funct Genomics **6**(2-3): 121-7.

Liu, Z. J., E. S. Vysotski, C. J. Chen, J. P. Rose, *et al.* (2000). "Structure of the Ca2+-regulated photoprotein obelin at 1.7 A resolution determined directly from its sulfur substructure." Protein Sci **9**(11): 2085-93.

Lombardi, A., C. M. Summa, S. Geremia, L. Randaccio, *et al.* (2000). "Retrostructural analysis of metalloproteins: Application to the design of a minimal model for diiron proteins." Proceedings of the National Academy of Sciences of the United States of America **97**(12): 6298-6305.

Madej, T., J. F. Gibrat and S. H. Bryant (1995). "Threading a Database of Protein Cores." Proteins **23**(3): 356-369.

Mccoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, *et al.* (2007). "Phaser crystallographic software." J Appl Crystallogr **40**: 658-674.

Mittl, P. R. E. and M. G. Grutter (2001). "Structural genomics: opportunities and challenges." Current Opinion in Chemical Biology **5**(4): 402-408.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, *et al.* (2007). "New developments in the InterPro database." Nucleic Acids Res **35**: D224-D228.

Murshudov, G. N., A. A. Vagin and E. J. Dodson (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-55.

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia (1995). "Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures." J Mol Biol **247**(4): 536-540.

Navaza, J. (1994). "Amore - an Automated Package for Molecular Replacement." Acta Crystallographica Section A **50**: 157-163.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones*, et al.* (1997). "CATH - a hierarchic classification of protein domain structures." Structure **5**(8): 1093-1108.

Orengo, C. A. and W. R. Taylor (1996). "SSAP: Sequential structure alignment program for protein structure comparison." Method Enzymol **266**: 617-635.

Otwinowski, Z. and W. Minor (1997). "Processing of X-ray diffraction data collected in oscillation mode." Method Enzymol **276**: 307-326.

Perrakis, A., R. Morris and V. S. Lamzin (1999). "Automated protein model building combined with iterative structure refinement." Nat Struct Biol **6**(5): 458-63.

Pflugrath, J. W. (1999). "The finer things in X-ray diffraction data collection." Acta Crystallogr D Biol Crystallogr **55**(Pt 10): 1718-25.

Potterton, L., S. McNicholas, E. Krissinel, J. Gruber*, et al.* (2004). "Developments in the CCP4 molecular-graphics project." Acta Crystallogr D **60**: 2288-2294.

Price, M. N., K. H. Huang, E. J. Alm and A. P. Arkin (2005). "A novel method for accurate operon predictions in all sequenced prokaryotes." Nucleic Acids Res **33**(3): 880-92.

Rould, M. A. (1997). "Screening for heavy-atom derivatives and obtaining accurate isomorphous differences." Macromolecular Crystallography, Pt A **276**: 461-472.

Sali, A. and T. L. Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints." J Mol Biol **234**(3): 779-815.

Schmid, K. J. and C. F. Aquadro (2001). "The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes." Genetics **159**(2): 589-598.

Schneider, T. R. and G. M. Sheldrick (2002). "Substructure solution with SHELXD." Acta Crystallographica Section D-Biological Crystallography **58**: 1772-1779.

Schuller, D. J. (1996). "MAGICSQUASH: More versatile non-crystallographic averaging with multiple constraints." Acta Crystallogr D **52**: 425-434.

Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting, *et al.* (2000). "SMART: a web-based tool for the study of genetically mobile domains." Nucleic Acids Res **28**(1): 231-234.

Sekar, K., V. Rajakannan, D. Velmurugan, T. Yamane, *et al.* (2004). "A redetermination of the structure of the triple mutant (K53,56,120M) of phospholipase A(2) at 1.6 angstrom resolution using sulfur-SAS at 1.54 angstrom wavelength." Acta Crystallographica Section D-Biological Crystallography **60**: 1586-1590.

Shah, A. K., Z. J. Liu, P. D. Stewart, F. D. Schubot, *et al.* (2005). "On increasing protein-crystallization throughput for X-ray diffraction studies." Acta Crystallogr D Biol Crystallogr **61**(Pt 2): 123-9.

Siew, N. and D. Fischer (2003a). "Analysis of singleton ORFans in fully sequenced microbial genomes." Proteins **53**(2): 241-251.

Siew, N. and D. Fischer (2003b). "Twenty thousand ORFan microbial protein families for the biologist?" Structure **11**(1): 7-9.

Siew, N. and D. Fischer (2004). "Structural biology sheds light on the puzzle genomic ORFans." J Mol Biol **342**(2): 369-373.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-960.

Sonnhammer, E. L. L., S. R. Eddy and R. Durbin (1997). "Pfam: A comprehensive database of protein domain families based on seed alignments." Proteins-Structure Function and Genetics **28**(3): 405-420.

Sonnhammer, E. L. L. and D. Kahn (1994). "Modular Arrangement of Proteins as Inferred from Analysis of Homology." Protein Sci **3**(3): 482-492.

Stevens, R. C. (2000). "Design of high-throughput methods of protein production for structural biology." Structure **8**(9): R177-R185.

Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif **41**(1): 207-34.

Teng, T. Y. (1990). "Mounting of Crystals for Macromolecular Crystallography in a Freestanding Thin-Film." J Appl Crystallogr **23**: 387-391.

Terpe, K. (2003). "Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems." Applied Microbiology and Biotechnology **60**(5): 523-533.

Terwilliger, T. C. (2000). "Maximum-likelihood density modification." Acta Crystallogr D Biol Crystallogr **56**(Pt 8): 965-72.

Terwilliger, T. C. (2003). "Automated main-chain model building by template matching and iterative fragment extension." Acta Crystallogr D **59**: 38-44.

Terwilliger, T. C. and J. Berendzen (1999). "Automated MAD and MIR structure solution." Acta Crystallogr D Biol Crystallogr **55**(Pt 4): 849-61.

Tranier, S., C. Iobbi-Nivol, C. Birck, M. Ilbert*, et al.* (2003). "A novel protein fold and extreme domain swapping in the dimeric TorD chaperone from Shewanella massilia." Structure **11**(2): 165-174.

Uson, I. and G. M. Sheldrick (1999). "Advances in direct methods for protein crystallography." Curr Opin Struct Biol **9**(5): 643-8.

Wang, B. C. (1985). "Resolution of phase ambiguity in macromolecular crystallography." Methods Enzymol **115**: 90-112.

Wang, B. C. (1993-2001). Phasing Macromolecular Structures. Strcuture Ananlysis by X-Ray Crystallography, Lecture Notes for ACA Summer Course in Crystallography, Chapter XXV.

Wang, B. C., M. W. Adams, H. Dailey, L. DeLucas*, et al.* (2005). "Protein production and crystallization at SECSG -- an overview." J Struct Funct Genomics **6**(2-3): 233-43.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant*, et al.* (2008). "Database resources of the national center for biotechnology information." Nucleic Acids Res **36**: D13-D21.

Xu, H., C. Yang, L. Chen, I. A. Kataeva*, et al.* (2005). "Away from the edge II: in-house Se-SAS phasing with chromium radiation." Acta Crystallogr D Biol Crystallogr **61**(Pt 7): 960-6.

Yang, A. S. and B. Honig (2000). "An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance." J Mol Biol **301**(3): 665-678.

Yang, C., J. W. Pflugrath, D. A. Courville, C. N. Stence*, et al.* (2003). "Away from the edge: SAD phasing from the sulfur anomalous signal measured in-house with chromium radiation." Acta Crystallogr D Biol Crystallogr **59**(Pt 11): 1943-57.

Zarembinski, T. I., L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim*, et al.* (1998). "Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics." Proceedings of the National Academy of Sciences of the United States of America **95**(26): 15189-15193.

Zhang, K. Y. J. and P. Main (1990). "The Use of Sayre Equation with Solvent Flattening and Histogram Matching for Phase Extension and Refinement of Protein Structures." Acta Crystallographica Section A **46**: 377-381.