STRUCTURES AND PARTIAL FUNCTIONAL STUDIES OF TWO *PYROCOCCUS FURIOSUS* PROTEINS BEYOND HIGH THROUGHPUT STRUCTURAL GENOMICS EFFORT

by

HUA YANG

(Under the Direction of Bi-Cheng Wang)

ABSTRACT

Structural genomics initiatives aim to determine the three-dimensional structure of all proteins. The Southeast Collaboratory for Structural Genomics (SECSG) focuses on *Pyrococcus. furiosus*, a hyperthermophillics archaeon, as a major model organism using a two-tiered approach to achieve this ultimate goal. The SECSG crystallomics group was formed in tier-2 approach to rescue tier-1 failed target.

Structural studies of PF0863 and PF0864 from *Pyrococcus furiosus* were initiated in the context of the structural genomics effort, but failed in the high-throughput structural determination. By using the non-high throughput salvaging practice, both protein structures were solved by single wavelength anomalous dispersion (SAD) method. The structure of PF0863 represents the first released structure of the CYTH domain superfamily defined from protein family (Pfam) database. The biochemical experimental results show that PF0863 has the activity as the nucleotidase. The structural analysis shows high three dimensional similarities between PF0863 and yeast RNA triphasphatase belonged to the mRNA-triPase family. This structural similarity urges us to reconsider the definitions of these two protein families and their

evolution relationship. The structure of PF0864 is the second available structure of transcriptional regulator belonged to Lrp/AscC family from *P. furiosus*. PF0864 can specifically bind to the DNA fragment upstream of the whole operon containing genes PF0865, PF0864, and PF0863, suggesting its autoregulatory function to its own gene expression and members in the same operon under environmental changes. By using genomic context comparison and structure docking methods, PF0863 and PF0864 show a high possibility to form a complex. However, the experimental methods using the size exclusion chromatography followed by native PAGE gel analysis and co-expression followed by GST-pull down assay did not show detectable interactions between PF0863 and PF0864.

INDEX WORDS: Protein crystallography, *Pyrococcus furiosus*, CYTH, Lrp/AsnC, protein-protein interaction

STRUCTURES AND PARTIAL FUNCTIONAL STUDIES OF TWO *PYROCOCCUS FURIOSUS* PROTEINS BEYOND HIGH THROUGHPUT STRUCTURAL GENOMICS EFFORT

by

HUA YANG

B.S. WuHan University, P.R. China, 2000

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2007

STRUCTURES AND PARTIAL FUNCTIONAL STUDIES OF TWO *PYROCOCCUS FURIOSUS* PROTEINS BEYOND HIGH THROUGHPUT STRUCTURAL GENOMICS EFFORT

by

HUA YANG

Major Professor: Bi-Cheng Wang

Committee: John P. Rose
Michael W.W. Adams

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007

DEDICATION

For my dearest parents

ACKNOWLEDGEMENTS

The past five years have been the greatest adventure in my life. I would like extend my sincere and warm thanks to my major advisor, Professor Dr. Bi-Cheng Wang for his generous support, wisdom, encouragement, guidance and patience during my graduate study.

I would also like to thank Dr. John Rose and Dr. Mike Adams from the bottom of my heart for their serving on my advisory committee. Their guidance has aided me to become a scientist from a student. My heartfelt thanks go to Dr. James Liu for the patient teaching and training on crystallography.

I am everlastingly grateful for all of my friends and co-workers in the Wang Lab and the biochemistry department these past five years. Finally, thank you to my family and loved ones for their unending and unwavering support and love throughout my entire scholastic career.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

Introduction and literature review

The significance of this project can be best understood in the context of the historical and scientific developments of genomics and structural genomics. The high-throughput X-ray crystallography activities at Southeast Collaboratory for Structural Genomics (SECSG) and the SECSG crystallomics group, which suggests a solution for high-throughput-failed targets by exclusively focusing on producing purer proteins and ultimately crystals, also provide important background for this project. Meanwhile, the characteristics and importance of model organism, *Pyrococcus furiosus*, need to be discussed.

1.1 Genomics

In the past two decades, genomics has not only served as the engine to revolutionize biology, but also has become a dominate paradigm, or even the catchword depending on one's prospective, of the now flourishing life science studies. The term "genomics" derived from the word "genome". Genome was first used by H. Winkler in 1920, and created by combining the words "**gen**es" and "chromos**ome**s". It refers to the complete set of genes and chromosomes of an organism. Genomics was first proposed by Thomas H. Roderick in 1986 to describe the scientific discipline of mapping, sequencing, and analyzing genomes. On September 1, 1987, the inauguration of a new journal *Genomics* quickly established the first institutional support of this new concept, with the first editorial titled "Genomics: A New Discipline, a New Name, a New Journal" (McKusick 1997).

Genomics was a relatively new field of biological investigation in the 1980s, but it soon took off in the 1990s. It began with the proposal of the Human Genome Project (HGP) after intensive discussion, debate, and planning from 1986 to 1990 among the academia and governmental agencies. The mapping and sequencing of the human genome as suggested by the HGP was officially initiated in 1990 in the United States by the Department of Energy (DOE) and National Institutes of Health (NIH). This largely U.S. sponsored international consortium also drew geneticists from China, France, Germany, Japan, and the United Kingdom. The whole HGP adventure was expected to be accomplished in fifteen years with an expenditure of $200 million per year. Parallel to the HGP, the genomic sequencing projects for model organisms such as yeast, mouse, and DOE-formulated Microbial Genome Initiative were also pursued. The Institute for Genomic Research (TIGR) published the whole genome sequence (~1.8Mb) for *Haemophilus influenzae* Rd, a free-living bacterium, in July 1995 (Fleischmann, Adams et al. 1995). This represented the first genome sequenced successfully using a shotgun sequencing approach, which is a fast and effective method for obtaining genomics sequences. From the prospective of academic development, this result was the first milestone toward the era of genomics (Venter, Smith et al. 1999). Since then, more than 400 eukaryotic and prokaryotic genomes have been completely sequenced, and hundreds of sequenced genomes will join them in the very near future (http://www.tigr.org). The complete sequencing of budding yeast *Saccharomyces cerevisiae*, which represented the first sequenced eukaryotic organism, was finished in 1996 (Goffeau, Barrell et al. 1996). The first multicellular organism, the worm *Caenorhabditis elegans*, was completely sequenced by *C. elegans* sequencing consortium in 1998 (C. elegans Sequencing Consortium 1998). In the year 2001, two versions of the draft of the human genome were published (Sachidanandam, Weissman et al. 2001; Venter, Adams et al.

2001). This constituted the cornerstone of genome-based biology and provided the richest intellectual resource in the history of biology. The breakthrough in agricultural research and the first economically important cereal crop, rice (*Oryza sativa*), on the other hand, was completely sequenced in 2002 (Goff, Ricke et al. 2002; Yu, Hu et al. 2002). The genome sequencing of model organism used in this study, *Pyrococcus furiosus*, was finished by the scientists from University of Maryland Biotechnology Institute in 2001 (Robb, Maeder et al. 2001). The genome information and characteristics of *P. furiosus*, the hyperthermophilic archaea—in direct context of this dissertation— will be discussed in the later part of this chapter.

The universal acceptance and wide use of the term "genomics" in the scientific community, however, failed to solve the definition problem of the very term once for all. The focus of genome analysis shifted from exclusively mapping and sequencing to including gene functions in 1995 (Hieter and Boguski 1997) suggesting that genomics be divided into structural and functional studies (McKusick 1997). Since then, various terms such as functional genomics, proteomics, and structural genomics have continued to circulate in the scientific community (Burley, Almo et al. 1999; Fields, Kohara et al. 1999; Martin and Nelson 2001). Also, due to the application of genome sequence information and genomic technologies, new diverse biological sub-disciplines as toxicology, pharmacology, medicine, physiology, and ecology have come into being and gradually established their footholds in academic institutions.

To take full account of these fascinating new developments and to construct a more comprehensive and indicative, albeit tentative, understanding of genomics, the term genomics is used here primarily to reach a genome-level comprehension based on the whole-genome sequence information and high-throughput genomic technologies regarding the molecular basis of the structure, function, and evolution of biological systems. Under this circumstance,

therefore, this dissertation will use the classification based on system attributes, and concentrate on structural genomics. (For a full review of the classification of genomics, see (Zhou 2004)).

1.2 Structural Genomics

The general system theory defines structure and function as two primary characteristics of any system. As mentioned above, structural genomics is the primary focus of this study, and it is hopefully through the lens of structure that the function is perceived. From the viewpoint of system attributes, structural genomics was initially and broadly defined as the genome-wide structural study of genes, proteins, and other biomolecules, including genome mapping, sequencing, and organization as well as protein structure characterization in biological systems.

In practice, the broadly defined term of structural genomics has shrunk considerably, and it has been used only with regard to part of the initial research agenda. In various ways, it has come to refer to the genome mapping, sequencing, and organization (Hieter and Boguski 1997), and genome-wide protein structural characterization and prediction (Kim 1998). After the complete sequence information of the entire human genome as well as of various model organisms and microbial pathogens became available, structural genomics was further narrowed down to describe the process of determining the three-dimensional structures of all proteins. Now as a widely accepted definition, this newly fashioned term justifies some further deliberation.

Since structural genomics has evolved into a multidisciplinary research project and an international enterprise to focus on making the three-dimensional atomic-level structure of most proteins easily obtainable from knowledge of their corresponding DNA sequences. Developing methods and technologies to speed up recombinant expression, purification, structure

determination, and model building has become the long-range goal. Under this situation, the high-throughput approach has become a dominant research idea.

The impact of high-throughput pipelines in American structural genomics is obvious. The Protein Structure Initiative (PSI-1) was a joint effort of federal government, universities, and industries aimed at dramatically reducing the costs and shortening the time required to determine a three-dimensional protein structure. In the year 2000, NIH established nine pilot structural genomics centers under the umbrella of PSI-1, see Table 1.1.

Although experimental structural genomics uses the same principles as those used on traditional structural biology, the two fields differ in terms of motivation, automation, and scale (Brenner 2001). Not only are proteins with well-characterized functions candidates for structure determination, but all the proteins from the whole genome become possible targets to work with in the structural genomics era. The logic of understanding function from structure has been turned on its head by using the structure to infer functions. Moreover, each stage of the experimental structural genomics process, including protein production, crystallization, data collection, phasing, model building, structure refinement and validation, needs to be refined and optimized, since any stage has the potential of becoming a bottleneck for the whole process. The processing of selected protein targets through the experimental structural genomics necessitates eliminating those targets in large quantity, with many of them starting at the same time and a portion of them failing at each stage of the process. Each pilot center chose protein targets using its own distinct criteria. Once the targets were locked, however, each center assigned priority to certain targets according to the same standards such as phylogenetic distribution (Fischer 1999), family size (Vitkup, Melamud et al. 2001), the likelihood of producing a new fold (Linial and Yona 2000), and functional relevance (Erlandsen, Abola et al. 2000).

After the five-year pilot phase (Table 1.2), more than 1,100 protein structures were determined, and more than 700 of them were unique structures, i.e., they shared less than 30 percent of their sequence with other known protein structures in Protein Data Bank (PDB) (Berman, Westbrook et al. 2000). The protein structures having new folds constituted about 16% of the structures solved from the PSI-1 pilot centers (Chandonia and Brenner 2006), which did not meet the expectation of 40% of the protein structures having a new fold (Brenner and Levitt 2000). Yet on the other hand, quantity can not supplant quality since every new fold identified from a protein family gives us new insights. Moreover, the impact of structural genomics on technology devolvement is remarkable (http://www.nigms.nih.gov/Initiatives/PSI/Background/PilotFacts.htm). These technical advances include: the "Sesame" laboratory information management system, auto-induction protocols, expression systems based on fusions, incorporation of a wheat germ cell-free expression system, fully integrated robotic crystallization systems, automated storage and crystal imaging units, small-volume crystallization chips, automated software for X-ray structure determination, automatic crystal mounting and crystal screening robots, automated nuclear magnetic resonance (NMR) data analysis, and automated post-structure functional analysis software. Those remarkable achievements have affected all areas related to structural genomics, including bioinformatics, molecular biology, biochemistry, NMR spectroscopy, and X-ray crystallography.

The follow up five-year production phase of structural genomics (PSI-2) beginning in July 2005 wasted no time in taking advantage of the pilot phase achievements. Centers in two categories were established. Four large-scale centers, which were established during the PSI-1 pilot phase, have constructed structural genomics pipelines for the production and structural determination of proteins in a high-throughput operation. Six new specialized centers were

expected to develop innovative methods, approaches, and technologies for producing and determining the structures of proteins that have been traditionally regarded as difficult (http://www.nigms.nih.gov /Initiatives/PSI/Centers/).

1.3 SECSG

The Southeast Collaboratory for Structural Genomics (Adams, Dailey et al. 2003), one of the original nine structural genomics PSI-1 pilot centers, is a consortium consisting of five partner institutions in the southeast: the University of Georgia, the University of Alabama at Birmingham, the University of Alabama at Huntsville, Georgia State University, and Duke University Medical Center. As an NIH-PSI-1 pilot center, SECSG focused on developing high throughput and cost effective pipelines for protein production, crystallization, and structure determination by X-ray crystallography and NMR (Wang, Adams et al. 2005).

SECSG focuses on determining structure from protein families in the Pfam database (Bateman, Coin et al. 2004) who lack a three-dimension model since they will not only enlarge the structural database, but also shed new lights on function and mechanism obtained from the structure. SECSG targets are mainly selected from *P. furiosus* and *C. elegans*, which serve as model organisms for exploring methods and technologies related to high throughput production of prokaryotic and eukaryotic proteins and structures. Potential targets also include selected human proteins from Mammalian Gene Collection (MGC) (Strausberg, Feingold et al. 1999).

The SECSG high throughput gene-to-structure pipeline in this center includes the following activities: (1) target selection; (2) cloning the coding sequence of targets into an appropriate expression vector; (3) sequence verification of the cloned gene; (4) protein expression and solubility screening; (5) large scale protein production and purification; (6) high

throughput crystallization and optimization, (6) data collection, (7) structure determination; (8) structure validation and (9) deposition into PDB.

The impact of high throughput approach on protein crystallography, the central tool used for structure determination by SECSG, has been tremendous. The advanced automated techniques have dramatically shortened the time needed to obtain three-dimensional structures. However, the high throughput crystallography has its own disadvantages, which are primarily determined by the inherent bottlenecks associated with protein crystallography. Thus, it is necessary to examine protein crystallographic context next.

1.3.1 Protein Crystallography

Crystallography, in older usage, is the scientific study of crystals based on their geometry. It is the experimental science of determining the arrangement of atoms in solids. Crystallographic methods now depend on the analysis of the diffraction patterns that emerge from a sample that is targeted either by an X-ray, neutron or electron. Protein crystallography is a technique that uses X-ray diffraction through the closely spaced lattice of atoms in a crystal to reveal the nature of that lattice. Protein crystallography should be more accurately named macromolecular crystallography because the targets include proteins, nucleic acids, and other molecules with the molecular weight higher than 35KDa (Dauter 2006).

Protein crystallography provides the most direct way of visualizing images of molecules. The well-ordered parts of high-resolution structures are among the most solid evidence obtainable in sciences. In addition, the three dimensional structures of the proteins offer detailed information on their activities, mechanisms, and possible conformational changes. The evolutionary relationships between molecules from widely separated systems can be shown graphically by their three dimensional structures, which can provide a wide view of the

similarities between different proteins where their amino acid sequences has no significant homology (Johnson, Sutcliffe et al. 1990).

The process of protein crystallography consists of the following steps: (1) crystallization, (2) data collection, (3) phasing, (4) model building, (5) refinement, (6) validation, and (7) presentation of the result. However numbers of potential bottlenecks exist in the process, of these growing a well diffracted crystal suitable for structure analysis and solving the phase problem are the two most serious bottlenecks: without crystals, no structure can be determined by crystallography; and phase information can not be measured directly from the diffraction data. Because these two bottlenecks are the keys to understanding the limitation of high throughput crystallography, the detail of the methods and techniques used for crystallization and phase solution will be introduced.

Crystallization, the first essential step in determining the X-ray structure of a protein, is a multi-parameter problem. The purity of protein sample is one of the most important factors in obtaining the diffraction quality crystals (better than 3Å resolution). Usually the protein sample should be pure >99%, fresh, and homogenous (monodispersed) (Ferre-D'Amare and Burley 1994) in solution. Protein crystallization occurs only when the protein achieves supersaturation, which means the concentration of protein in solution is greater than its limit of solubility (Figure 1.1). The crystallization process consists of two major events, nucleation and crystal growth. In the supersaturated state, small aggregates are formed, which are the nuclei for crystal growth. For crystal growth, supersaturation must be reduced to avoid production of too many nuclei which will lend to the formation of many small crystals. A point is then usually reached when the protein molecules in solution and crystal are in equilibrium at which point crystal growth stops (McPherson 1999). Commonly used crystallization methods include vapor diffusion, batch, and

dialysis. Batch crystallization is the oldest and simplest method for protein crystallization. The principle is that the precipitating reagent is instantaneously added to a protein solution, suddenly bringing the solution to a state of high supersaturation (Drenth 1999). Using this method, the Nobel Prize in Chemistry in 1946 honored Sumner "for his discovery that enzymes can be crystallized", and Northrop and Stanley "for their preparation of enzymes and virus proteins in a pure form" (Nobelprize.org). The most popular crystallization method today is the vapor diffusion method using either hanging drop (Davis and Segal 1971) or other sitting drop variations. In the vapor diffusion experiment, small volumes of precipitant and protein are mixed together and the drop equilibrated against a larger reservoir of solution containing precipitant or other dehydrating agents. In the dialysis crystallization experiment, protein is equilibrated against a larger volume of precipitant solution through a dialysis membrane either in solution (Zeppezauer, Eklund et al. 1968), in a gel (García-Ruiz and Moreno 1994), or in a capillary (Ng, Gavira et al. 2003). Before the birth of the structural genomics, setting up crystallization trials by hand was a laborious and time-consuming process. Automation of crystallization was the first step in the high-throughput X-ray structure analysis process made by the use of robotic devices (Weselak, Patch et al. 2003). The modified microbatch under oil method (Chayen, Shaw-Steward et al. 1992) is very amenable to automatic crystallization. The experiment is carried out by introducing a drop of protein solution and a drop of precipitant solution into a vessel containing water-immiscible oil (Figure 1.2). Because their density are higher than that of oil, both the protein and solution droplets move to the bottom of the well and mix. The oil prevents water evaporation and reduces the amount of oxygen that can reach the droplet. In this method, only small volume protein sample is needed to obtain the crystals. By using water-permeable oil slow evaporation of the droplets can be achieved in a manner analogs

to vapor diffusion (D'Arcy, Elmore et al. 1996). The commonly used oil mixtures are paraffin and silicon in a 70:30 ratio.

The result of crystallization is hard to predict. As a multi-parameter problem, changing protein concentration, varying nature and concentration of precipitant conditions adjusting to different pH, and using different crystallization temperature, may all have the effect on producing diffraction quality crystals (Drenth 1999). Sometimes even serendipity plays a role, too. Therefore, while high throughput crystallization does speed up the set up of crystallization trials, it can not be regarded as an omnipotent method.

The phase problem has long been considered as protein crystallography's Achilles' heel. Phase information can not be obtained from a single diffraction experiment for macromolecules. Therefore, the calculation of the electron density map, the goal of a crystallographic experiment, using the Fourier transform equation will be impossible without the phase information. The electron density Fourier transform equation is used by all protein crystallographers in structure solution (Figure 1.3). This equation requires four components: volume of a unit cell, an atomic position value, an intensity measurement, and a phase difference. In the diffraction experiment, we measure the intensities of waves scattered from planes in the crystal. The amplitude of the wave $|F_{hkl}|$, also known as structure factor, is proportional to the square root of the intensity measured on the detector. To calculate the electron density at a position (xyz) in the unit cell of a crystal requires us to perform the Fourier summation over all the *hkl* planes. The amplitudes can be measured, but all phase information is lost in the diffraction experiment. Therefore, phase problem becomes an inherent bottleneck in crystallography. Since the crystallographer must discover a way to find the true phase from the infinite number of solutions. The general principle of phasing is the requirement of presences of heavy or anomalously scattering atoms

(large number of electrons) to provide initial phase estimation. Several methods exist to accomplish this goal (Table 1.3).

In the past, multiple isomorphous replacement (MIR) (Crick and Magdoff 1956; Harker 1956) was the most popular method for phase determination for macromolecular aspects when intensity estimations were in general not very accurate and tunable X-ray sources were not available. With the accurate measurement of intensity by area detectors and improved data collection techniques coupled with tunable synchrotron X-rays, the structure solution method shifted its emphasis to multiple wavelength anomalous dispersion (MAD) (Hendrickson, Smith et al. 1985; Kahn, Fourme et al. 1985). The introduction of selenomethionine into proteins (Hendrickson, Horton et al. 1990) resulted in the MAD phasing method being the dominated method of structure determination until recently.

Single wavelength phasing of macromolecules was pioneered by Hendrickson & Teeter (Hendrickson and Teeter 1981) and Wang (Wang 1985). With today's progress in data collection techniques and the current trend toward high throughput structure determination, the simpler single wavelength anomalous dispersion (SAD) method is becoming more and more popular (Blow 2003) (Figure 1.4). For better understanding of this method, we need to trace back and recall what anomalous intensity differences are. In 1913, Friedel found that the diffraction spot intensity at a point (h,k,l) should be identical to the intensity at point (-h,-k,-l) which became known as Friedel's law. Pairs of diffraction spots obey Friedel's law is illustrated in Figure 1.5. The structure factor is represented as a vector on an Argand diagram with a horizontal "real" axis and a vertical "imaginary" axis. In 1949, Bivjoet identified what he called "abnormal scattering," scattering that breaks the Friedel's law intensity correlation, of an iodide ion to distinguish between the real and mirrored version of cholesteryl iodide and later suggested

that this difference could be used with isomorphous replacement. A few years later, Bivjoet differences were observed in the diffraction pattern from the iron atom associated with myoglobin (Ingram and Kendrew 1956). At that time, this deviation from Friedel's law was not expected and referred to as "anomalous" scattering. Anomalous scattering is a misnomer for it is a naturally occurring resonance phenomenon observed when the wavelength of the X-rays approaches the absorbance edge of an atom. The equation in Figure 1.6a shows the summation of scattering components that accounts for the total intensity, $F_{anomalous}$, of a single type of atom where F_{normal} is the normal Thompson scattering, $\Delta f'$ is the added real anomalous scattering component that is always in the plane of the normal scattering (with a phase of either 0 or 180°), and $\Delta f''$ is the added imaginary component that is always 90° ahead of the real component, and it is graphically illustrated in Figure 1.6b. The latter two terms represent the anomalous scattering that occurs at the absorption edge when the X-ray photon energy is sufficient to promote an electron from an inner shell. The disruption of Friedel's law is caused by the phase shift of the imaginary component when dealing with two types of atoms; F_+ no longer has the same intensity, vector length, phase, or relative angle to the real axis, as F_- (Figure 1.6c). By introducing one heavy atom, higher Z-number than carbon, nitrogen, oxygen, and sulfur into the unit cell, one data set collected at the absorption edge of that atom can provide enough information to use SAD method to get the structure solution. An illustration of the SAD method is graphically shown in Figure 1.7. The experimental procedure of producing heavy atom derivative crystals will be described in Chapter 3.

Molecular replacement (MR), another common phasing method was first described by Michael Rossmann and David Blow in 1962 (Rossmann and Blow 1962; Rossmann 1972). This method is currently responsible for half of the structures deposited in the PDB. To achieve

success of this method, the sequences between the model structure and the unknown structure have to be at least 35% identical. The traditional way of performing this method uses a known model having a similar structure predicted by sequence identity and involves a three-dimensional rotation search using the model structure to get correct orientation of the unknown structure followed by a three-dimensional translation search to get the correct placement of the oriented model in the unknown unit cell.

The solutions for the phase problem have not changed substantially for almost twenty years with its focus primarily on using the previously introduced methods under the general principles of isomorphous replacement and anomalous diffraction. Many automated software package for X-ray crystallographic structure determination are now available as account of the structural genomics effort and the enormous progress made in computing technology. The high throughput structure determination thus can be shortened to weeks, days, even hours for a well diffracting crystal contains a heavy scattering atom. However, the introduction of heavy atoms into the native protein crystal is a risky process since it may cause diminish the crystals' diffraction quality. On the other hand, selenomethionine (SeMet) labeled protein can almost guarantee heavy atom incorporation, but the technique will only work when the protein sequence contains one or more methionine. Also in some cases not all methionines are replaced by SeMet. Moreover, the protein characteristics may change after being labeled with SeMet, such as solubility and crystallization conditions. Therefore, the phase problem is still a bottleneck on the way to structure determination.

Today, due to the high throughput structural genomics efforts and the maturity of protein crystallography methods, there is a tendency of solving protein structures automatically with minimum human intervention (Fu, Rose et al. 2005; Leonard, Sainz et al. 2005; Liu, Lin et al.

2005; Panjikar, Parthasarathy et al. 2005). However, it is obvious that not all protein structures can be obtained by fully automatic approaches. Some unusually difficult structures or crystals displaying various kinds of atypical behavior will always exist (Dauter, Botos et al. 2005), thus traditional crystallography with non-high throughput effort will still be of value in the future.

1.3.2 Crystallomics

The high throughput protein-to-structure pipeline developed by SECSG has been introduced above; it represents SECSG's tier-1 approach. In 2003, the SECSG crystallomics group was formed to support its tier-1 efforts by providing non-high throughput second tier activities to rescue failed tier-1 targets. The crystallomics group provided scaled-up amounts of tier-1 proteins-repeats for further crystallization trials, where necessary, and supplied labeled/modified proteins for crystal optimization and structure determination purposes (Wang, Adams et al. 2005). The crystallomics group uses two important biochemical principles, protein purity and surface entropy modification, to rescue tier-1 failed targets. The final products of crystallomics are diffraction quality crystals, and getting pure proteins is only the intermediate step. The effort that deemphasizes the high throughput pipeline approach adopted by crystallomics group includes the alternative protein purification, affinity tag removal, reductive methylation, surface mutagenesis, SeMet labeling, and optimization in crystallization. For example, by reexamining 50 targets that failed in tier-1 processing (either producing no crystals or crystals of poor diffraction quality), 9 structures were solved and deposited into PDB (Liu, Shah et al. 2005). The salvaging effort, a part of crystallomics, can be regarded as the supplement for high throughput effort.

1.3.3. New Refinement Procedures to Ensure High Quality Structural Results

Another related issue is structure refinement and validation. The quantity of structures is one important standard of measuring the achievement of high-throughput structural genomics effort, yet quality should never be sacrificed for a single-minded pursuit of quantity. SECSG always aims at providing structural models of the highest quality. Combined procedures of newly developed structure-validation tools with refinement programs are used. The current approach uses (1) updated versions of the standard Ramachandran side-chain rotamer database and bond-angle criteria (Lovell, Davis et al. 2003), (2) crystallographic R, R_{free} (Brunger 1992), and difference map peaks, (3) hydrogen-bonding and analysis of side-chain and imidazole orientation (Word, Lovell et al. 1999), and (4) H-atom addition and all-atom steric clashes (Lovell, Davis et al. 2003). All recent submissions from the University of Georgia's SECSG crystallography core have undergone the automatic correction of Asn/Gln/His flips available in MOLPROBITY site, and MOLPROBITY's rotamer, Ramachandran, and clash information have been incorporated early on in the refinement process (Arendall, Tempel et al. 2005).

The historical and scientific context reviewed above indicates the importance of not depending on the high-throughput methods exclusively. Now we should briefly review the background of *P. furiosus*, the model organism used in this dissertation work.

1.3.4. *Pyrococcus furiosus*

A phylogenetic tree (Figure 1.8), based on 16S-rRNA analyses, shows the three domains of life as bacteria, archaea, and eukaryotes (Woese, Kandler et al. 1990). Many archaea are extremophiles. They are organisms, usually unicellular ones, living in or requiring an "extreme" environment (Rothschild and Mancinelli 2001). Many different extremophiles classes exist, each corresponding to the way in which its environmental niche differs from those of the

majority of terrestrial mesophile organisms. Hyperthermophiles are a subset of extremophiles, including organisms that have optimum growth temperatures of at least 80 ºC with maximum growth temperatures of 90 ºC and above (Blochl, Rachel et al. 1997). In the past decades, many different types of hyperthermophile have been isolated in the geological hot spots around the world, with all of them being prokaryotes and most belonging to the domain of archaea (Stetter 1996).

Pyrococcus furiosus, a hyperthermophile species of archaeon, belongs to Euryarchaeota phylum (Figure 1.7). It was originally isolated anaerobically at the beach of Porto Levante, Vulcano Island, Italy within the heated marine sediments with temperature between 90 ºC and 100 ºC (Fiala and Stetter 1986). The genus name *Pyrococcus* means "fireball" which refers to the round shape of the extremophile and its ability to survive in temperatures of around 100 ºC. The species name *furiosus* means "rushing" and refers to the extremophile's motility. The appearance of *P. furiosus* is mostly regular motile cocci of 0.8 µm to 1.5 µm diameters with 50 flagella at one end and often found in pairs. The optimum growth temperature is 100 ºC with the variation between 70 ºC and 103 ºC. The organism grows between pH 5 and 9 with an optimum at pH 7 (Fiala and Stetter 1986).

P. furiosus is notable because its chromosomal integrity can be maintained at temperatures up to 103 ºC with very little accumulation of DNA breaks. Also, it can withstand radiation doses up to 1.5 k Gray (Gy) (Robb, Maeder et al. 2001). Temperature is a critical factor for biomolecules. Generally as the temperature approaching 100 ºC, proteins and nucleic acids will be denatured, and the fluidity of membranes will reach the lethal state. In addition, radiation can seriously damage the nucleic acids. Thus, finding out the ways that proteins can cope with high temperature and their mechanisms to avoid and/or repair radiation damage will

provide valuable data on molecular biology, especially protein folding. Protein structures from *P. furiosus* offer information in these two aspects. From the viewpoint of evolutionary biology, the discovery of extremophiles increased the phylogenetic clarification, and deepened our understanding of chances versus necessities on the molecular level in the evolutionary pathways (Rothschild and Mancinelli 2001).

Because of the significant characteristics of *P. furiosus* and the rich information we can obtain from its study, SECSG selected *P. furiosus* as the prokaryotic model organism to pursue high throughput structural genomics studies. The sequencing of the complete genome of *P. furiosus* was completed in 2001 by scientists at the Biotechnology Institute of the University of Maryland (Robb, Maeder et al. 2001). The whole genome of *P. furiosus* has 1,908 kilobases that code for 2,065 proteins labeled PF0001-PF2065. About 700 of the open reading frames (ORFs) are predicted to be organized in operons, suggesting that they encode either multisubunit complexes or include accessory proteins for assembly of the active enzyme. A complete expression library of *P. furiosus*, representing all proteins from the simplest cytoplasmic protein to the most complex membrane protein assembly, is the ultimate goal of the SECSG protein production core, which will lead us to the final mission of determination all the three dimensional structures for each protein (http://www.secsg.org).

1.4 Significance of This Work

The international high throughput structural genomics efforts have profoundly influenced the scientific world. As a dominate approach of protein crystallography, however, the high-throughput operation should not be regarded as a panacea capable of overcoming the deficiencies of traditional crystallography. For example, structural studies of gene PF0863 (open reading frame *Pfu*-838710) and gene PF0864 (open reading frame *Pfu*-839272) from *Pyrococcus*

furiosus were both initiated in the context of the structural genomics effort. Yet both failed to be solved by the high-throughput structural determination. Therefore, by using non high-throughput effort to solve the structure of PF0863 and PF0864, we can redress certain shortcomings of high-throughput effort.

More importantly, the discovery of the crystal structure of PF0863 is significant. It is the first model of CYTH domain superfamily in Pfam (Bateman, Coin et al. 2004), which is defined by the catalytic domains of CyaB-like adenylyl cyclase and thiamine triphosphatase. This structure fills another gap in our knowledge of CYTH domain superfamily and sheds new lights on understanding the functions of this Pfam through the lens of structure. The non-high-throughput protein purification approaches produce diffraction quality crystals, which lead to the final structure determination of PF0863. A PF0863 dimer occupies the crystallographic asymmetric unit. Each monomer of the dimer contains an 8-stranded anti-parallel β barrel that forms a topologically closed tunnel. The structural analysis shows that PF0863 has the same fold as yeast RNA triphosphatase Cet1p belonged to RNA triphosphatase family in Pfam. The structural similarity urges us to reconsider the definitions of these two protein families and their evolution relationship. The biochemical experimental results show PF0863 has the activity as the nucleotidase.

PF0864 sequence displays weak homology to transcriptional regulation family: Lrp/AsnC regulators. The structure determination went through traditionally tedious heavy atom soaking process to solve the phase problem. Molecular replacement, the favorite high throughput structure solving method, failed because of relatively low sequence identity (<35%) with the available model at that time. The crystal structure of PF0864 shows N-terminal HtH binding motif involving in the DNA-binding and C-terminal RAM domain possibly involving in the

effector binding. It shows high structural similarity with another crystal structure from *Pyrococcus furiosus* in the Lrp/AsnC family. The DNA-binding essay shows PF0864 can specifically bind to the DNA fragment upstream of the whole operon containing genes PF0865, PF0864, and PF0863, suggesting its autoregulation function to its own gene expression and members in the same operon under environmental changes. Because of the variety in these family members in different species, the exact binding sequence of DNA and how it functions as a transcriptional regulator *in vivo* need more study.

The protein-protein interaction study of PF0863 and PF0864 began after the individual structures of those two proteins were finished. A prediction of this project was that by using genomic context comparison and structure docking methods, there might be a high possibility that PF0863 and PF0864 can form a complex. However, the experimental methods using the size exclusion chromatography followed by native PAGE gel analysis and co-expression followed by GST-pull down assay did not show detectable interactions of PF0863 and PF0864.

Structures are well conserved during evolution (Bajaj and Blundell 1984). The three-dimensional structural studies of PF0863 and PF0864 not only reveal their biological functions and provide the opportunity to find homology with undetectable sequence similarity, but also open the door to studies on their possible physical and functional interactions.

Table 1.1: Nine structural genomics PSI-1 pilot centers

Center	Key ideas	Website
Midwest Center for Structural Genomics (MCSG)	Novel protein folds and technology development	www.mcsg.anl.gov
Northeast Structural Genomics Consortium (NESGC)	Complementarities of NMR and crystallography; coverage of structure space	www.nesg.org
New York Structural Genomics Research Consortium (NYSGRC)	Yeast proteins with novel folds; technology development	www.nysgrc.org
The Joint Center for Structural Genomics (JCSG)	Large-scale automation; proteins from *Thermotoga maritime* and *Caenorhabditis elegans*	www.jcsg.org
Berkeley Structural Genomics Center (BSGC)	Complete structural genomics of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*	www.strgen.org
Southeast Collaboratory for Structural Genomics (SECSG)	Development of SAD technology; *Pyrococcus furiosus*, *Homo sapiens* and *Caenohabditis elegans* proteins	www.secsg.org
Mycobacterium tuberculosis Structural Genomics Consortium (TB)	*Mycobacterium tuberculosis* proteins; new folds; large scale collaboration	www.die-mbi.ucla.edu/TB/
Center for Eukaryotic Structural Genomics (CESG)	Novel eukaryotic proteins, with *Arabidopsis thalianaas* a model genome	www.uwstructuralgenomics.org
Structural Genomics for Pathogenic Protozoa (SGPP)	Structural genomics of protozoan pathogens	www.sgpp.org

Table 1.2: Protein Structure Initiative (NIH-affiliated centers) - Phase 1 (October 1 2000—August 31, 2005)

Center	All targets	Cloned	Crystals	Diffracted	Structures(novel)		In PDB (novel; unique)	Deposits after Oct 1, 2000 (novel; unique)	% unique
					NMR	X-Ray			
MCSG	15565	5730	888	363	0	296 (274)	296 (274; 235)	291 (269; 230)	79.0
NESGC	12213	5484	163	116	93	116 (97)	198 (169; 138)	186 (157; 128)	68.8
NYSGRC	2145	1538	397	196	0	195 (157)	178 (146; 106)	171 (139; 100)	58.5
JCSG	6594	3650	1167	268	8	221 (180)	198 (160; 104)	198 (160; 104)	52.5
BSGC	911	812	94	65	3	58 (50)	52 (45; 37)	43 (37; 30)	69.8
SECSG	14786	14378	223	118	2	74 (52)	71 (51; 29)	71 (51; 29)	40.8
TB	1758	1547	209	120	2	107 (70)	67 (44; 25)	62 (40; 23)	37.1
CESG	6582	4476	104	40	18	34 (22)	47 (33; 27)	47 (33; 27)	57.4
SGPP	19503	10154	175	45	0	28 (17)	22 (15; 10)	22 (15; 10)	45.5
Total PSI	75104	45391	3311	1307	125	1114 (919)	1111 (937; 711)	1074 (901; 681)	63.4

The data were generated based on XML files released by the centers as of August 31, 2005.
(Adopted from http://olenka.med.virginia.edu/mcsg/html/results_psi_1/index.html)

Only distinct target sequences are taken into account for each center and in the total count (hence numbers of "distinct" targets reported above for centers where sequences are duplicated or missing in XML files may be lower than those reported by the centers; note also that the number of targets in the total count may be less then the sum of targets for the centers due to target overlaps). Columns “Structures” and “In PDB” show number of distinct target sequences marked with this status in XML files. In cases where individual domains are deposited column “In PDB” shows both the total number deposits and the number of distinct sequences. The number of deposits after the start of PSI (October 1, 2000) is calculated by excluding the targets with earlier deposition date (thus it includes solved targets marked as 'In PDB' in XML files but still awaiting actual deposition and not showing up in the PDB). “Novel structures” are those for which there are no matches with more than 50% relative identity to earlier deposits in PDB. “Unique” are those with no matches above 30% relative identity to earlier PDB structures. The number of novel and unique structures cannot be accurately determined for centers which do not provide information about PDB ids for their deposits in XML files. Cases where 10% or more of solved structures for a center have no PDB id will be indicated by a question mark. “Uniqueness” and “novelty” of a SG deposit may change after its deposition, due to a subsequent release of an earlier PDB deposit.

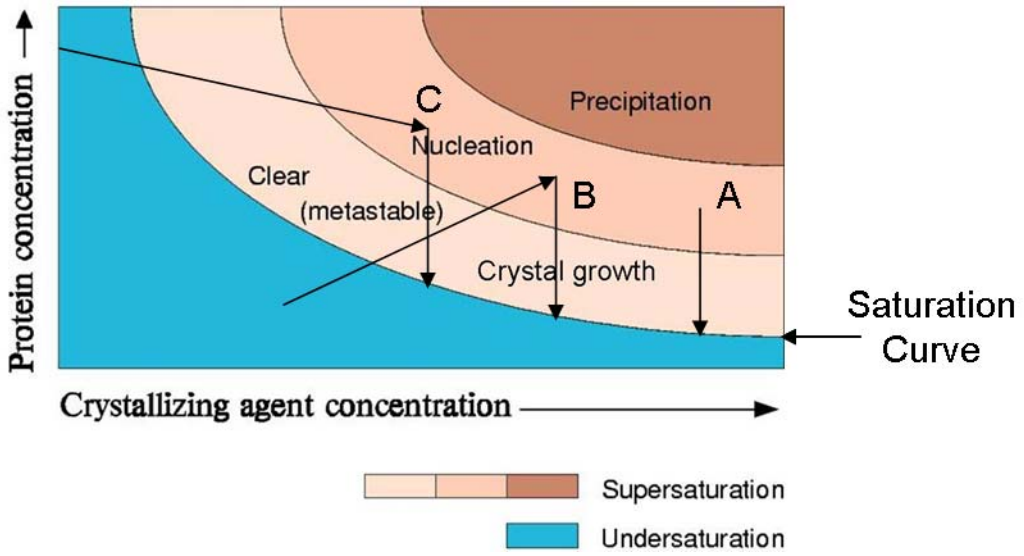


Figure 1.1: The phase diagram showing zones for crystal nucleation, growth and precipitation.

The solubility curve separates the undersaturation and supersaturation states of a protein. Nucleation zone is where the crystallization starts. The metastable zone is where the crystal growth after nucleus formation. A, B, C represent crystal formation process for batch method, vapor diffusion method and dialysis method, respectively.

Modified from http://www-structmed.cimr.cam.ac.uk/Course/Crystals/Theory/phase_diagzones.html.

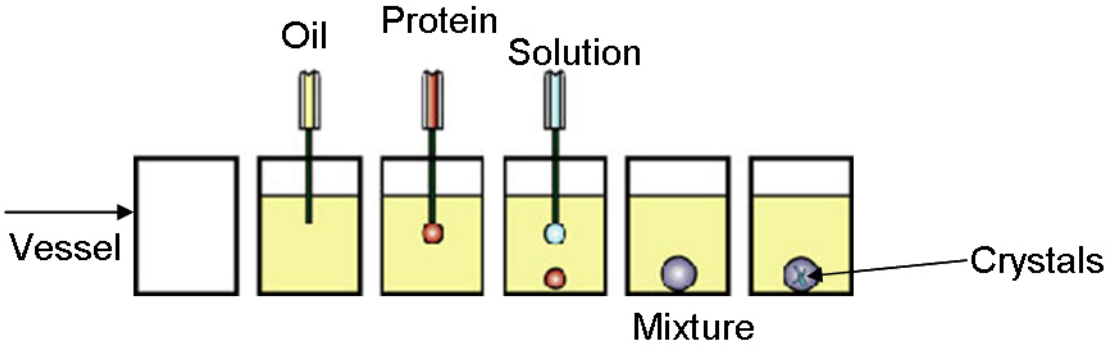


Figure 1.2: The experimental procedure for microbatch-under-oil method.

The small volume protein and solution are mixed under oil in the vessel. After a suitable incubation period, crystals can be observed in the experiment drop.

$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\left|F_{(h,k,l)}\right|\exp\left[-2\pi\cdot i(hx+ky+lz-\alpha_{(h,k,l)})\right]$$

Figure 1.3: The electron density equation.

The Fourier transform equation used for calculating electron density (ρ) at any coordinates in the unit cell (x,y,z). $|F_{(h,k,l)}|$ is the structure factor amplitude of reflection (h,k,l), including the temperature factor. $\alpha_{(hkl)}$ is the phase angel. V is the volume of the unit cell. Because an intensity $I_{hkl}=F_{hkl}^2$, only phase information, $\alpha_{(hkl)}$, is lost in a single crystal X-ray diffraction experiment.

Table 1.3: The phasing methods

Commonly use methods for biomacromolecule structures determination. Modified from The Crystal Structure Analysis for Chemists and Biologists (Glusker, Lewis et al. 1994).

Method	Note	Use
Patterson function	Map of interatomic vectors. Analysis is complicated unless there are only few atoms or heavy atom is present.	Used to locate heavy atom in isomorphous replacement for macromolecules. Can also be used for small molecules even if they contain no heavy atom.
Direct Methods	Aim for no negative areas in electron-density map. Probabilities of phases analyzed.	Used for small molecules.
Molecular replacement	Known part of structure positioned in unit cell.	Used on large molecules. Eliminates need for heavy-atom derivatives.
Isomorphous replacement	Intensity differences for isomorphous crystals. Best if several derivatives are made. Replaced atom located by Patterson methods.	General method for macromolecules.
Anomalous scattering	Needs good data. Requires anomalous scatterers, metal and sulfur.	Good with a tunable X-ray source, such as synchrotron radiation.

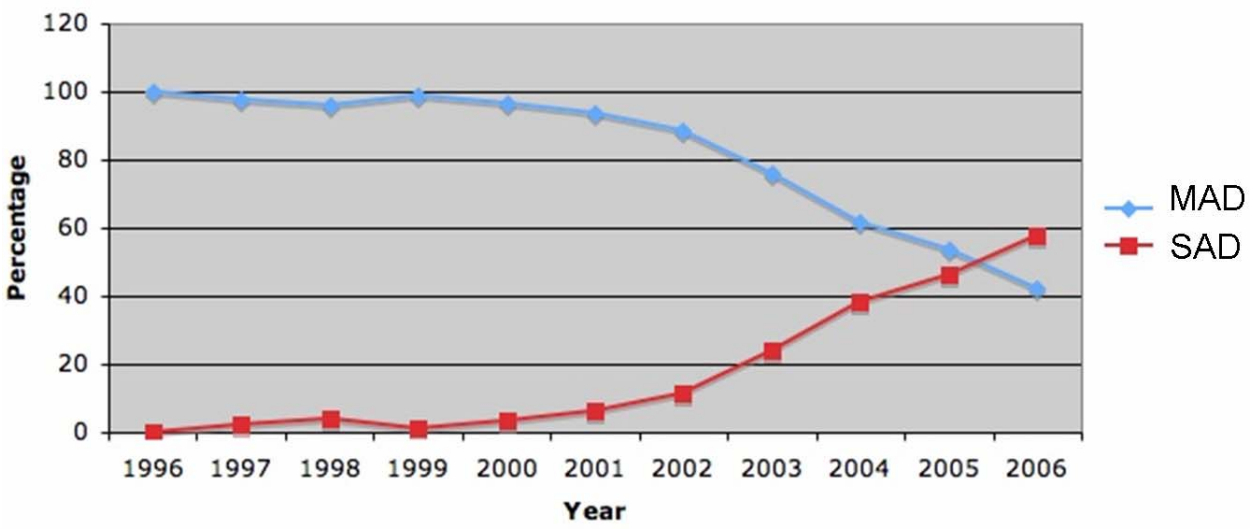


Figure 1.4: The statistics of PDB percentage of MAD vs SAD structural deposits.

The SAD method is more popular today. The number of structures solved using SAD method is bigger than that using MAD method since later 2005.

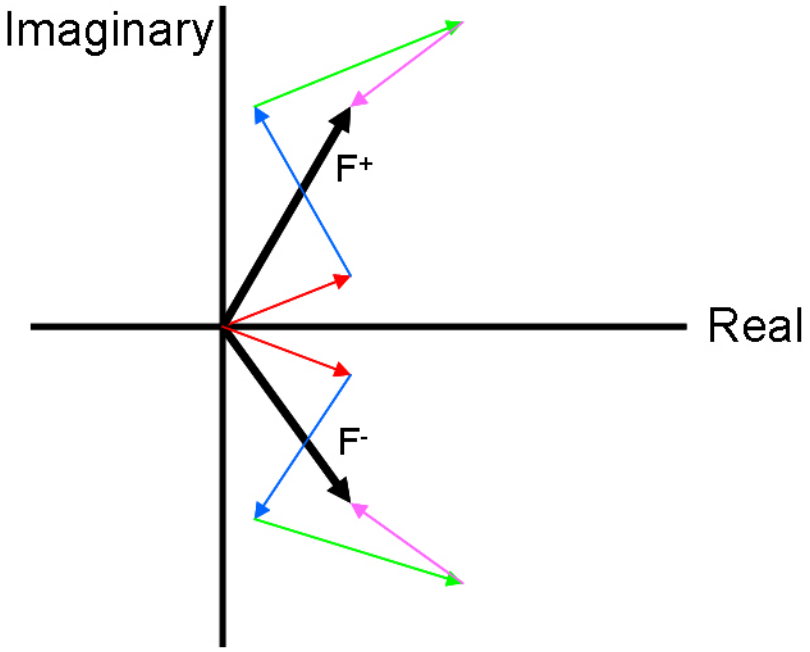
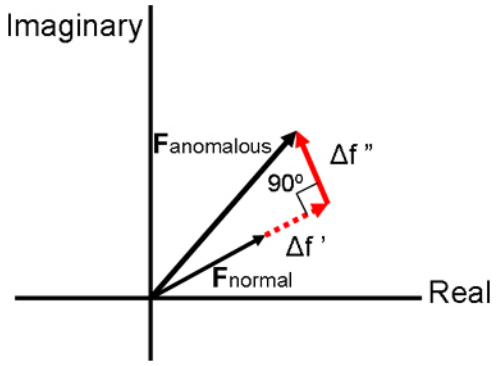


Figure 1.5: The Friedel's law illustration

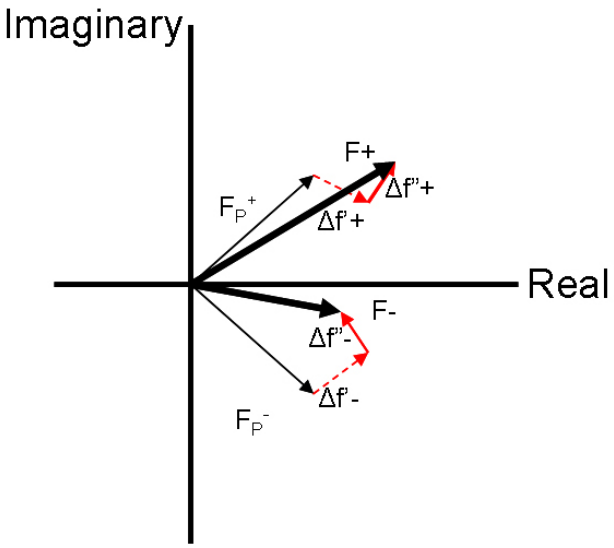
$F^+=F^-$. Structure factor (F) is the sum of all the atomic scattering vectors in the unit cell representing using different colors.

Fanomalous = Fnormal + Δf' + *i*Δf"

a



b



c

Figure 1.6: Anomalous scattering and breaking Friedel's law

(a) The summation equation of anomalous scattering. Real, Δf', and imaginary, Δf", components are added to normal Thompson scattering. **(b)** Vector summation of the equation in (a) for a single type of atom. **(c)** The break of Friedel's law by anomalous scattering in the case of two types of atoms, where one type displays anomalous scattering and the other does not (F_P). The final F_{hkl} for the positive and negative (h,k,l) value does not have the same magnitude (i.e. measured intensity) and the relative phases to the real axis are different. The Δf' vector merges the Thompson and real component of the anomalous scattering into a single vector.

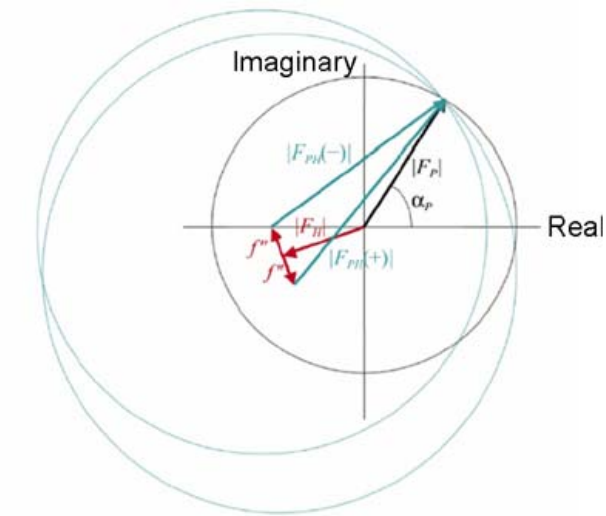


Figure 1.7: The SAD illustration.

After introducing one heavier atom into the crystals, we can use its phase ($|F_H|$) and the f" to calculate the final phase solution.

Adapted from (Taylor 2003)

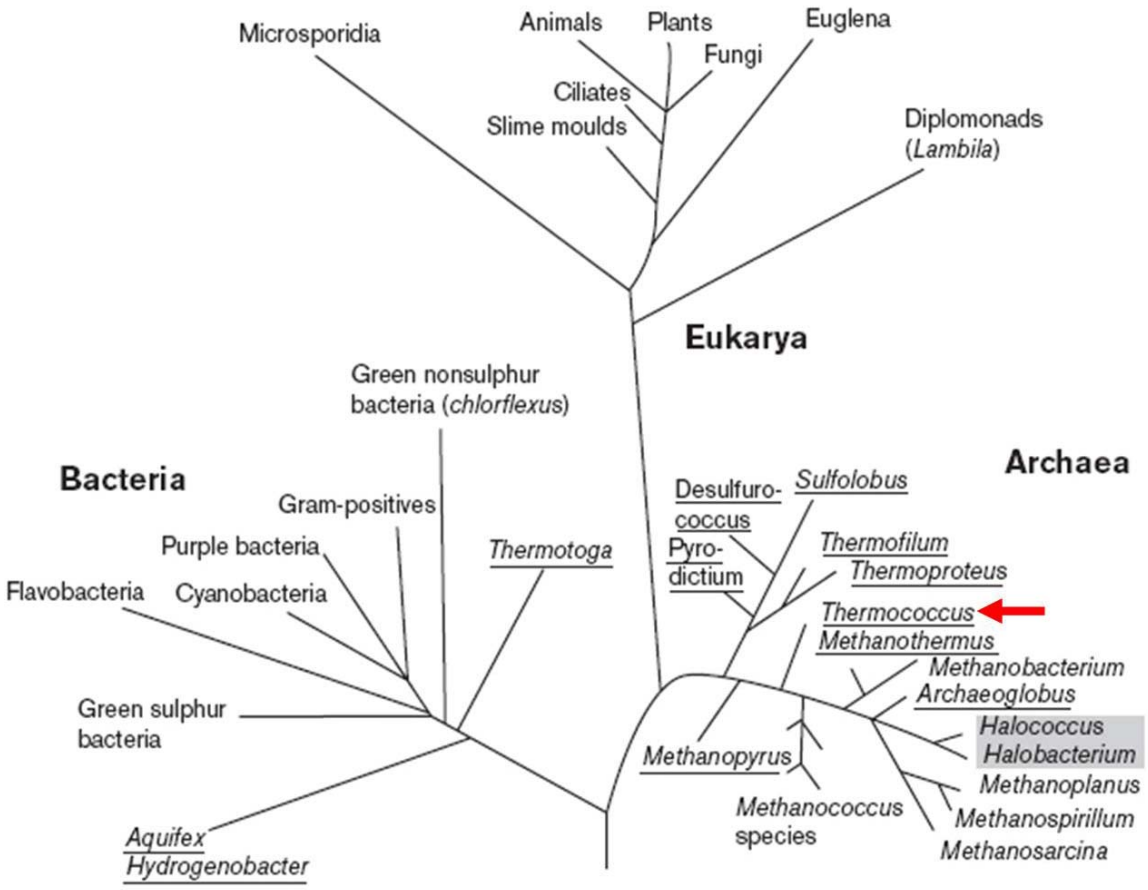


Figure 1.8: Phylogenetic tree of life.

Three domains of life are represented by branches. Thermophilic and hyperthermophilc species are underlined. Halophilic species are shaded. *Pyrococcus furiosus* belongs to the Thermococcus class, marked by red arrow.

Adapted from (Hough and Danson 1999).

Chapter 2

Crystal structure of PF0863, a putative adenylate cyclase from *Pyrococcus furiosus*: the first model of a Pfam CYTH domain superfamily

2.1 Introduction

Open reading frame (ORF) 0863 from *Pyrococcus furiosus* encodes a 20.4 KDa protein (*Pfu*-838710), which has been annotated as a putative CyaB-like adenylyl cyclase (Robb, Maeder et al. 2001), first identified in *Aeromonas hydrophila* (Sismeiro, Trotot et al. 1998). Adenylyl cyclase (AC) catalyzes the reaction of adenosine triphosphate (ATP) to cyclic AMP (cAMP), which functions as secondary messenger to control many major cellular functions (Danchin, Pidoux et al. 1993). The activity of adenylyl cyclase can be found from prokaryotic to eukaryotic species. Three well defined classes of adenylyl cyclase are based on the *cya* gene sequences analysis from various organisms: the enterobacteria class, including *Escherichia coli*; the "toxic" class, including calmodulin-activated enzymes from *Bordetella pertussis* and *Bacillus anthracis*; and the universal class, including homologues from bacteria to human; in which nine isoforms are found in mammals (AC-1 to AC-9) (Danchin 1993; Nowak and Zawilska 1999). In *Aeromonas hydrophila*, the protein coded by *cyaB* gene is a second adenylyl cyclase from this organism. The protein sequence analysis of this *cyaB* coded protein does not match any of the three well defined ACs, and its unique biochemical characteristics, such as an optimal temperature of 65ºC and an optimal pH of 9.5, show that it can be classified as the first member of a fourth class of ACs (Sismeiro, Trotot et al. 1998). The protein databases search find

significant similarities between the adenylyl cyclase from *cyaB* gene and the gene products among hyperthermophilic archaea (Sismeiro, Trotot et al. 1998).

The CyaB-like adenylyl cyclases have been reported to be distant homologs of the soluble mammalian thiamine triphosphatases (Lakaye, Makarchikov et al. 2002). These two families of enzymes define a novel superfamily of catalytic domains called the CYTH domain (Iyer and Aravind 2002). CYTH domain superfamily is one member of the protein family (Pfam) (Bateman, Coin et al. 2004) with the accession number of PF01928. Currently there are 314 members in this Pfam, and they span all kingdoms of life. They are functionally identified as members of the adenylate cyclase family. The secondary structure prediction using JPRED (Cuff, Clamp et al. 1998) and PHD (Rost and Sander 1993) define this superfamily's catalytic core as a novel α+β scaffold with 6 conserved acidic residues and 4 basic residues (Iyer and Aravind 2002).

No three dimensional structure of CYTH domain had been determined when I started working on PF0863, which was identified to be a member of the this domain superfamily by the multiple alignments (Iyer and Aravind 2002). The protein sample purified using high throughput protocol by *P. furiosus* protein production core of SECSG led by Dr. Adams at University of Georgia did not provide diffraction quality crystals (<3Å) to give the structure solution. The PF0863 clone was provided by Dr. Adams lab and the new purification protocol was performed to get purer protein sample for crystallization trial. Purer protein sample and higher temperature crystallization condition provided crystals diffracted to 2.6Å using home source X-ray generator. The phase was obtained using Pt-SAS. Final model was refined and deposited into PDB with the ID 1YEM. The structure of PF0863 is the first released structure of the CYTH domain superfamily. It has 8 anti-parallel β strands that form a topologically closed tunnel, which, in

this structure, contains three phosphates inside. The structural analysis combining with its biochemical activity analysis done by Dr. Adams group provide insights into the functions and structure-function relationship of the CYTH domain superfamily. The three-dimensional structural similarity with yeast RNA triphosphatase Cet1p also sheds the light on the evolution relationship between those two protein families.

2.2 Material and Methods

2.2.1 Protein Expression, Purification and Biochemical Assay

The cloning and biochemical analysis were done by Dr. Adams group. The ORF encoding PF0863 (*Pfu*-838710, see www.secsg.org) was cloned into an expression plasmid pET24dBam and expressed in *Escherichia coli* host strain BL21 Star DE3 pRIL (Sugar, Jenney et al. 2005; Weinberg, Schut et al. 2005) to create the expression clone pETPF0863. For biochemical analyses, the protein was expressed in M9 defined medium (Sambrook and Russell 2001) supplemented with 110 μM Fe or 50 μM Zn. The recombinant protein contained an N-terminal His_6 purification tag (AHHHHHHGS-) and was purified according to the high-throughput protocols established for *P. furiosus* protein production by the SECSG (Adams, Dailey et al. 2003; Sugar, Jenney et al. 2005). This involved a single column chromatography step: immobilized metal (Ni) affinity (EMD Biosciences, Madison, WI) which rendered a pure protein. The mass of the purified protein was determined by electrospray ionization mass spectrometry at the University of Georgia Department of Chemistry facility, as well as using analytical gel filtration with a Superdex75 column. Metal content was determined either by inductively-coupled plasma emission spectroscopy (ICP) at the University of Georgia Chemical Analysis Laboratory, or by a colorimetric determination (Fe only) as described (Lovenberg, Buchanan et al. 1963).

To measure the released phosphate concentration with ATP, GTP, ADP, and GDP as substrates, a reaction mixture of 100 µl including 25 mM of EPPS at pH 8, 20 mM of $MgCl_2$, 2 mM of NTP or NDP, and 0.25 µM of PF0863 was incubated in 80 ºC for 15 minutes. 40 µl of the reaction was added to 260 µl of dH_2O, then 700 µl of freshly made 1 to 6 mixture of 10% ascorbic acid and 0.42% ammonium molybdate in 1N H_2SO_4 was added to make the mixture 1000 µl. The mixture was incubated at 45 ºC for 25 minutes, and its absorbance was measured at 820 nm. The phosphate concentration in the mixture was calculated with a standard curve.

To identify the reaction product, HPLC analysis was deployed. An XTerra column RP 18 (4.6 mm x 100 mm, 3.5 mm particle size) coupled to a Waters 2690 HPLC was used for the analysis. The column was first equilibrated with 65 mM KH_2PO_4, 0.9 mM tetrabutylammonium phosphate (TBAP) at pH 3.2 (Buffer A) with flow rate of 0.6 ml/min at 40 °C. 10 µl of each the above mixtures was loaded onto the column after purification with microcon YM-3. The bound nucleotides were eluted with a linear gradient from Buffer A to 20% (v/v) acetonitrile over 12 mL. ATP, ADP, AMP, cAMP, GTP, GDP, GMP, and cGMP were used as standards.

2.2.2 Re-purification, Crystallization and Data Collection

To obtain the diffraction quality crystals, PF0863 was re-purified. The expression plasmid pETPF0863 was transformed into *E.coli* BL21-CodonPlus® (DE3)-RIPL Competent Cells (Stratagene). The transformation was plated onto an LB agar plate containing 50µg/ml Kanamycin and 35µg/ml Chloramphenicol, and the plate was incubated at 37°C overnight. Several positive transformants were grown in 5ml LB media with isopropyl-β-D-thiogalactoside (IPTG) induction to test for the expression of recombinant protein. For large scale purification, the cell was grown in 5L 2YT medium (Sambrook, Fritsch et al. 1989) at 37ºC to an OD_{600} of 0.6. The culture was then induced with 1mM IPTG and incubated overnight at 20ºC. Six grams of

cell pellet were collected. After the cells were lysed by sonication on ice using a Branson Sonifier Cell Disruptor 450 at power level 7 with 6 bursts of 30 seconds on and 30 seconds off, and centrifugation at 12,000rpm for 30 minutes, the supernatant was heated in a 70ºC water bath for 60 minutes. The supernatant was recovered by centrifugation at 12,000rpm for 30 minutes and filtered. The filtrate was next passed over a DEAE Sepharose column (56mL bed volume). The flow rate used was 2 mL/min. The loading buffer was 25mM Tris-HCl pH 7.5, 10mM EDTA (ethylenediaminetetraacetic acid), 20mM NaCl, 1mM PMSF (phenyl methyl sulfonyl fluoride), and 1mM β–mercaptoethanol, and the elution buffer was 25mM Tris-HCl pH 7.5, 10mM EDTA, 1000mM NaCl, 1mM PMSF, and 1mM β–mercaptoethanol. After sample loading, the column was washed with 1 bed-volume of loading buffer, followed by elution with 0% to 100% gradient of the elution buffer over 2 bed volumes. Peak fractions (10 fractions of 3 mL) were analyzed with SDS-PAGE and loaded onto a hydroxyapatite column (20mL bed volume). The column was packed by ourselves with resin from BioRad (Catelog # 157-0021), and the flow rate was 2 mL/min throughout this step. The loading buffer for hydroxyapatite column was 25mM phosphate buffer pH 7.0, and the elution buffer was 25mM phosphate buffer pH 7.0 and 1000mM K_2HPO_4. Two bed-volume wash was done after sample loading, and elution was performed with 0% to 100% gradient over 3 bed-volumes with the elution buffer. Peak fractions (10 Fractions of 2 mL) were analyzed using SDS-PAGE, concentrated with Millipore Amicon Ultra-15 Centrifugal Filter Unit of 10,000 MW cutoff to 2mL, and loaded onto a gel filtration column. For gel filtration chromatography, a Superdex G30 (320mL bed volume) was used with 20mM Tris-HCl pH 8.0, 2mM DTT, 200mM NaCl as running buffer, and the flow rate was 0.5 mL/min. Fractions (8 Fractions of 3 mL) were analyzed by SDS-PAGE, pooled, and concentrated with Millipore Amicon Ultra-15 Centrifugal Filter Unit of 10,000 MW cutoff.

Protein concentration was calculated based on extinction coefficient (Gill and von Hippel 1989) of 20340 M^{-1} cm^{-1}at 280nm absorption to be 15 mg/mL. Aliquots of 100µl were stored at -80ºC and thawed immediately prior to crystallization trials.

Crystallization experiments were performed by the modified microbatch under oil method (Chayen, Shaw-Steward et al. 1992; Baldock, Mills et al. 1996) with MicroWell™ MiniTrays (Nunc™) . The initial crystallization conditions were identified from the Crystal Screen (Hampton Research, Aliso Viejo, CA) when the protein was screened against 384 conditions (Crystal Screen, Crystal Screen 2, MembFac, PEG/ION Screen, and Crystal Screen Cryo from Hampton Research, and Wizard I and II from Emerald BioSystems, and MemSys from Molecular Dimensions). For optimization, the protein was heated to 75ºC for 6 minutes, filtered by centrifugation for 10 minutes at 12,000rpm, and then heated again for 1 min at 75ºC just before the crystallization setup. Protein (0.5µl) at 15 mg/mL was mixed with 0.5µl crystallization reagent mixture (100mM HEPES/NaOH, pH: 6.9-8.2, 800-1200mM NaH_2PO_4, 800-1200mM KH_2PO_4) in the setup. The wells were covered with 70:30 paraffin:silicone oil mixture to retard dehydration. The crystallization tray was incubated at 40ºC. Crystals reached dimensions of 200×200×300µm (Figure 2.1) after 3-4 days. The possible substrates (0.2µl), GTP and GDP at 5mM, were added into the crystallization wells with crystals. Also, co-crystallization of the substrates with PF0863 was tested. GTP and GDP were mixed with PF0863 with 1:1 molar ratio individually and incubated at 80ºC for 15 minutes. Then the mixtures were set up crystallization screening using the same method and conditions as described above.

A native single crystal was harvested with a cryoloop (Hampton Research, HR4-747) and briefly immersed in a 1 µL drop containing 2M Li_2SO_4 as a cryoprotectant. A platinum

derivative crystal was obtained from a quick soak in 25mM K_2PtCl_4. The crystals were flash cooled, retrieved, and stored in liquid nitrogen. Data that led to the structure solution were collected at cryogenic temperatures on a Smart6000 at 1.5418 Å. High resolution data were collected at beamline 22-ID (SER-CAT), Advanced Photon Source, Argonne National Laboratory using a MAR 165 CCD detector and 1.06 Å X-ray. Data were indexed, integrated, and scaled using the HKL2000 software suite (Otwinowski and Minor 1997), and the resulting statistics is in Table 1.

2.2.3 Structure Determination and Refinement

Two anomalous scatterers (soaked Pt) were located, and initial phases based on these sites were obtained with Solve/Resolve (Terwilliger 2002) which were incorporated into SCA2Structure pipeline (Liu, Lin et al. 2005) using the single wavelength anomalous scattering option. The experimental phases were improved using non-crystallographic symmetry averaging with DM (Cowtan and Zhang 1999) in CCP4 (1994). The initial model automatically traced about 30% by Resolve was completed manually using XFIT (McRee 1999) and was refined using REFMAC5 (Murshudov, Vagin et al. 1997). NCS restraints were employed in all stages of the refinement except for the terminal residues where two monomers have different conformations. The structure validation was performed using MOLPROBITY (Lovell, Davis et al. 2003) and PROCHECK (Laskowski, Moss et al. 1993). The refined model is available from the Protein Data Bank (Berman, Westbrook et al. 2000), entry 1YEM.

2.3 Results and Discussion

2.3.1 Analysis of the Gene and Sequence

The primary annotation of PF0863 from the complete genome sequence of *P. furiosus* is a hypothetical protein. It may have regulatory functions involved in the small molecule

interactions based on the classification in the Institute for Genomic Research (TIGR) cellular role category. The PF0863 gene encodes a protein of 171 residues with a predicted molecular weight of 20417 Da. The sequence-based PSI, PHI-BLAST shows that PF0863 belongs to the CYTH domain superfamily defined by the catalytic domains of CyaB-like adenylyl cyclase and thiamine triphosphatase (Iyer and Aravind 2002). In the Pfam database, the CYTH domain superfamily has 314 members (Bateman, Coin et al. 2004) belonging to prokaryote, archaea, and eukaryote. The sequences of all the members are functionally identified as the adenylate cyclase, which catalyzes the conversion of ATP to 3', 5'-cyclic AMP and pyrophosphate. No detectable homology to any other protein of known function has been identified, and no protein structure has been determined. The whole-family member sequence alignment used T_Coffee program with further refinement based on the PSI-BLAST HSPs (Iyer and Aravind 2002). The secondary structure prediction using JPRED (Cuff, Clamp et al. 1998) and PHD (Rost and Sander 1993) defined this superfamily's catalytic core as a novel α+β scaffold.

PF0863 is the last ORF in a putative three ORF operon (Figure 2.2). An operon is a group of nucleotide sequences containing an operator, a promoter, and one or more structural genes that are controlled as a unit to produce messenger RNA (mRNA). The first discovery of operon was in 1961 (Jacob and Monod 1961) and primarily found in prokaryotes. There is one paralog of PF0863 in *P. furiosus*, PF1859 (35.5% identity; 53.6% similarity), also the third ORF in a putative operon with two additional ORFs (Figure 2.2). PF1859 is also annotated by the Interpro database (Mulder, Apweiler et al. 2003) as a CyaB-like adenylyl cyclase.

2.3.2 High Throughput Purification and Biochemical Characterization (Done by Dr. Adams lab)

The clone (pETPF0863) was sequenced before expression, and found to match the published sequence database exactly. Because of the N-terminal His-tag construction on the expression vector, the purification steps used Ni^{+} affinity followed by gel filtration chromatography. For biochemical characterizations, the recombinant His-tagged protein was purified from induced *E. coli* extracts from cells grown in Fe- or Zn-supplemented minimal media using Ni-NTA affinity column chromatography. The protein yield was 34 mg (Zn) and 80 mg (Fe). Metal analysis (ICP) indicated the recombinant protein contained only 0.34 - 0.54 moles of nickel per mole of monomer in either condition. Colorimetric estimation of iron content also indicated that the protein did not contain iron. The experimentally determined monomeric mass by mass spectrometry was 21,388 Da, which closely matched the predicted average mass (21,384 Da) with His-tag. The native mass, determined by FPLC using an analytical Superdex 75 column, was found to be around 40 KDa, closely matching the predicted mass if the protein was dimer (predicted dimer mass, 42,787 Da).

Recombinant PF0863 did not exhibit adenylyl cyclase activity but did show nucleotidase activity, hydrolyzing nucleosides tri- and diphosphates. Nucleotidase activity was tested by measuring the amount of Pi released (Hutchins, Holden et al. 2001). Several nucleoside di- and triphosphates were tested as substrates with GDP being the most efficiently converted substrate. The enzyme required Mg^{++} for activity, and no activity was obtained using Ca^{++} or Mn^{++}. The optimal temperature for the nucleotidase activity assay was determined to be 85ºC. Kinetic analyses of PF0863 using ADP and GDP in the nucleotidase activity assay gave K_m and V_m values of 1.5 mM and 5 μmoles $min^{-1}mg^{-1}$, respectively, with GDP and 1.6 mM and 2 μmoles

$min^{-1}mg^{-1}$, respectively, with ADP. The PF0863 catalyzed reaction products were also analyzed by HPLC (using ATP, ADP, GTP, and GDP as substrates). Results showed that GTP was converted to GDP, GDP was converted to GMP, ATP was converted to ADP, and ADP was converted to AMP. No cAMP or cGMP was detected (data not shown).

2.3.3 Re-purification and Crystallization

The purified protein sample using high-throughput protocol produced protein crystals but they diffracted to only 3.9Å at the Advanced Photon Source. The re-purification steps used DEAE sepharose column, hydroxyapatite column, and gel filtration superdex G30 column. The three-step purification produced a high purity sample for the crystallization (Figure 2.3). The crystal quality was significantly improved using the purer protein sample and diffracted to 2.6Å using a home source X-ray with 1.5418 Å wavelength. It is an excellent example to emphasize the importance of protein purity which is slightly sacrificed in the high throughput efforts.

The often-used crystallization temperature is 18°C. Based on our previous experience with crystallizing *P. furiosus* proteins and the consideration of the characteristics of hyperthemophiles, the crystallization trays were set up at both 18°C and 40°C. The higher temperature at 40°C crystallization produced better diffraction quality crystals for data collection. As introduced in Chapter 1, crystallization is multi-parameter problem. Temperature is an important factor affecting crystallization results. Alternative higher temperature crystallization for hyperthermophile proteins is a reasonable consideration.

The crystals cracked when GTP or GDP was added into the crystallization wells. They could not be used for any data collection trial. The co-crystallization of GTP or GDP with PF0863 did not produce crystal after screening against 384 conditions. Crystallization result is hard to predict, especially for co-crystallization with substrates. Even though different

crystallization methods and various crystallization conditions may produce crystals in the future, there is no guarantee that we can achieve this goal.

2.3.4 Structure Description

Two PF0863 molecules form a homodimer in one asymmetric unit (Figure 2.4a). The analytical gel filtration chromatography also showed PF0863 as a dimer in solution. Residues 1-164 in molecule A and residues1-166 in molecule B were clearly defined in the electron density map. The N-terminal His-tag was not visible in either molecule. Each monomer of PF0863 structure has a barrel structure formed by 8 anti-parallel β-sheets. β3-β4-β5 form a half circle, and β2, β6-β7-β8, and β1 form the other half circle of the barrel structure. The collection of β sheets in the formation of the barrel gives a "tornado-like" shape from the top view (Figure 2.4b). Three major alpha helices are located outside the barrel in each monomer. The connection between β1 and β2 by using α1 circumvents β6-β7-β8. α2 follows β5 encompassing β3-β4-β5, then connects to β6. α3 is right after β8. The C-terminal end contains the loop and coil region which forms the bottom of the barrel structure and creates a closed tunnel. Based on the multiple sequence alignment of CYTH domain superfamily, there are 26 residues with an amino acid conservation at 90% consensus (Iyer and Aravind 2002). Most of the conserved residues are located on the β-stands which form the barrel. It suggests that the family members of CYTH domain superfamily may all have the same barrel structure. The secondary structural prediction suggests CYTH domain contains 6 α-helix and 6 β-stands (Iyer and Aravind 2002), the real three dimension crystal structure contains the barrel structure formed by 8 anti-parallel β-sheets instead. Even though computational development is fast growing and becoming more powerful on protein structure prediction, the experimental structure result is still more valuable and accurate to help understanding its function and structure-function relationship.

The dimer is formed between the ends of the β5 sheet of two monomers with a non-crystallographic 2-fold symmetry. The residues Glu76A/Ile77B and Ile77A/Glu76B on the fifth β sheets form the hydrophobic dimerization core. The sequence alignment of the archaea CYTH domain superfamily shows highly conserved residues involved in the dimer construction. The two monomers are twisted with respect to each other forming approximately 80 degree twists along the diagonal of the unit cell. The surface of the overall structure is more negatively charged (Figure 2.5). A curve made by the dimer formation has an outer surface with a positive-negative-positive charged pattern along the twist. The inner surface of the curve, which is also near the entrance of the barrel tunnel, is covered by negatively charged residues while the inside of the barrel tunnel has more positively charged residues.

The extra electron density in the barrel was observed. The results from biochemical assays show that PF0863 has nucleotidase activity and does release phosphate, and GDP seems to be the most efficiently converted substrate of PF0863. Using GDP to fit the density map was performed without success. The reason may be caused by the low occupancy and relatively low resolution, or maybe GDP is not the most favorable substrates. At last, after careful observation of the shape of electron density map, three phosphates were assigned in each monomer. Also, in the crystallization conditions, free phosphates are present in the well solution. The hydrogen bonds were formed among those phosphates with K6, R45, R47, K60, K100, and R102 (Figure 2.6). K6, R45, R47, and K100 are highly conserved in the CYTH domain superfamily, and they were predicted to be involved in the substrate binding (Iyer and Aravind 2002). The interaction between those basic residues and the phosphates in the barrel defines the catalytic core and mimics PF0863 binding to its substrate.

2.3.5 Comparison with a Similarly Folded Structure

A search for similar protein folds with DALI (Holm and Sander 1993) revealed a significant structural similarity between PF0863 and yeast RNA triphosphatease Cet1p which catalyzes the first step in mRNA processing, the cap formation, in which the 5'-triphosphate end of pre-mRNA is hydrolyzed to 5'-diphosphate (Lima, Wang et al. 1999), on the barrel structure (Figure 2.7). Unlike the structure prediction on CYTH domain to be a novel α+β fold (Iyer and Aravind 2002), this barrel structure is not new. DALI Z score between two structures, which represents strength of structural similarity in standard deviations above expected, is the higher the more similar (Holm and Sander 1993). The three dimensional comparison between yeast RNA triphosphatase Cet1p (PDB ID: 1D8I) and PF0863 has the DALI Z score of 9.4. The positional root-mean-square deviation of superimposed CA atoms between those two structures is 3.4Å; the total number of equivalent residues is 136 for which their identity is 14%.

Both barrel structures from those of yeast RNA triphosphatase Cet1p and PF0863 have hydrophilic side chains, which dominate the inside; the structural based sequence alignment is shown in Figure 2.8. In yeast RNA triphosphatase, there are two conserved motifs A (ELEMKF) and C (EVELE) involved in one divalent cation binding (Lima, Wang et al. 1999). In PF0863 structure, almost exact overlap of EXEXK as the motif A and DXE as the motif C are present. The finding of metal-binding motif in PF0863 is consistent with the biochemical assay result of the nucleotidase activity requiring Mg^{++}. E2, E4, and K6 on β1, R45 and R47 on β2, K60 on β3, E72 on β5, K100 on β6, and D125 and E127 on β8 are highly conserved residues among CYTH domain superfamily (Iyer and Aravind 2002). They also show side-chain identity or similarity with yeast RNA triphosphatase Cet1p on the barrel fold. The mutation on E305, E307 ,E496, and K456 in yeast RNA triphosphatase Cet1p, which are counterparts to E2, E4, E127, and K100

of PF0863, show critical effect on its activity. Six conserved acidic positions predicted in the CYTH domain suggest that it may coordinate two divalent metal ions. But only in archaea family members, one predicted acidic position is missing, which corresponds to the position on the β7 of PF0863 structure. This information may suggest that in archaea, there may be only one divalent cation required for the function.

Till today, two other protein structures from CYTH domain are available in the Pfam database (November 2006). One is a putative adenylate cyclase Q87NV8 from *Vibrio parahaemolyticus* by the Northeast Structural Genomics Consortium (NESG), PDB ID 2ACA. Another one is a hypothetical protein NE1496 from *Nitrosomonas europaea* solved by Midwest Center for Structural Genomics (MCSG), PDB ID 2FBL. *V. parahaemolyticus* is a facultative anaerobic, gram negative bacterium. *N. europaea* is also a gram negative bacterium. All three structures from two different domains of life have the same homodimer in one asymmetric unit providing more evidence to support that the CYTH domain superfamily members may be biologically active as a dimer. The sequence identity between PF0863 2ACA, and 2FBL sequences on the basis of their matched secondary structure elements are 20.9% and 29.9%, respectively. The three dimensional structural alignment between PF0863, 2ACA, and 2FBL has the RMSD value of 2.09Å and 3.69Å, respectively. Structural genomics projects from different species generate a large number of crystal structures. Those two structures from other structural genomics center also do not have functional description, but the more structures available in this Pfam, the more information we can gather for the further study.

2.4 Conclusion

The CyaB protein from *Aeromonas hydrophila* has been shown to possess adenylyl cyclase activity. While orthologs of this enzyme have been found in some bacteria and archaea

(including PF0863 and PF1859), they show no detectable relationship to the classical nucleotide cyclases (Sismeiro, Trotot et al. 1998). The actual biological functions of these proteins are not clearly understood because they are also present in organisms in which there is no evidence for cyclic nucleotide signaling. The phylogenetic distribution of the CYTH domain suggests that it is an ancient enzymatic domain that was present in the Last Universal Common Ancestor (LUCA) and was involved in nucleotide or organic phosphate metabolism (Iyer and Aravind 2002).

The primary biological function of CYTH domain superfamily is predicted to be related to polyphosphate and nucleotide metabolism, while the generation of cyclic AMP and thiamine triphosphate hydrolysis seems to be secondary activities (Iyer and Aravind 2002). No cyclic nucleotide generating activity has ever been detected in archaea (Schultz and Klumpp 1994), and the biochemical assay did not detect that PF0863 has the functions as an adenylyl cyclase either.

For structure determination, the protein sample purified by a three step purification procedure using the DEAE ion exchange column, hydroxyapitite column, and gel filtration column yielded protein that produced diffraction quality crystals. This emphasizes that purity is a crucial factor for getting quality crystals for X-ray diffraction studies (Liu, Shah et al. 2005). The crystallization at 40°C was the better temperature to produce high quality crystals for data collection. This result provides evidence that using higher temperature to crystallize hyperthermophile proteins is a good alternative.

As the first crystal structure of CYTH domain superfamily, PF0863 structure shows a fold which is different from the prediction for this family. The tunnel of the barrel structure formed by 8 anti-parallel β-sheets is likely to be the catalytic core of PF0863. The dimer construction may be functionally required in the CYTH domain superfamily since three members

from bacteria and archaea all have dimer formations. The three dimensional structural similarity with yeast RNA triphosphatase Cet1p and biochemical characterization of PF0863 from *P. furiosus*, which exhibits no adenylyl cyclase activity that exists in the CyaB protein from *A. hydrophila*, shows that this *P. furiosus* protein is a nucleotidase.

CYTH domain exists in three domains of life, but not present in yeast (Iyer and Aravind 2002). Also, CYTH domain does not have any relationship with the RNA triphosphatase family. The significant three dimensional similarities between yeast RNA triphosphatase Cet1p and PF0863 shed a light on the re-consideration on definitions of these two protein families and their evolution relationship. The new name "triphosphatase tunnel metalloenzyme" superfamily was proposed recently (Gong, Smith et al. 2006). The firstly described yeast RNA triphosphatase Cet1p was the prototype of this superfamily, and CYTH domain was classified as a branch under the triphosphate tunnel fold. Can archaeal origin be the evolutionary precursor in "triphosphatase tunnel metalloenzyme" superfamily? Further detailed functional study of CYTH domain superfamily is needed to give clearer understanding in the future.

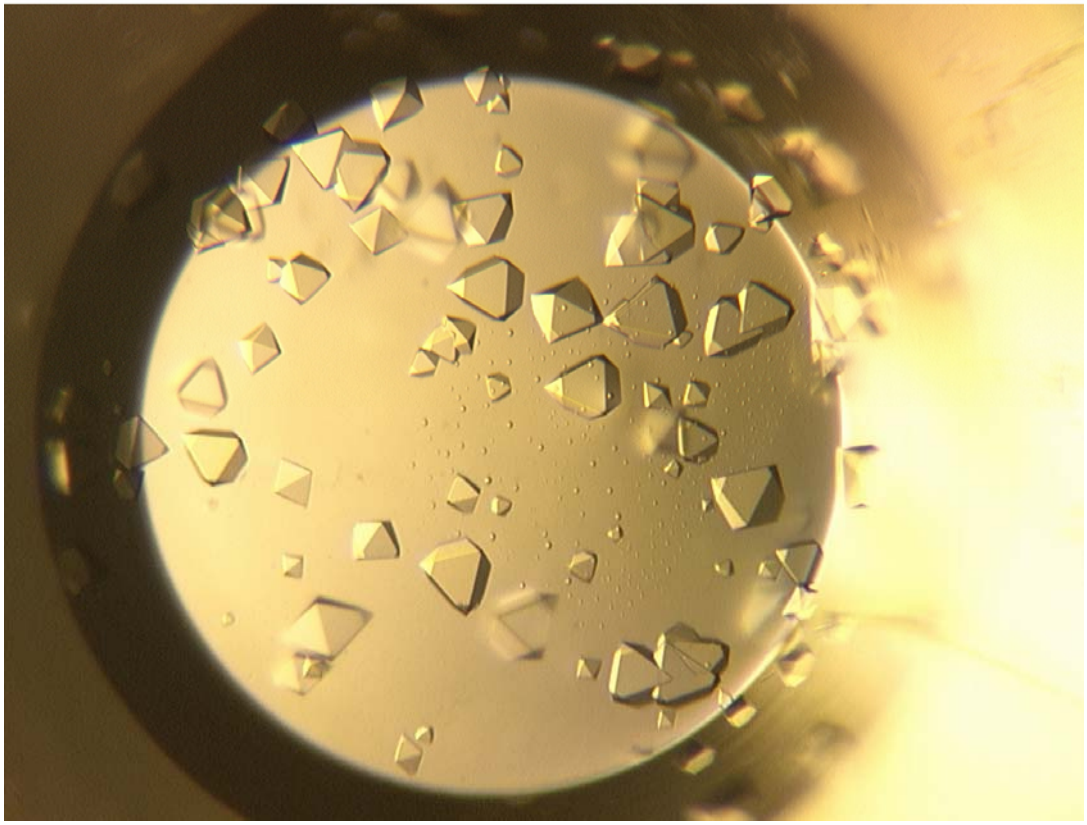
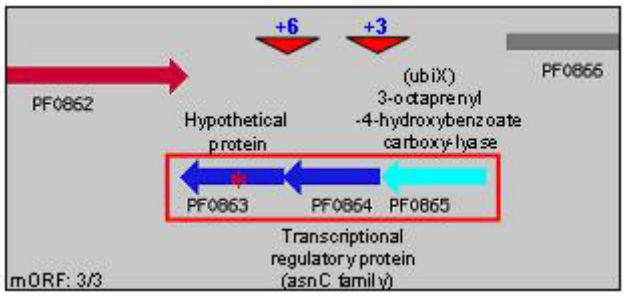
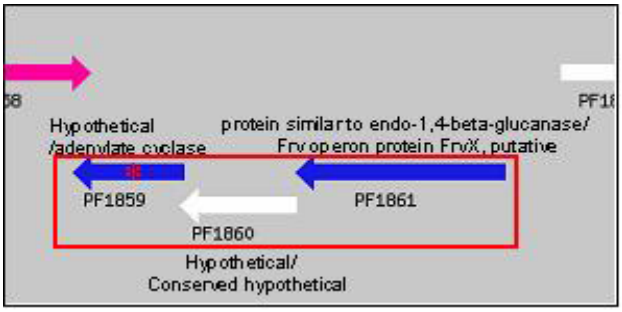


Figure 2.1: Crystals of PF0863.

0.5µl protein at 15mg/mL was mixed with 0.5µl crystallization reagent mixture (100mM HEPES/NaOH, pH: 6.9-8.2, 800-1200mM NaH_2PO_4, 800-1200mM KH_2PO_4) in the setup. The wells were covered with 70:30 paraffin:silicone oil mixture to retard dehydration. The crystallization tray was stored at 40ºC. Crystals reached dimensions of 200×200×300µm after 3-4 days.



a



b

Figure 2.2: Gene region view.

a. Region view of PF0863, PF0864, and PF0865 operon.
b. Region view of PF1859, PF1860, and PF1861 operon. PF0863 and PF1859 are both belonged to CYTH domain superfamily.

Figures are modified from TIGR-CMR.

Table 2.1: Statistics of PF0863 from the crystallographic analysis

Crystal

Space group:	$P3_121$
a=b	97.02Å
c	127.59Å
α=β	90º
γ	120º

Data processing statistics

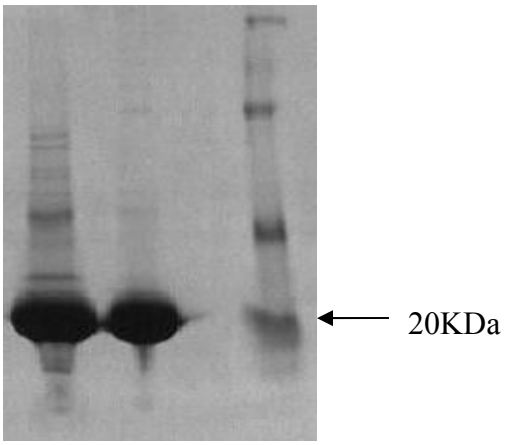
Wavelength (Å)	1.5418	1.06
Resolution (outer shell, Å)	50- 2.73 (2.91-2.73)	84.51-2.3 (2.38-2.3)
Completeness (%)	99.91	89.2
R_{sym}	0.079 (0.52)	0.065 (0.16)
Redundancy	16.29 (8.34)	9.2 (3.5)

Refinement statistics

Resolution range (Å)	84.51-2.3 (2.36-2.3)
Reflections used (free)	26652 (1430)
R-factor (R_{free}, %)	23.0(25.8)
Mean B factor (Å2)	46.56
RMSD bond lengths (Å)	0.007
RMSD bond angles (°)	1.431

Final model

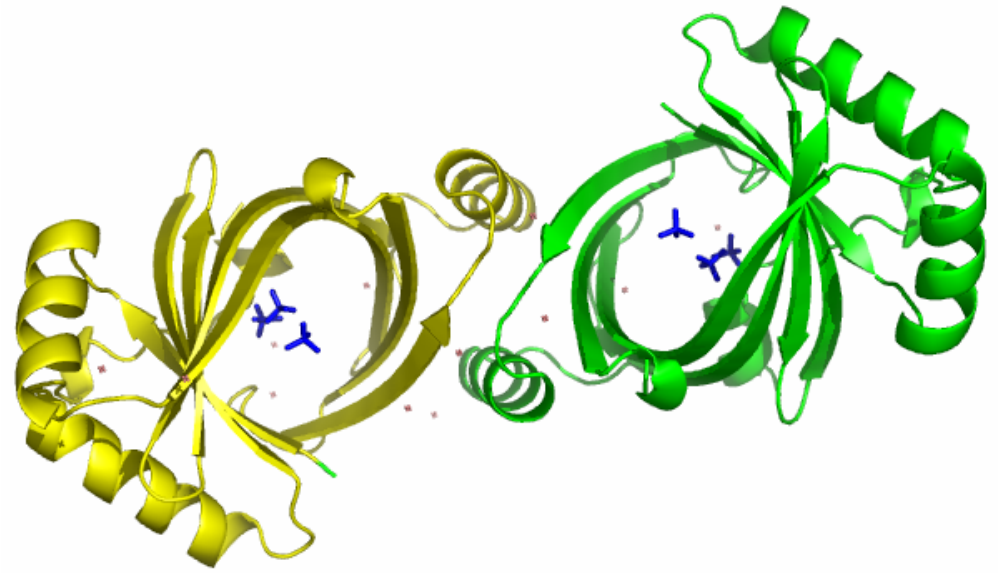
Residues	1-163 (Chain A); 1-166 (Chain B)
Protein atoms (solvent)	2770 (51)
PDB ID	1YEM



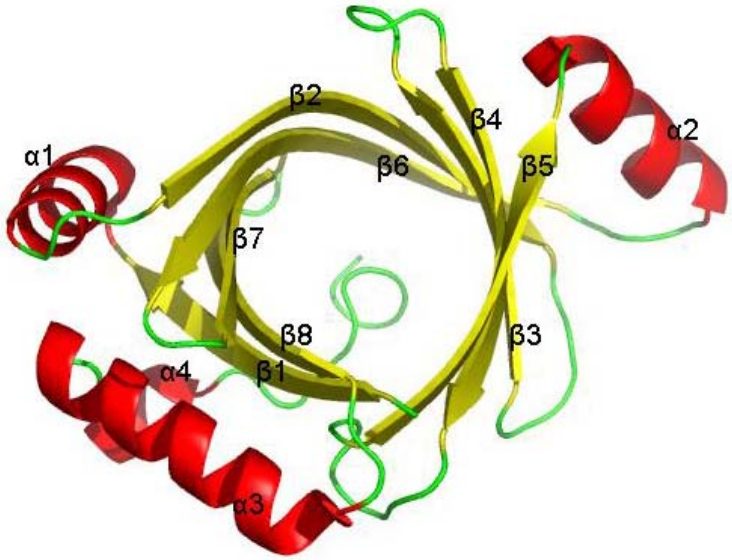
M: marker
1: high-throughput purified sample
2: re-purified sample

Figure 2.3: SDS-PAGE of purified PF0863 using different purification steps.

From the results of two different purification protocols, the three-step purification produced purer protein sample (lane 2) than the high throughput purification product (lane 1). The purer sample provides high diffraction quality data leading to the structure determination.



a



b

Figure 2.4: The overall structure of PF0863.

a. The dimer of PF0863 viewed along the non-crystallographic 2-fold axis.
b. The top view of chain A with the α-helices and β-strands shown as labeled coils and arrows.

Each monomer has a barrel structure formed by 8 anti-parallel β-sheets. The dimer is formed between the ends of the β5 sheet of two monomers with a non-crystallographic 2-fold symmetry.

[All structure figures were produced using PyMOL (DeLano 2002)]

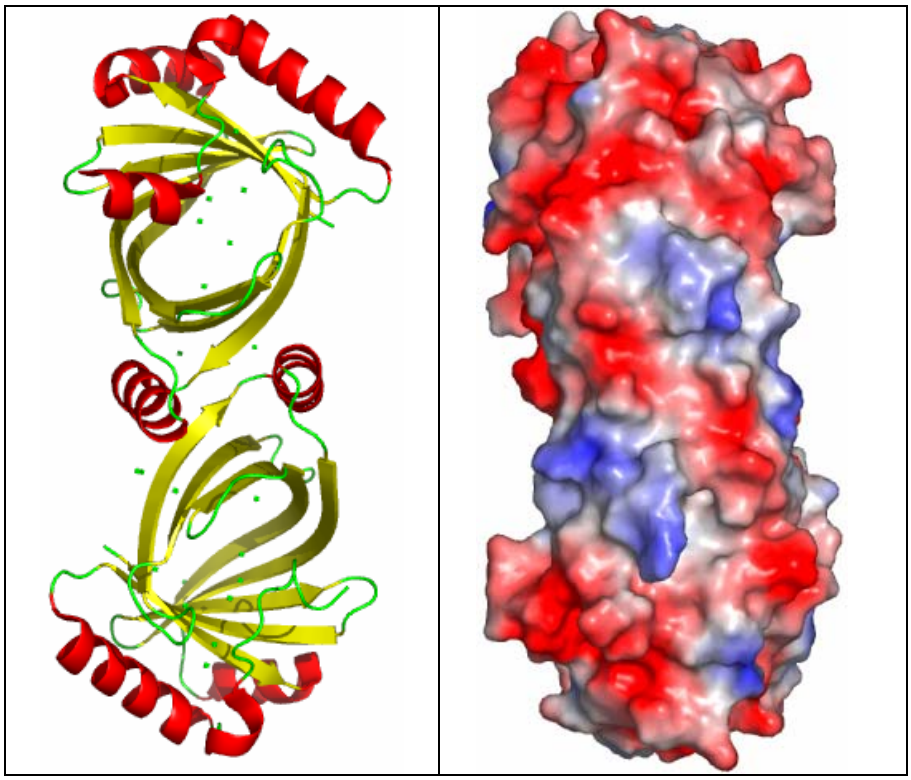


Figure 2.5: **Surface representation of the overall structure**.

Negatively charged surfaces are colored in red and positively charged surfaces in blue. Electrostatic surfaces were calculated using the program APBS (Baker, Sept et al. 2001) in PyMOL (DeLano 2002).

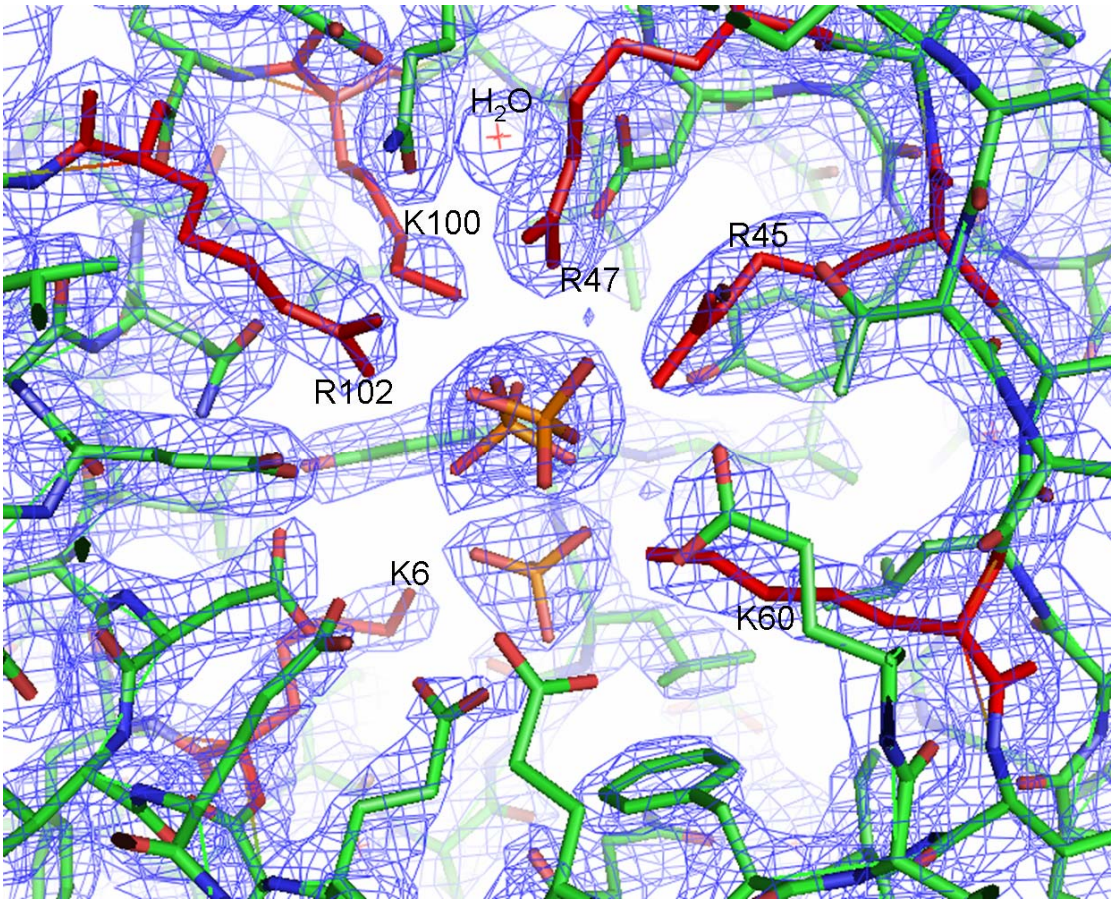


Figure 2.6: The top view of the barrel structure with three phosphates inside.

The phosphates and interacting residues are shown in stick. Those interacting residues are highly conserved in the CYTH domain superfamily.

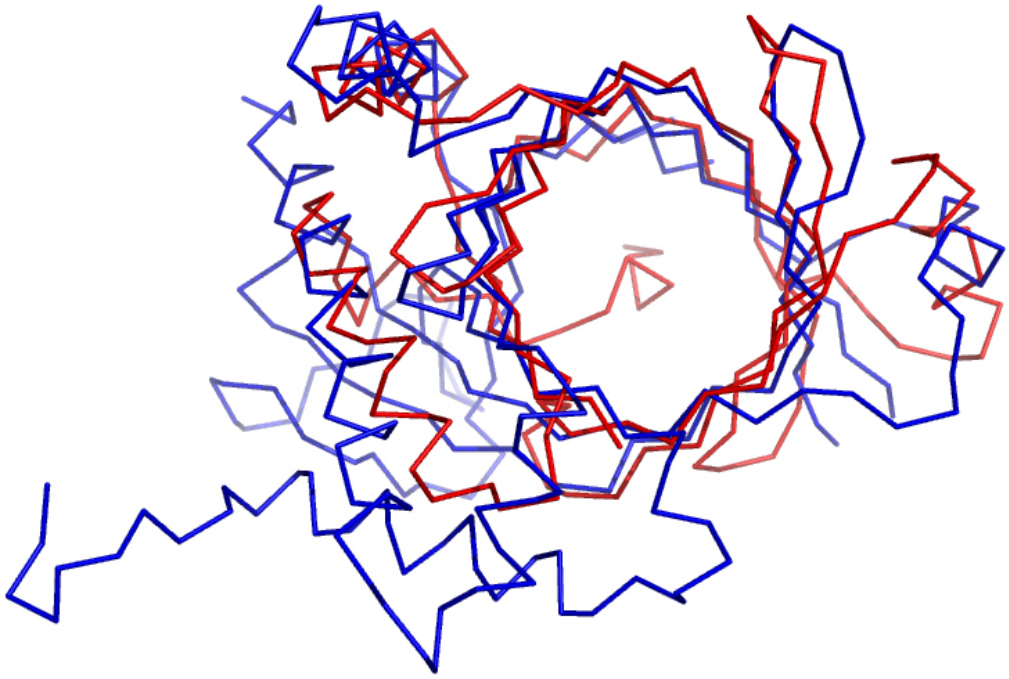


Figure 2.7: The structural overlap of yeast RNA triphosphatase Cet1p (1D8I) without the first 60aa (blue) and PF0863 (red).

The three dimensional alignment shows they have the same eight β strands barrel structure. The superimposition on the barrel structure has an RMSD of 1.22Å on 72 residues.

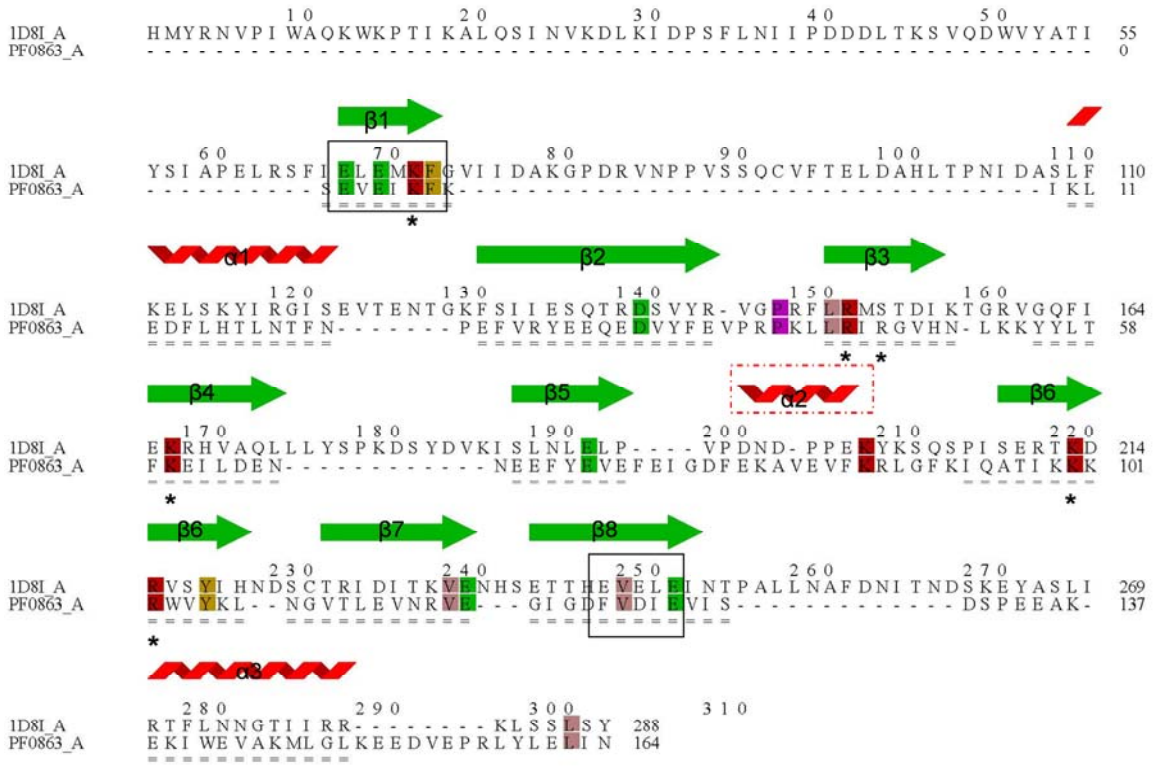


Figure 2.8: Structural-based sequence alignment between yeast RNA triphosphatase Cet1p (1D8I) and PF0863.

The two structures both have eight β stands. In PF0863 structure, there is an α helix after β5, but it is absent in 1D8I. The box in β1 represents the conserved motif A (ELEMKF) in 1D8I. The box in β8 represents the conserved motif C (EVELE) in 1D8I. The residues with asterisk underneath are involved in the interactions with PO_4 inside the barrel structure in PF0863. (INDONESIA package: D. Madsen, P. Johansson, and G.J. Kleywegt, unpublished data).

Chapter 3

The crystal structure of PF0864, one transcriptional regulator of Lrp/AsnC family from *Pyrococcus furiosus*

3.1 Introduction

PF0864 encodes a protein of 18.5kDa that is annotated as an Lrp/AsnC family transcriptional regulator in *Pyrococcus furiosus*. The transcriptional system in archaea, one of the three domains of life (Woese, Kandler et al. 1990) that includes *P. furiosus*, is the combination of that in eukarya and bacteria. The basal archaeal transcription machinery is a simplified version of that in eukaryotes, especially the RNA polymerase II system (Zillig, Stetter et al. 1979; Bartlett, Thomm et al. 2000), whereas most of the transcriptional regulators resemble those of bacteria (Aravind and Koonin 1999; Kyrpides and Ouzounis 1999; Geiduschek and Ouhammouch 2005). Although many of these bacterial-like archaeal regulators have been characterized at the molecular level (Bell and Jackson 2001), the understanding of the evolution in gene transcriptional regulation among all three domains of life can be further increased from research in this area.

The leucine-responsive regulatory protein (Lrp) family of transcriptional regulators is widely distributed among bacteria and archaea. It was first discovered in *Escherichia coli* (Anderson, Quay et al. 1976; Tuan, D'Ari et al. 1990) and was found in *P. furiosus* in 1995 (Kyrpides and Ouzounis 1995). Based on the most extensively studies from *E.coli*, Lrp family member function as a global regulator of amino acid metabolism and related processes, responding primarily to leucine (Newman and Lin 1995). Recent DNA microarray analysis

revealed that Lrp affects the transcription of at least 10% of all *E. coli* genes (Tani, Khodursky et al. 2002). AsnC is a specific asparagine-dependent activator of asparagine synthase (asnA), which can also autoregulate its own gene expression in an asparagine-independent way (Kolling and Lother 1985). Because AsnC shows notable sequence similarity (25% identity) to Lrp (Willins, Ryan et al. 1991), they are classified as part of the same evolutionary protein family, named the Lrp/AsnC family of transcriptional regulators. They are also termed feast/famine regulatory proteins (FFRPs) (Calvo and Matthews 1994; Suzuki 2003) to summarize the general functions of Lrp. Members of Lrp/AsnC family are typically DNA-binding proteins with the molecular weight around 15KDa. The multimeric state of this family members can be found as dimers, tetramers, octamers, and hexadecamers in solution (Willins, Ryan et al. 1991; Madhusudhan, Huang et al. 1995; Jafri, Evoy et al. 1999; Brinkman, Dahlke et al. 2000; Chen, Rosner et al. 2001). A number of binding sites exist in the target promoter region; they usually lack obvious palindrome pattern, which suggests the binding to DNA may be co-operative (Brinkman, Ettema et al. 2003).

Prior to the work on PF0864, a couple of Lrp/AsnC family structures were solved from archaeal organisms: *Pyrococcus. furiosus* LrpA (Leonard, Smits et al. 2001) and *Pyrococcus. horikoshii OT3* FL11 (Koike, Ishijima et al. 2004). Both structures reveal an N-terminal helix-turn-helix (HtH) motif, which is a typical DNA-binding domain. The C-terminal domain with the βαββαβ topology, which resembles the ACT domain family (Chipman and Shaanan 2001), is defined as RAM domain (Ettema, Brinkman et al. 2002). The ACT domain was coined (Aravind and Koonin 1999) after three of the allosterically-regulated enzymes in which this sequence domain is found: **a**spartate kinase, **c**horismate mutase, and **T**yrA (prephenate dehydrogenase). ACT domain is proposed to be a conserved regulatory binding fold, which can be found in a

broad range of metabolic enzymes that are regulated by amino acid concentration (Chipman and Shaanan 2001). RAM domain, referring to allosteric **r**egulation of **a**mino acid **m**etabolism, has significant difference with ACT domain in the effector-binding modules (Ettema, Brinkman et al. 2002).

PF0864 was not the high priority target in structural genomics project because its homologs' structures have already been determined as mentioned above. Also it did not produce favorable results with high throughput cloning and expression screening. Therefore, the structure study of this protein was paused; along with missing the opportunity to study the possible interaction with its gene neighbor until now. The crystal structure of PF0864 from *P. furiosus* was determined to 2.4 Å. Because of the failure in using high-throughput favorite molecular replacement method to solve the structure and the lack of sulfur in the protein, this structure determination process went through a traditional thorough heavy atom soaking screening experiment to solve the phase problem. The Au-SAS method gave the final solution of the crystal structure. Two obvious domains, the N-terminal resembles the HtH DNA-binding motif, and the C-terminal domain has the βαββαβ fold as the RAM domain, are observed in PF0864 structure. The DNA-gel shifting assay shows the specific binding activity of PF0864 to the putative promoter region upstream of the whole operon, which includes the genes of PF0865, PF0864, and PF0863.

3.2 Material and Methods

3.2.1 Sequence Analysis

Using PF0864 sequence as a query, PSI, PHI-BLAST (Altschul, Madden et al. 1997) was carried out. The multiple sequence alignment of PF0864 with selected Lrp/AsnC family members was performed using the INDONESIA package (D. Madsen, P. Johansson, and G.J.

Kleywegt, unpublished data). The analysis on genomic level sequence alignment was generated from TIGR comprehensive microbial resources (http://cmr.tigr.org/tigr-scripts/CMR/shared/GenePage.cgi?locus=NTL01PF0864).

3.2.2 Expression and Purification of Recombinant PF0864

A modified pET24 plasmid with a 6 residue N-terminal His-tag was used as the expression vector. The *P. furiosus* gene was under control of a strong T7 promoter (Adams, Dailey et al. 2003). The clone (pETPF0864) was sequenced before expression and found to match the published sequence database exactly. Plasmid was transformed into *E.coli* BL21 (DE3) RIPL competent cells (Stratgene, Torrey Pines Road La Jolla, CA) for expression. Cells containing the recombinant PF0864 construct were grown in 5 ml LB medium containing 50μg/ml Kanamycin and 35μg/ml Chloramphenicol. After overnight incubation at 37°C with shaking at 200 rpm, the culture was used to inoculate 1 liter of LB medium. The 1L culture was incubated at 37°C on a shaker until the OD_{600} was 0.6. The culture was then induced by adding IPTG to a final concentration of 1mM, and incubated for 4 more hours at 37°C with shaking. Cells were harvested by centrifugation at 5,000 rpm for 15min. The pellet (~ 5 grams) was resuspended in 30 ml working buffer (25mM sodium phosphate and 100 mM NaCl at pH 7.6). 5mM PMSF and 2mM β–mercaptoethanol were added to the cell suspension to increase the lysis efficiency and decrease the protease activity. The suspension was then sonicated on ice using a Branson Sonifier Cell Disruptor 450 at power level 7 with 6 bursts of 30 seconds on and 30 seconds off. The resulting lysate was centrifuged at 12,000 rpm for 30 min. The supernatant was transferred to another centrifuge tube and incubated in a Fisher ISOTEMP 228 water bath at 65°C for 30 minutes. The supernatant was again collected by centrifugation at 12,000 rpm for 30 min, and then was put on ice ready for affinity chromatography.

A 5 ml HiTrap affinity column from Amersham Biosciences was charged with Ni^{2+} by first washing the matrix with 2 bed volumes of 1 M NaOH followed by 10 bed volumes of water, then 2 bed volumes of 0.5 M EDTA followed by 10 bed volumes of water, and charging the matrix with 2 bed volumes of 0.1 M $NiCl_2$ followed by 10 bed volumes of water wash, and finally equilibrated with 5 bed volumes of sample loading buffer, containing 25mM sodium phosphate, 100 mM NaCl, and 20mM Imidazole at pH 7.6. Using the AKTA Prime chromatography system from Amersham Biosciences, the supernatant was loaded onto the 5 ml Ni HiTrap column previously equilibrated. The column was washed with sample loading buffer until the UV absorption at 280nm wavelength was about zero, followed by elution with a linear gradient from 20 mM to 500 mM Imidazole in 100 ml. Fractions were collected and analyzed by SDS-PAGE. The recombinant protein obtained after Ni column was loaded onto HiLoad 16/60 Superdex 75 gel filtration column from Amersham Biosciences after concentration. Elution was carried out using 20mM HEPES and 100mM NaCl, pH 7.6, at 1 ml/min. Fractions were analyzed by SDS-PAGE. The mass of purified protein was determined by liquid chromatography mass spectrometry (LC-MS) at the University of Georgia, Department of Chemistry facility. Protein concentration was calculated based on extinction coefficient of 3840 M^{-1} cm^{-1} (Gill and von Hippel 1989) at 280nm absorption. The final purified PF0864 were 550µl at 5.6mg/mL concentration. Aliquots of 100 µl were stored at -80ºC and thawed immediately prior to crystallization trials.

3.2.3 DNA-binding by PF0864

The DNA fragment H1 and H2 were designed using sequence information from genome sequence of *P. furiosus* (NCBI) between nucleotides 839221-840301. The DNA fragments used as bait in the DNA binding experiments were amplified from genomic DNA generously provided

by Dr. Frank Jenny. The following primers, H1 Forward-AAAATTTTAGATATTTT GGGGATTCC, H1 Reverse–ATCCTTTTTTAATAACCTTAGTATGTTC; H2 Forward-TATTCTTTGCTAGCTCTACC, and H2 Reverse-TATATAGCCCCACTTCG were ordered from Integrated DNA Technologies (IDT). Polymerase Chain Reaction (PCR) was carried out using a Peltier Thermal Cycler (MJ Research) to amplify H1 (120bp) and H2 (207bp) using their primers designed as above. The amplification protocol was as follows: incubation for 3 minutes at 94°C, followed by 30 cycles of the following: 30 seconds at 94°C, 30 seconds at 55°C, and 60 seconds at 68°C, and finished with 10 minutes at 72°C. The PCR product was visualized by electrophoresis on a 2% agarose gel using TAE (Tris-Acetate- EDTA) buffer and ethidium bromide, followed by purification using a Qiagen PCR purification kit. PF0864 was in HEPES Buffer pH 7.3 with 200mM KCl, 2.5mM $MgCl_2$, 5% PEG-400, and 1mM EDTA for the DNA binding experiments. All binding experiments were carried out at 37°C for 30 minutes. Firstly, H1 and H2 were mixed with excess protein sample separately. Then modified gradient ratios of protein-DNA fragment mixture were performed for the gel shifting. The DNA-protein mixture was then loaded onto a native 4-20% gradient Criterion Tris- HCl Acrylamide Gel purchased from BioRad. The gel was then stained with ethidium bromide to detect DNA (BioRad Hercules, CA). After observation using the UV light, the gel was then stained with coomassie blue to detect the protein.

3.2.4 Crystallization

The initial crystal screening of PF0864 was carried out using the Honey bee robot from Cartesian Technologies with sitting-drop vapor diffusion method. 384 crystallization conditions from Crystal Screen, Crystal Screen 2, MembFac, PEG/ION Screen, and Crystal Screen Cryo from Hampton Research, Wizard I and II from Emerald BioSystems, and MemSys from

Molecular Dimensions were used. Screens were set up by adding 100 µl of crystallization reagents into the reservoirs of Greiner Crystal Quick 96-well, 3-drop plate, mixing 200 nl of protein solution with 200 nl of the crystallization reagent, covering the whole tray with the ClearSeal Film (Hampton Research, Aliso Viejo, CA), and incubating at 18 ºC.

The crystallization optimization was performed by the modified microbatch under oil method (Baldock, Mills et al. 1996) using the Oryx 6 crystallization robot from Douglas Instruments. 0.5 µl of protein solution with 0.5 µl of the crystallization reagent were mixed, drops were covered with mixed oil containing 70% of Paraffin and 30% Silicone oil, and the crystallization tray was incubated at 18 °C.

The crystals were grown in condition 100mM NaAc/HCl pH 4.5 and 16% w/v PEG4000 in the optimization at 18 °C after 3-4 days with size of 150×100×100µm (Figure 3.1).

3.2.5 Heavy Atom Soaking

Crystals with good diffraction quality were used for the heavy atom soaking screening. The heavy atom compounds were all from Hampton Research Heavy Atom Screen kits. The soaking procedure was performed by placing heavy atom compound powder at least 10 times smaller than the crystal size in the crystallization well with a needle. Total of 15 heavy atom compounds were used. They are number 1, 9, and 16 from Heavy Atom Screen M1, number 2, 8, and 18 from Heavy Atom Screen M2, number 1 and 2 from Heavy Atom Screen Au, number 1, 14, and 20 from Heavy Atom Screen Hg, and number 1, 2, and 5 from Heavy Atom Screen Pt. The powder soaked crystallization trays were kept at 18 ºC for different time period: 1 hour, 4 hours, and overnight before the crystal harvesting.

3.2.6 Data Collection

After heavy atom soaking, the crystals were harvested from the drop using 18mm Hampton CrystalCap® Copper magnetic pins with the loop sizes chosen to best match the crystals' dimensions. Once harvested, the crystal was briefly (1-2 sec) immersed in a small drop (~2µl) of its well solution containing 30% glycerol as a cryoprotectant. The cryo-protected crystal was recovered from the cryoprotectant drop using the same Hampton pin, flash-froze, and stored in liquid nitrogen.

The final structure determination data set was collected at the Southeast Regional Collaborative Access Team (SER-CAT), beamline 22ID, Advance Photon Source (APS), Argonne National Laboratory. The wavelength of the X-ray was 0.9724 Å (12758ev), and the single SAS dataset was collected on a MAR300 detector. The crystal was rotated 180 degrees in the beam at 1 degree oscillation.

3.2.7 Structure Determination

Data were indexed, integrated, and scaled using the HKL2000 software suite (Otwinowski and Minor 1997); the statistics are in Table 3.1. Phasing of PF0864 was done through the SECSG web-based structure solution pipelines (Liu, Lin et al. 2005). The pipeline uses an array of separate programs to screen parameter space to achieve the best solution. The program package SOLVE/RESOLVE (Terwilliger and Berendzen 1999; Terwilliger 2000; Terwilliger 2002) was used to identify and refine the heavy atom positions and calculate the initial electron density map. The ARP/wARP software traced the initial model (Perrakis, Morris et al. 1999). The Matthew coefficient calculation showed the dimer formation in the asymmetric unit, then the experimental phases were improved using non-crystallographic symmetry averaging with DM (Cowtan and Zhang 1999). The ARP/wARP traced about 40% of the model

automatically. The final model was completed using Coot (Emsley and Cowtan 2004), TLS (translation/libration/screw) motion determination was carried out using the websever, http://skuld.bmsc.washington edu/~tlsmd/, and TLS restraint refinement (Winn, Isupov et al. 2001) was performed using REFMAC5 (Murshudov, Vagin et al. 1997). NCS restraints were employed till the last stage of the refinement. The structure validation was performed using MOLPROBITY (Lovell, Davis et al. 2003), then it was deposited into PDB bank, ID 2IA0.

3.2.8 Structural Comparison and Alignment

The structures of PF0864 was compared with *P. furiosus* LrpA (1I1G) (Leonard, Smits et al. 2001), *P. horikoshii shinkaj OT3* FL11(1RI7) (Koike, Ishijima et al. 2004), *Escherichia. coli* AsnC (2CG4), and *Bacsillus.subtilis* LrpC (2CFX) (Thaw, Sedelnikova et al. 2006). The three dimensional structure alignment and structural based sequence alignment was performed using the INDONESIA package (D. Madsen, P. Johansson, and G.J. Kleywegt, unpublished data).

3.3 Results and Discussion

3.3.1 Sequence Analysis

The PSI, PHI-BLAST search using the PF0864 sequence as a query gave the result as shown in Figure 3.2a. The result is consistent with the TIGR annotation (http://cmr.tigr.org/tigr-scripts/CMR/Cmr-HomePage.cgi) for PF0864 as a putative transcriptional regulator belonging to Lrp/AsnC family. It has an N-terminal helix-turn-helix (HtH) DNA-binding domain, which shares a high degree of similarity with HtH motifs of bacterial transcriptional regulators, such as the nitrogen assimilation regulatory proteins (NtrC) from species like Azobacter, Rhodobacter, and Rhizobium. PF0864 displays weak homology to a number of protein families involved in transcriptional regulation: Lrp/AsnC regulators (amino acid metabolism), ArsR (stress response to heavy metals), and MarR (antibiotic resistance, oxidative stress, response to aromatic

compounds). The sequence search against PDB database showed two available structures with limited sequence identities (Figure 3.2b). The proteins from *P. horikoshii shinkaj OT3* FL11(1RI7) (Koike, Ishijima et al. 2004) and *P. furiosus* LrpA (1I1G) (Leonard, Smits et al. 2001) share 32% sequence identity among 140 residues and 39 % sequence identity among 64 residues respectively with PF0864 (162 amino acids). They were both used as models for structure determination of PF0864 using molecular replacement methods.

In *P. furiosus*, total of 13 Lrp/AsnC familiy members have been found based on the genome analysis from TIGR-CMR. They are PF0054, PF0113, PF0250, PF0739, PF0864, PF1022, PF1231, PF1543, PF1601, PF1732, PF1734, PF1893, and PF2053. PF1893 and PF1732 have only the C-terminal domain, which implies that they do not bind DNA like their counterparts do, and their role is remain unclear. The structure solved previously from *P. furiosus* LrpA (1I1G) (Leonard, Smits et al. 2001) is PF1601, which has 26.8% full length sequence identity with PF0864.

3.3.2 DNA-binding Gel Shifting

PF0864 as a member in the Lrp/AsnC family has the N-terminal DNA-binding domain. The Lrp/AsnC family member in *P. furiosus* was proposed to have the auto-regulation function of its own gene (Brinkman, Dahlke et al. 2000). Using this information, two DNA fragments were designed. H1 fragment contains 60bp from PF0864 gene and 60bp upstream. The H2 fragment contains 44bp from PF0865 gene and 163bp upstream. Both H1 and H2 are the putative promoter regions for the auto-regulation of PF0864. The control DNA fragment was picked from non-related PCR product, its sequence had no similarity with either H1 or H2. Figure 3.3a shows all of H2 were shifted with excess PF0864, but the control DNA and H1 did not show any movement under the same conditions. In the gradient protein-H2 fragment binding

experiment as shown in Figure 3.3b, the molar ratios of H2:PF0864 are 1:0.1, 1:0.3, 1:1, and 1:3 in lanes 1 to 4. The NativMark ^{TM} (Invitrogen) allows accurate molecular weight estimation of proteins using native gel electrophoresis with Tris-Glycine gels. It was used to detect the protein molecular weight in the DNA-binding state. The H2-PF0864 complex is around 242 KDa position. The molecular weight of the H2 fragment is calculated to be around 127 KDa, and the mass of His-tagged PF0864 is 19362Da from mass spectrometry result, so it seems that PF0864 form at least a hexamer to bind the H2 fragment. Since it is believed that the Lrp/AsnC family is functional in the dimer form (Ettema, Brinkman et al. 2002), there should be three recognition sites with non-strict palindrome repeats on H2 fragment. Although more experimental data is needed to clarify this assumption, the binding of PF0864 to H2 fragment is specific, and it may regulate the expression level of the whole operon containing PF0865-PF0864-PF0863.

3.3.3 Structure Determination

Solving the phase problem is one of the significant bottlenecks on the way to the final structure. At first, molecular replacement method, the extensively used high-throughput structural solving method was considered to solve the structure of PF0864, but it failed. The reason may be that the best model available at that time, PDB ID: 1RI7, the Lrp/AsnC family member from *P. horikoshii OT3* (PH1519), had only 32% identity in 140 residues with PF0864. Usually accurate structure models can only be built for sequences that are at least 35% identical (Martin, MacArthur et al. 1997). Then the direct crystallography using sulfur phasing was excluded because there is no cystine or methionine in the protein sequence except the first methionine at the N-terminus. So, although screening for heavy-atom derivatives is a time-consuming and cumbersome process, it became the only way to get the phase information of PF0864.

From many heavy atom soaked crystals using different heavy atom compounds at different soaking time period, only one crystal with Gold (I) Potassium Cyanide ($KAu(CN)_2$) soaking for 4 hours gave the anomalous data set to get the final structure. Two Au sites were found, refined, and used to calculate the initial electron density map.

During the refinement stage, the TLS refinement helped dramatically to lower the R-free value. TLS motion determination (TLSmd) analyzes a protein crystal structure for evidence of flexibility, such as local or inter-domain motions, and then divides the protein chains into multiple segments that are modeled as rigid bodies. The output file from TLSmd contains the TLS groups and values of T, L, and S tensors, which is used as an input file to run the TLS restraint refinement in REFMAC. For PF0864 structure, since there are clearly two domains at N- and C- terminals connected with the hinge region, the TLS refinement put in the flexibility which helped lower the R-free values.

3.3.4 Structure Description

Two PF0864 molecules form a homodimer in one asymmetric unit (Figure 3.4). The results from analytical gel filtration chromatography also showed PF0864 as a dimer in solution. Residues 5-162 in both molecules A and B were clearly defined in the electron density map. The N-terminal His-tag was not visible in either molecule. Chain A and chain B are equivalent with the RMSD of 0.534 Å. Each monomer of PF0864 structure has N-terminal DNA-binding domain consisting α1 to α3. From residues 49-66, one β strand and loop region connect the N-terminal domain to the C-terminal domain. The C-terminal domain is an effector-binding domain, which has been termed RAM domain. It contains β2α4β3β4α5β5β6α6, where number 2 to 5 β strands form the four stranded anti-parallel pattern. The β6 is anti-parallel with the β3 from the other monomer. After β6, the α6 goes back up to the connection region between N- and

C- terminal domain, right under the β1 and above the whole C-terminal domain. The βαββαβ fold has significant similarity with the ACT domain, an ever-present allosteric regulatory domain in many metabolic enzymes, but the effector-binding sites between them are different. Also, the RAM domain is mainly associated with transcriptional regulators, while the ACT domain is usually found as a regulatory module of metabolic enzyme (Ettema, Brinkman et al. 2002). The proteins used for studying phyletic distribution of RAM domain were mainly Lrp-like regulators. The typical HtH-RAM conformation in Lrp-like regulators was found in some bacteria (*Streptomyces sp.*), where the duplicated HtH-RAM exists. This analysis results support the view that native form of Lrp-like transcriptional regulator is at least a dimer configuration (Ettema, Brinkman et al. 2002).

The dimer formation representing the functional units is stabilized by the hydrophobic core between β2 to β5 strands of each monomer. In addition, the main chain hydrogen bonding between the anti-paralleled β6 and β3 of the other monomer contributes to stabilize the dimer formation. Further stabilization is achieved in the hydrogen bonding interaction between β1 of each monomer. They come together forming a 2-stranded anti-parallel β ribbon. At the C-terminal, α6 from each monomer fills the empty area between the β ribbon and C-terminal RAM domain.

3.3.5 Structural Comparison with the Homologs

At the final refinement stage of PF0864 structure before the deposition into PDB, three-dimensional homology search using DALI (Holm and Sander 1993) found two more new structures from Lrp/AsnC family: *E. coli* AsnC (2CG4) and *B.subtilis* LrpC (2CFX) (Thaw, Sedelnikova et al. 2006). The three-dimensional structure alignment (Figure 3.5) and the structural-based sequence alignment of PF0864 with *P. furiosus* LrpA (1I1G) (Leonard, Smits et

al. 2001), *P. horikoshii shinkaj OT3* FL11(1RI7) (Koike, Ishijima et al. 2004), *E. coli* AsnC (2CG4), and *B.subtilis* LrpC (2CFX) (Thaw, Sedelnikova et al. 2006) (Figure 3.6) are performed. They are all very similar in three-dimensional, whereas the sequence identities are relatively low. The N-terminal DNA-binding domain shows higher sequence similarity than the C-terminal effector-binding domain. Residues D7, D10, I13, and L17 on α1, E27 on α2, R42 on α3, G50, and I52 from PF0864 are all identical among all the aligned sequences. Those identical residues are all on the HtH fold, which is their DNA binding region. An alignment gap exists between β2 and α4 between the previously solved structures and PF0864. In our structure, a longer loop exists in this area, which unfortunately is disordered in the crystal structure. After this region, the sequence similarity becomes faint in the C-terminal domain, where different effectors may bind. The overlap of C-terminal region shows that the PF0864 anti-parallel β strands give smoother look than the other structures. At the C-terminal, all the previously solved structures have a long loop region going back up to the N-terminal region. But in PF0864 structure, one α-helix turns almost 90 degree and sticks into the middle of the dimer, which gives a more compact feature of the overall structure comparing with the other structures.

The high structural similarity combined with the relatively lower sequence identity suggests that the members of Lrp/AsnC family may undergo different activation mechanisms to regulate the expression of various genes.

3.3.6 Structure-function Relationship

The helix-turn-helix domain can be found in both basal and specific transcription factors. It contains an open right-handed tri-helical bundle, where the 3rd helix forms the major DNA-protein interface. The 3rd helix inserts itself into the major groove of the DNA to form the close contacts (Brennan and Matthews 1989). However, the individual residues involved in DNA

contacts may widely vary across the fold. There are many different variants of the HtH domains (Aravind, Anantharaman et al. 2005); PF0864 has the simple tri-helical HtH, in which α1&α2 and α2&α3 form a two helix motif (HtH). The sequence alignment in this region combining the mutant data from *E.coli* Lrp shows T38, H40, and R42 on third helix of PF0864 have direct involvement in DNA-binding. Although the sequence is not same, the folding position is identical. That may explain the various DNA recognition sequences by the family members. Until now, no DNA-binding structure has been determined yet. The LrpA-DNA model (Leonard, Smits et al. 2001) shows straight piece of B-form DNA binding to the LrpA dimer. In FL11 cylinder (Koike, Ishijima et al. 2004) and recently published *E.coli* AsnC and *B. subtilis* LrpC (Thaw, Sedelnikova et al. 2006), they model the dodecameric and octameric formation of the protein structure binding to the curved B-form DNA fragment. Since the exact nature of the interaction of those proteins with their target DNA is still not clear, and the binding promoter region lack obvious inverted repeat elements, the DNA binding model can only give us the possible gene regulation mode. Each Lrp/AsnC family members may have its own regulatory activities under different environmental changes (Yokoyama, Ishijima et al. 2006).

The DNA-binding and gene expression regulation by Lrp/AsnC family members is working in the absence or presence of effectors. The effector binding at C-terminal RAM domain may cause modulation of DNA affinity, DNA bending, Lrp-like protein oligomeration, and Lrp-teritiary structure (Ettema, Brinkman et al. 2002). Those conformation changes may or may not be suitable for the DNA-binding. Only in the *E.coli* AsnC structure, an asparagine exists in the cleft between the turn from strand β3 to β4 of one monomer and strand β5 of another (Thaw, Sedelnikova et al. 2006). In PF0864 structure, no extra density has been found in the C-terminal effector binding domain.

3.4 Conclusion

PF0864 is annotated as transcriptional regulator from Lrp/AsnC family in TIGR. The sequence identities of PF0864 to this family are relatively low, but the three-dimensional shows the same folds with the other characterized family members. The crystal structure of PF0864 shows N-terminal HtH binding motif involving in the DNA-binding and C-terminal RAM domain possibly involving in the effector binding. The DNA-binding essay shows PF0864 can specifically bind to the DNA fragment upstream of the whole operon containing genes PF0865, PF0864, and PF0863. It is certain that PF0864 is a DNA-binding protein belonging to the Lrp/AsnC family, which may autoregulate its own gene expression and members in the same operon under environmental changes. But the exact binding sequence of DNA and how it functions as a transcriptional regulator *in vivo* need more study.

PF0864 is not the high priority target in the structural genomics project at SECSG because the homolog structure has been solved before our work. However, by reviewing the ultimate goal of structural genomics, each protein target should be valued of its three-dimensional structure information. The keep accumulating structural data and the comparison among each structure from different organisms are providing more understanding to each protein family. The phasing determination of PF0864 structure used the traditional heavy atom soaking to introduce heavier atom for solving the phase problem in crystallography method. Even through nowadays high-throughput structural solving methods are powerful and convenient, the regress to used methods are necessary and supplementary

Based on the general eukaryotic-like and muti-component nature of transcriptional machinery observed in Archaea (Bell and Jackson 1998), it is suggested that the Lrp/AsnC family members may interact with other proteins to form protein-protein complexes during the

transcriptional regulation via the RAM domain (Ettema, Brinkman et al. 2002) in the C-terminal domain of each available family member, including PF0864.

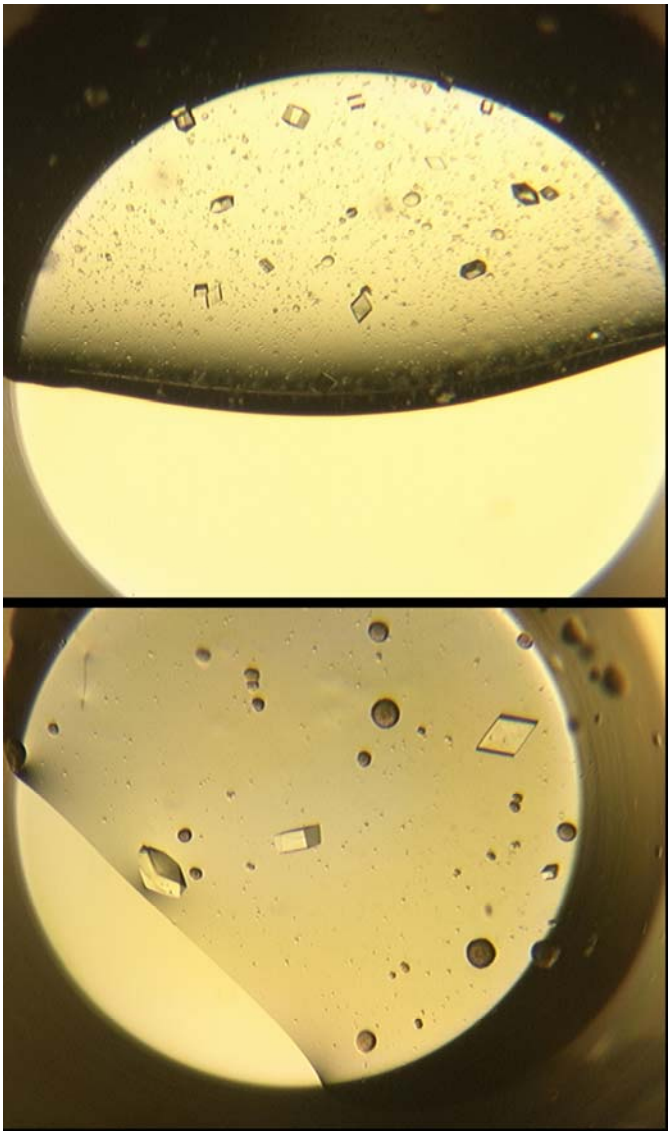
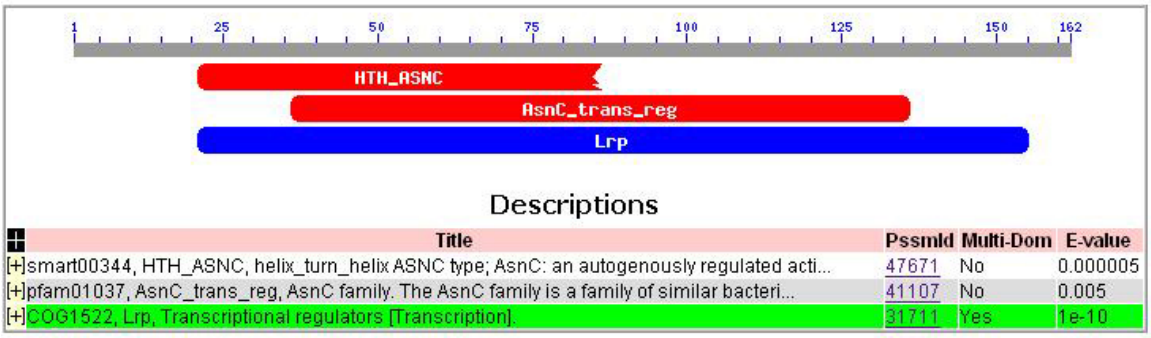


Figure 3.1: Crystals of PF0864.

The top picture was taken from the initial crystal screen using MF-5 condition. The bottom picture was taken from the optimization well using 100mM NaAc/HCl pH 4.5 and 16% w/v PEG4000.



a

```
>gi|18655415|pdb|1I1G|A Chain A, Crystal Structure Of The Lrp-Like Transcriptional Regu
From The Archaeon Pyrococcus Furiosus
 gi|18655416|pdb|1I1G|B Chain B, Crystal Structure Of The Lrp-Like Transcriptional Regulat
From The Archaeon Pyrococcus Furiosus
Length=141

 Score = 45.4 bits (106),  Expect = 6e-06, Method: Composition-based stats.
 Identities = 25/64 (39%), Positives = 43/64 (67%), Gaps = 4/64 (6%)

Query  6   LDDLDRNILRLLKKDARLTISELSEQLKKPESTIHFRIKKLQERGVIERYTIILGEQLKP  65
           +D+ D+ IL +L+KDAR   +E++++L   E+ +  R+K L+E+G+IE YTI    ++ P
Sbjct  2   IDERDKIILEILEKDARTPFTEIAKKLGISETAVRKRVKALEEKGIIEGYTI----KINP  57

Query  66  KHLA  69
           K L
Sbjct  58  KKLG  61


>gi|47168788|pdb|1RI7|A Chain A, Crystal Structure Of A Protein In The LrpASNC FAMILY
FROM The Hyperthermophilic Archaeon Pyrococcus Sp. Ot3
Length=171

 Score = 55.5 bits (132),  Expect = 6e-09, Method: Composition-based stats.
 Identities = 45/140 (32%), Positives = 82/140 (58%), Gaps = 15/140 (10%)

Query  4    IHLDDLDRNILRLLKKDARLTISELSEQLKKPESTIHFRIKKLQERGVIERYTIILG-EQ  62
            + LD++D+ I+++L+ D +  + E+S+      ESTIH RI+KL+E GVI+++T I+  E
Sbjct  23   VPLDEIDKKIIKILQNDGKAPLREISKITGLAESTIHERIRKLRESGVIKKFTAIIDPEA  82

Query  63   LKPKHLALIVLEV--GKPVIEDFLERYISYISSTLSALPGVLFVAK-SGEDKIIALVGKN  119
            L    LA I+++V  GK           S ++S L+  P ++ V + +G+  ++  +  +
Sbjct  83   LGYSMLAFILVKVKAGK----------YSEVASNLAKYPEIVEVYETTGDYDMVVKIRTK  132

Query  120  NKDELVKFIEENITSIPNLK  139
            N +EL  F++  I SIP ++
Sbjct  133  NSEELNNFLDL-IGSIPGVE  151
```

b

Figure 3.2: The PHI, PSI-Blast result using PF0864 sequence as a query.

a: As a member of Lrp/AsnC transcriptional regulator family, in PF0864 sequence, there is an HtH motif at the N-terminal domain involving in the DNA-binding activity.

b: Two homologs structures were found using PF0864 sequence to search against PDB database.

Table 3.1: Statistics of PF0864 from the crystallographic analysis

Crystal

Space group:	$P2_12_12_1$
A	58.074Å
B	70.372Å
C	96.536Å
$\alpha = \beta = \gamma$	90º

Data processing statistics

Wavelength (Å)	0.9724
Resolution (outer shell, Å)	50- 2.4 (2.49-2.40)
Completeness (%)	91.68
R_{sym}	0.114 (0.477)
Redundancy	6.1 (3.5)

Refinement statistics

Resolution range (Å)	56.89-2.37 (2.43-2.37)
Reflections used (free)	14545 (776)
R-factor (R_{free}, %)	19.1(26.8)
Mean B factor (Å2)	20.00
RMSD bond lengths (Å)	0.023
RMSD bond angles (°)	2.248

Final model

Residues	5-162 (Chain A); 5-162 (Chain B)
Protein atoms (solvent)	2436 (101)
PDB ID	2IA0

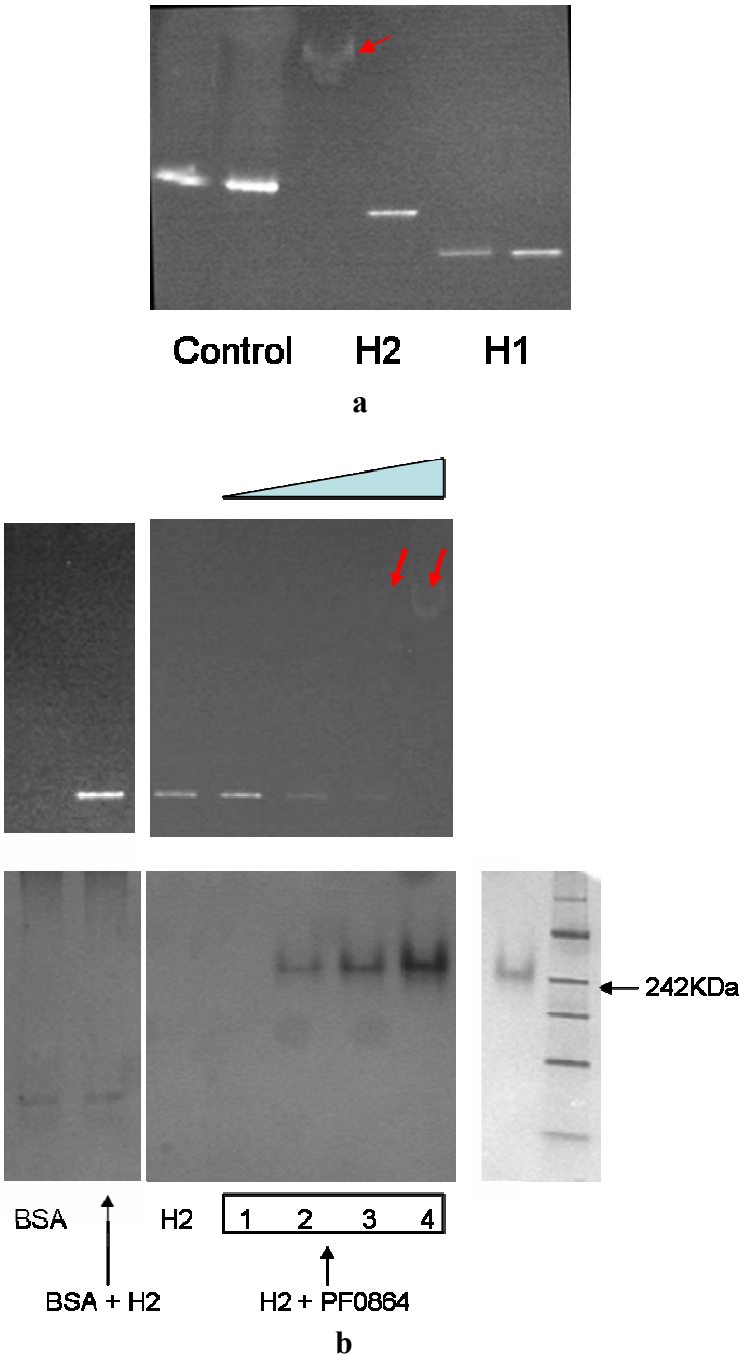


Figure 3.3: The DNA binding gel shifting.

a. The control lanes use unrelated DNA fragment. The H1 and H2 are designed DNA fragments. In each pair, the left lane is with the PF0864, and the right lane is without PF0864

b. The gradient DNA gel shifting using the H2:PF0864 in molar ratios 1:0.1, 1:0.3, 1:1, and 1:3 from lane 1 to 4. The upper gel was stained in ethidium bromide to detect the DNA. The lower gel was stained in coomassie blue to detect the protein.

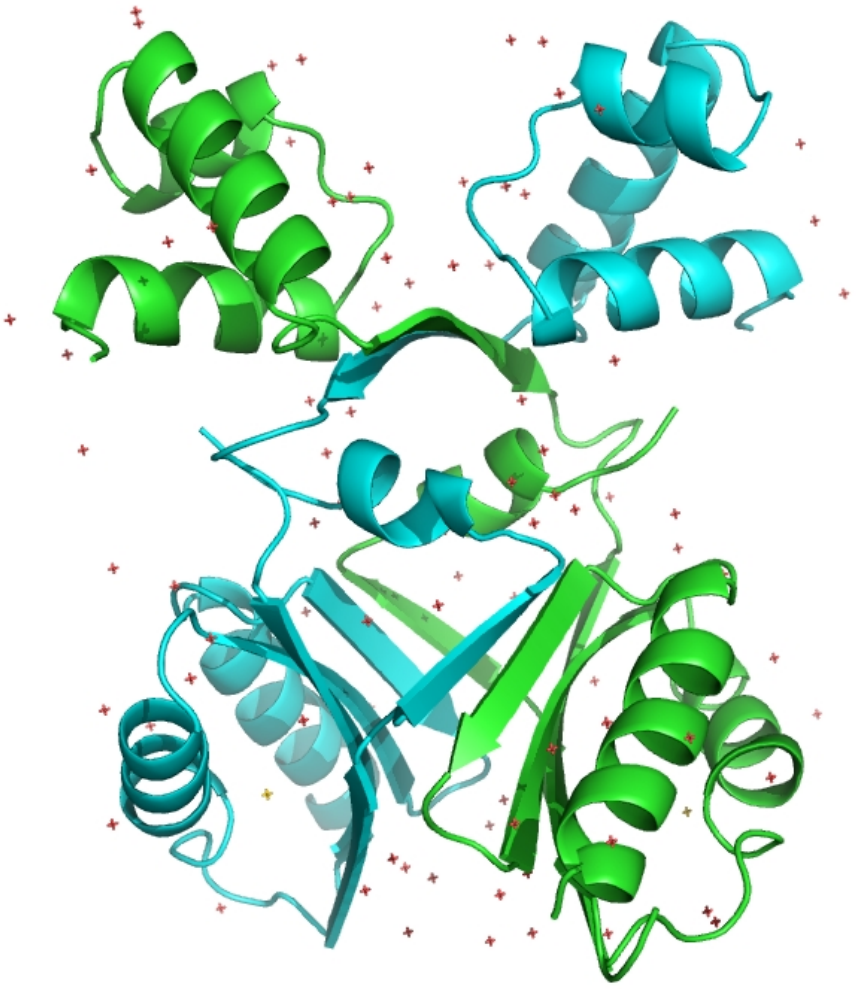


Figure 3.4: Structure of PF0864.

The cartoon diagram generated with PyMOL (DeLano 2002). Chain A and chain B are colored green and blue, respectively. The red cross marks represent the water molecules. The protein has two domains: the N-termianl is HtH-DNA binding domain, and the C-terminal is RAM domain.

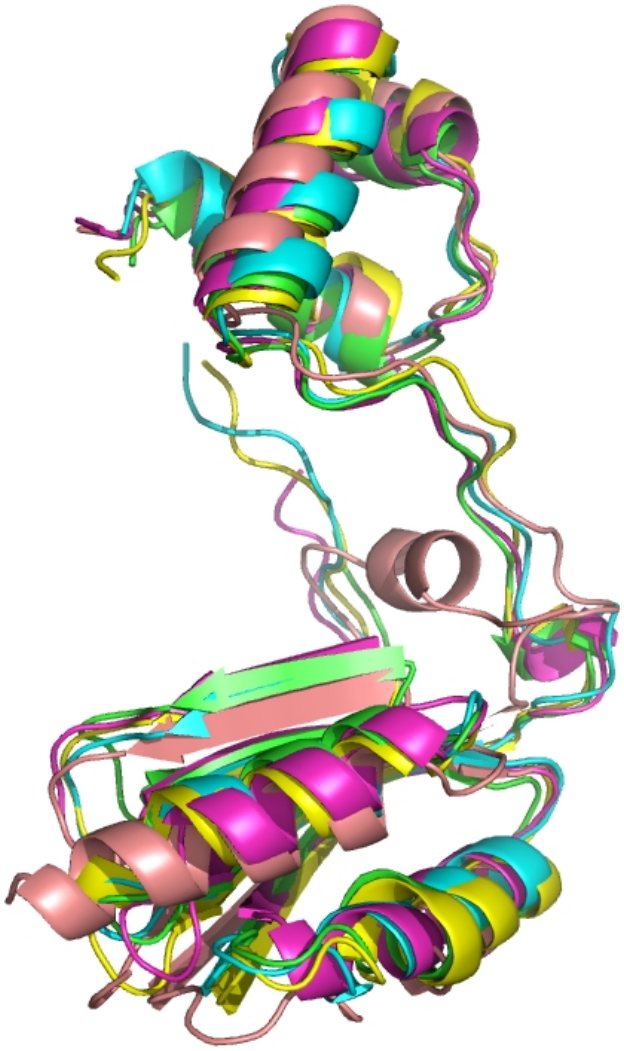


Figure 3.5: The three-dimensional structure alignment among the Lrp/AsnC family members.

The three-dimensional structures are aligned using INDONESIA package. The cartoon figure is made by Pymol (DeLano 2002). The green represents the *P. furiosus* LrpA (1I1G) (Leonard, Smits et al. 2001), the cyan represents *P. horikoshii shinkaj OT3* FL11(1RI7) (Koike, Ishijima et al. 2004), the yellow represents *E. coli* AsnC (2CG4), the magenta represents *B.subtilis* LrpC (2CFX) (Thaw, Sedelnikova et al. 2006), and the salmon represents PF0864.

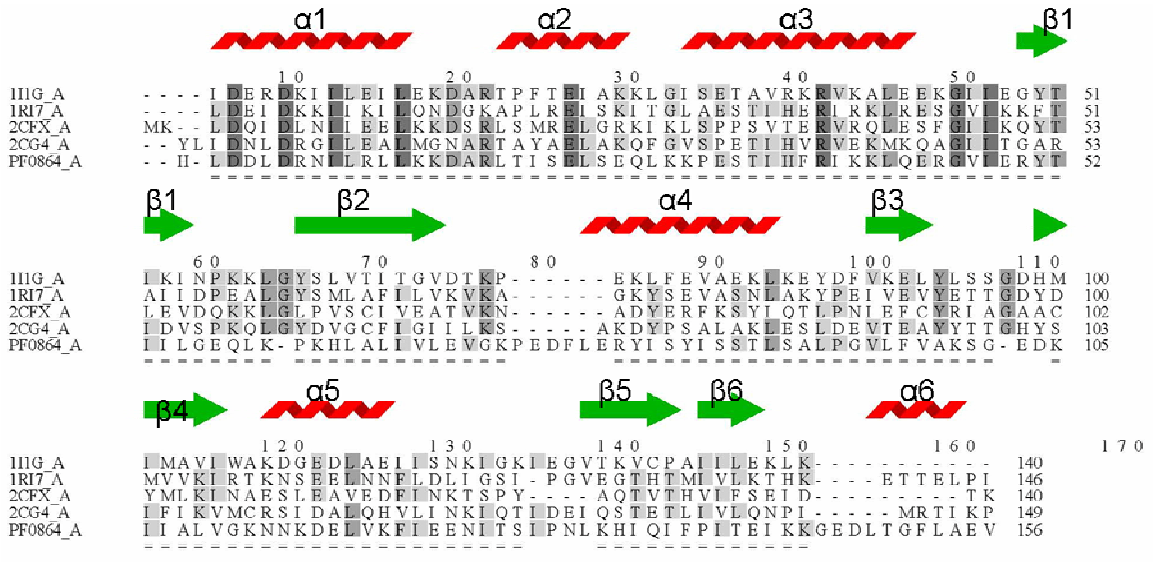


Figure 3.6: The structural-based sequence alignment.

The structures of PF0864 was compared with *P. furiosus* LrpA (1I1G) (Leonard, Smits et al. 2001), *P. horikoshii shinkaj OT3* FL11(1RI7) (Koike, Ishijima et al. 2004), *E. coli* AsnC (2CG4), and *B.subtilis* LrpC (2CFX) (Thaw, Sedelnikova et al. 2006). The secondary structure is shown on top of the sequences. The N-terminal has tri-helical bundle, and the C-terminal has the βαββαβ fold. The aligned residues are shown in different shades of Grey. (Made by INDONESIA package)

Chapter 4

The study of PF0863-PF0864

4.1 Introduction

Bacterial genes with related functions are often transcribed simultaneously from the same operon. In this study, PF0863 and PF0864 form a contiguous gene pair in a conserved gene cluster. Individual structures of these two encoded proteins have been solved successfully. Also the structural and functional analyses of each protein have been discussed in the previous two chapters. Elucidation of their possible interaction and related functions becomes an immediate interest in this work.

Protein-protein interactions are defined as the specific interplays between two or more proteins. In a living cell, non-covalent interactions among proteins are constantly forming and dissociating. Therefore, it has been understood that all proteins in a living cell are connected in a huge network (Schwikowski, Uetz et al. 2000). Within such a network, protein-protein interactions are involved in many important cellular processes such as signal transduction, transport, cellular motion, and most regulatory mechanisms (Alberts 2002). Finding and understanding those interactions is a major goal of functional genomics and proteomics (Legrain, Wojcik et al. 2001).

In recent years, many experimental techniques have become available for discovering protein-protein interaction networks of several organisms. The yeast two-hybrid screen (Fields and Song 1989) is one of the *in vivo* methods of detecting protein-protein interactions. It can provide the first hint for identification of interacting protein partners. The main weakness of the

two-hybrid screen is the high rate of false positive indications (Deane, Salwinski et al. 2002). The combination of large-scale affinity purification with mass spectrometry (MS) , *in vitro*, can detect and characterize multiprotein complexes (Sobott and Robinson 2002). However, the accuracy and coverage are limited too in this approach. Due to their inherent inconsistency, results from the experimental methods are usually consolidated by the deduction from the computational methods. The computational methods can be used to infer protein-protein interactions, to design and validate experimental studies, and also to predict detailed structures of protein complexes of interaction partners (Szilagyi, Grimm et al. 2005).

Computational prediction methods can be divided into two categories: the structural-independent methods and the structural-dependent methods (Galperin and Koonin 2000; Huynen, Snel et al. 2003; Russell, Alber et al. 2004). The structure-independent methods build on the known biological knowledge to predict protein-protein interactions. The comparative analysis of genomes can provide the information on the conservation of different types of genomic context, which is useful for prediction of functional interactions between gene products (Galperin and Koonin 2000). The co-evolution of interacting proteins in a coordinated way reveals the physical interactions among them (Pazos, Olmea et al. 1997). The structure-based prediction methods include (1) modeling protein-protein interactions by similar, known structures of protein complexes (Sali and Blundell 1993), (2) a threading-based method (Lu, Lu et al. 2002) and (3) the widely used protein-protein docking (Sternberg, Gabb et al. 1998). All those approaches look at the structural details of the putative interaction and use experimentally determined or even predicted structures. The relationship between the structure-dependent methods and the structure-independent methods is mutually supplementary.

The individual structures and structure-function relationship analysis of PF0863 and PF0864 have been discussed in Chapter 2 and Chapter 3, respectively. They both have regulatory functions. PF0863 shows the activities and structural similarity of a nucleotidase. PF0864 is a member from Lrp/AsnC transcriptional regulator family having the specific DNA-binding activity to its own putative promoter region. The whole genome microarray data (Schut, Zhou et al. 2001; Schut, Brehm et al. 2003; Weinberg, Schut et al. 2005) of *P. furiosus* indicate that all three ORFs (PF0863-PF0865) in this predicted operon are co-down regulated under stress conditions such as cold, -Fe, and peroxide shock. Combining this information, and on the basis of their gene analysis and structural features, it can be assumed that PF0863 and PF0864 form a complex.

Computational docking methods are first performed to determine whether the interaction between PF0863 and PF0864 is possible from the viewpoint of structure-dependent methods. The prediction is tested experimentally using size-exclusion chromatography followed by native-PAGE of the individually purified proteins and the mixture. In addition, re-cloning of PF0863 gene with glutathione-S-transferase (GST)-tagged fusion is performed. Co-expression of PF0864 and GST-PF0863 for the GST pull-down assay are followed to investigate their possible interactions.

4.2 Material and Methods

4.2.1 Two Genes Analysis

The individual sequence analyses of PF0863 and PF0864 have been discussed in Chapter 2 and Chapter 3, respectively. The homologs of PF0864-PF0863 cluster in the same operon can be found in other organisms that belong to *Thermococci* class. Their genome organization comparison generated from TIGR-CMR is shown in Figure 4.1.

4.2.2 Docking

The protein-protein docking predicts the structure of a multimeric protein complex from two or more separately determined protein structures (Szilagyi, Grimm et al. 2005). Many docking programs are developed by different research groups (Smith and Sternberg 2002), but all of them rely on the same assumption that interacting proteins have a certain degree of shape complementarities. The general docking procedure is shown in Figure 4.2.

3D-Dock suite was downloaded (free for academic users) from http://www.bmm.icnet.uk/docking/download.html and installed in a Linux environment. FTDock (Fourier Transform Dock) was run first to obtain a prediction of the binding geometry of two molecules by performing rigid-body docking (Gabb, Jackson et al. 1997). The result from FTDock was then scored with RPScore (Residue level Pair potential Score) using a empirically-derived, single distance constraint pair potential (Moont, Gabb et al. 1999). The top ten complex models are shown in Figure 4.3.

The web-based server of ZDock (http://zdock.bu.edu/) and GRAMM (http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/) were also used as supplementary methods to predict the possible interactions between PF0863 and PF0864. The model from GRAMM is presented in Figure 4.4.

4.2.3 Gel Filtration Chromatography of Two-protein Mixture and Native Gel Analysis

Individually purified PF0863 and PF0864 were mixed using a 1:1 molar ratio. Superdex 200 10/300 GL column (GE Healthcare) was calibrated using gel filtration standard from Bio-Rad (catalog # 151-1901) before loading the sample mixture. After 2 bed volumes (2X24ml) of equilibration using the buffer 20mM HEPES, pH 7.6 with 200mM NaCl at 0.2ml/min, 50µl of PF0863 and PF0864 mixture was injected onto the column, and then eluted for one bed volume

with the same speed. The chromatograms from AKTA Prime chromatography system were shown in Figure 4.5. The peak fractions were analyzed with two 4-20% Tris-HCl gels (Bio-Rad) for native PAGE and SDS-PAGE (Figure 4.6).

4.2.4 Gateway Cloning of PF0863

DNA sequences of PF0863 was cloned using the Gateway® Cloning Technology based on specific recombination between homologous DNA sequences (Invitrogen) (Hartley, Temple et al. 2000). The pETPF0863 plasmid was used as the template for gene amplification by the polymerase chain reaction (PCR). Primers were designed using XPression Primer 3.0 software (Forward-GAAAACCTGTACTTCCAAGGCGGGTCAGGTATGGAAGTTGAAATAAAGTT TAAGATTAAG, Reverse-GGGGACCACTTTGTACAAGAAAGCTGGGTTCATGAAGAG CGTCCAGATAAC, and universal-GGGGACAACTTTGTACAAAAAAGTTGGCGAAAACC TGTACTTCCAAGGC). A Tobacco Etch Virus (TEV) protease cleavage site (underlined) followed by a spacer region (GGGTCAGGT) was designed in front of PF0863 gene to remove the affinity and the Gateway® tag during protein purification. For PCR accuracy, the high fidelity and specificity AccuPrime Pfx DNA polymerase (Invitrogen) (Takagi, Nishioka et al. 1997) was used. The concentration of each primer was diluted to 5μM and 100ng plasmid was used in the PCR mixture. Two-step PCR was performed: only the forward and reverse primers were used in the first 5 cycles, then the universal primer was added to continue for 30 more cycles. The following temperature protocol was applied during the reaction: 95°C for 3 min; then 5 (30 after universal primer was added) cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 60 s; and 72°C incubation for 10 minutes. In Gateway® cloning, the entry clone was created by BP recombination reaction using Gateway® BP clonase enzyme mix (Invitrogen). The reaction mixture containing 1μl PCR product (168μg/μl), 1μl pDONR-221 entry vector (150μg/μl)

(Invitrogen), 2µl 5X reaction buffer, 2µl BP clonase, and 4µl TE buffer pH8.0 for the total volume of 10µl, was incubated at 25°C overnight and stopped by adding 1µl proteinase K and incubating at 37°C for 10 minutes. 1 µl of the above reaction sample was used to transform into TOP10 competent cells (Invitrogen). A single colony was isolated, and plasmid DNA was extracted and purified using the QIAprep Spin Miniprep Kit (Qiagen). The correct clone was verified by sequencing and used in the LR reaction with destination vector pDest-565 to create the expression clone. The LR recombination reaction mixture contained 1µl entry vector (158 µg/µl) with the correct gene insert, 1µl destination vector (172 µg/µl), 2µl 5X reaction buffer, 2µl LR clonase, and 4µl TE buffer pH8.0 for the total volume of 10µl. The LR reaction was incubated at 25°C overnight and stopped with proteinase K treatment as described above. The pDest-565 vector is constructed with both 6xHis tag and GST-tag at the N-terminus of the gene insert.

4.2.5 Co-Expression of GST-PF0863 and PF0864 and GST Pull-down Assay

The expression vectors, pETPF0864 (Kan^r) and PF0863 in pDest-565 (Amp^r), were co-transformed into *E.coli* BL21 (DE3) RIPL cell strain (Stratagene). The transformants were selected on LB agar plate containing 100µg/ml Ampicillin, 50µg/ml Kanamycin, and 35µg/ml Chloramphenicol. Three single colonies were picked for the small scale (3ml) expression test separately at 37 ºC in LB medium. When the cell density reached OD_{600}=0.6, isopropyl-β-D-thiogalactoside (IPTG) was added to the final concentration of 1mM for expression induction. Cells were grown for additional three hours. The expression level and solubility were checked by SDS-PAGE (Figure 4.7).

For GST pull-down assay, the strain with confirmed expression was grown in 500ml LB. Cells were harvested by centrifugation (5000 rpm, 15 min). 3.2 grams of the cell pellet were

split in half for testing in both normal (150 mM NaCl) and high (500 mM NaCl) salt conditions. The pellet was resuspended in lysis buffer (150 mM/500 mM NaCl, 20 mM Tris-HCl, 5 mM PMSF, 5 mM β-mercaptoethanol, pH 7.6), and lysed by French press. The cell debris was removed using centrifugation (12,000rpm, 30 min). The supernatant was loaded onto a Ni affinity column (HiTrap 5ml chelating HP; Amersham Biosciences) previously equilibrated with buffer A (150 mM/500 mM NaCl, 20 mM imidazole, 20 mM phosphate buffer, pH 7.6). Because they both have His-tags, they can be eluted at 2ml/min with a linear gradient from 20 to 500 mM imidazole over 50 ml. For identification purpose of His-GST-PF0863, a small scale TEV digestion trial was proceeded using 20 µl peak fraction sample from Ni column with 3 µl TEV protease at room temperature for 3 hours. Then all peak fractions from the Ni column elution were collected and loaded onto an equilibrated GSTrap HP (5ml) column (Amersham Biosciences) using the binding buffer (1XPBS, pH7.6, 5mM DTT) with the flow rate of 0.5 ml/min. The elution buffer has 50mM Tris-HCl, 10mM reduced glutathione, pH8.0, 5mM DTT, and step elution was preformed at 2ml/min for 20ml. The SDS-PAGE was used to analyze each purification step and TEV digestion (Figure 4.8a). To prevent the GST from interfering with the interaction between PF0863 and PF0864, all the His-GST-PF0863 was digested by TEV protease during dialysis in the Ni buffer A at 4ºC overnight. Then the proteins mixture was loaded again onto the Ni affinity column using the same buffer conditions as mentioned above. The flowthrough and peak fractions were analyzed using SDS-PAGE (Figure 4.8b).

4.3 Results and Discussion

4.3.1 Sequence Analysis

PF0865, PF0864 and PF0863 form a three-gene cluster in the operon (Figure 4.1). There is a 6bp spacer between PF0863 and PF0864 and a 3bp spacer between PF0864 and PF0865.

The first ORF in the operon, PF0865, encodes a small, basic protein (177aa; 19.2 kDa; pI 9.3) annotated as 3-octaprenyl-4-hydroxybenzoate carboxylase in TIGR. The encoded protein has an N-terminal flavoprotein domain (ID=PF02441 in Pfam) that is found in diverse flavoprotein enzymes (Kupke, Stevanovic et al. 1992; Daniel and Errington 1993; Clausen, Lamb et al. 1994). In addition, it has a polyprenyl p-hydroxybenzoate and phenylacrylic acid decarboxylase domain (TIGR00421, ubiX-pad) that spans the entire coding region. This family represents a distinct clade within the flavoprotein family mentioned above and includes aromatic acid decarboxylases (Kupke, Stevanovic et al. 1992). The cellular role of PF0865 is proposed to be involved in biosynthesis of cofactors and vitamins. The neighbor PF0864 is an autoregulator belonging to the Lrp/AsnC family, and PF0863 has nucleotidase activity. It is not clear how to relate the potential functions of all three ORFs into any well-known regulatory network. However, the fact that this same pattern of three-gene operon homologs have been found in sequenced genomes of *Thermococci* class including *P. horikoshii* (PH1012-PH1014), *P. abysii* (PAB0651-PAB0653), and *Thermococcus kodakarensis KOD1* (TK0507-TK0509) indicates that possible functional interactions exist among these gene products. It has been well demonstrated that many neighboring genes in bacterial or archaeal genomes have a propensity to encode proteins to form physical or functional interactions with each other (Dandekar, Snel et al. 1998; Overbeek, Fonstein et al. 1999; Huynen, Snel et al. 2000; Szilagyi, Grimm et al. 2005).

4.3.2 Docking Results Analysis

3D-Dock (Sternberg, Gabb et al. 1998) is the major program used for this study, along with the predictions from web-based Zdock (Chen, Li et al. 2003) and GRAMM (Vakser 1995) servers. These three programs all implement the groundbreaking searching algorithm done by

Katchalski and coworkers in 1992 (Katchalski-Katzir, Shariv et al. 1992). Their work introduced Fourier correlation into the most popular search scheme, the grid representations.

In the 3D-Dock suite, FTDock followed by RPScore were performed to predict the possible complex conformation of PF0864 and PF0863. Since PF0864 and PF0863 were believed to function in dimer formation, in FTDcok, the two dimers onto orthogonal grids were performed with a global scan of translational and rotational space. The scoring method is primarily a surface complementary score between the two grids. This step needs huge calculations involving the Fourier Transforms; in this study, the result required more than 4 days to complete. The surface complementary was the only score used in the FTDock. Then RPScore was run using the result from FTDock. This program uses an empirical pair potential matrix to re-score each possible complex from FTDock. The pair potentials are at an amino acid residue level. Each potential corresponds to the empirically derived likelihood of a trans-interface pair of two residue types, and the only limitation is the distance cut off. At present, the most useful matrix used is generated from 103 non-homologous interfaces found in the PDB with the aid of SCOP 1.50 (http://scop.mrc-lmb.cam.ac.uk/scop/). Figure 4.3 shows the overlap of top 10 complex models from RPScore. Since PF0863 model was the static model in the global docking search, it was shown in the middle of the picture. The ten different docking models have PF0864 either on the top region or the bottom region of the PF0863 model. Interestingly, in all the complex models, PF0864 uses the C-terminal RAM domain to form interactions with PF0863. The previous study also suggested that the C-terminal domain of Lrp/AsnC family members may be involved in forming the macromolecule complexes during the transcriptional regulation (Leonard, Smits et al. 2001). The results from ZDOCK server are very similar to that from 3D-Dock. The GRAMM result (Figure 4.4), however, shows that PF0864 uses its hinge region

between N- and C- terminal domain to "grab" the PF0863 structure. It gives another implication on how these two proteins may interact with each other, and those interacting regions can be the potential "hot spots" to study.

All of the docking methods based on the structure features of PF063 and PF0864 to provide predictions on their possible interactions with each other, however, visualize their interactions and characterize how they interact with each other need the verification from experimental methods.

4.3.3 Gel Filtration and Native Gel

Purified His-tagged PF0863 (21.4KDa) and PF0864 (19.3KDa) were mixed together in 1:1 molar ratio, and complex formation was analyzed by size exclusion chromatography. The molecular standard (Bio-rad, 151-1901) was run through the column using the same buffer before loading the protein sample. After elution, only one peak was observed in the chromatogram around the 40KDa (Figure 4.5). The peak fractions were analyzed using both native-PAGE and SDS-PAGE. If PF0863 and PF0864 interact, one band on the native-PAGE and two bands on the SDS-PAGE for the same fraction were expected. However, as shown in Figure 4.6, two protein bands were on both the native-PAGE and SDS-PAGE. The reasons for only one peak came out of gel filtration chromatography at around 40KDa position may be that each of the two proteins forms a homodimer of its own as mentioned in previous chapters, or only weak interaction between these two proteins, or native gel condition not favorable to the complex.

Even though this experimental process did not find interactions of PF0863 and PF0864, it does not completely exclude the possible interactions of these two in biological environments. Since each protein has been expressed, properly folded, and purified individually, the

interactions between them may not be detected when being observed in their mature forms. Therefore, co-expression and GST pull-down assay was performed for further verification.

4.3.4 Co-expression and GST Pull-down Assay

In order to test the interactions between PF0863 and PF0864 by affinity tag pull-down assay, the PF0863 gene was reconstructed into pDest-565 vector using Gateway® cloning technology (Invitrogen). This expression vector has both his-tag and GST-tag at the N-termius. After co-transformation of pET0864 and pDest-565-PF0863 into the *E.coli* BL21 (DE3) RIPL cell strain (Stratagene), both proteins were expressed as confirmed by SDS-PAGE (Figure 4.7). The cell lyses was firstly loaded onto a Ni^+ column. Since both proteins have His-tags, they can be purified together by Ni^+ affinity chromatography. Next, a small scale TEV digestion test was performed using a certain amount of sample from one peak fraction (Figure 4.7). The TEV protease can successfully remove the GST-tag from the PF0863 and give clear bands with correct molecular weights on the SDS-PAGE, confirming the product of GST-PF0863 fusion. All the peak fractions of Ni^+ chromatography were pooled and loaded onto a GST column. If PF0863 interacted with PF0864, both proteins should be bound onto the GST column and eluted together with reduced glutathione. The flowthrough and peak fractions from the GST column were analyzed using SDS-PAGE (Figure 4.4). The whole procedure of this pull-down assay is taken under low salt (150mM) and high salt (500mM) buffer conditions, because it has been suggested that some *P. furiosus* proteins interact only in high salt condition (personal communication, Dr. Frank Jenny). However, in both cases, PF0864 without the GST-tag was only observed in the flowthrough, while PF0863 with the GST-tag was mainly found in the elution peaks. Because the possibility of GST fusion of PF0864 at the N-terminus may cover the binding site of PF0864 to PF0863, after the TEV digestion, the sample was loaded again onto the

Ni affinity column. At this time, His-tagged PF0864 (19361Da) and His-GST (29616Da) were observed in the elution peaks, while PF0863 (20675Da) was only seen in the flowthrough. GST pull-down assay again failed to prove any detectable interactions between PF0863 and PF0864.

4.4 Conclusion

Analysis of PF0863 and PF0864 from the genome level identified a conserved pattern of three-gene operon in several *Thermococci* genomes. The individually solved structures of PF0863 and PF0864 show complementary features. The structural-based docking of PF0863 and PF0864 suggests the possible interacting sites within these two proteins. From those predictions, PF0863 and PF0864 have a high possibility of forming a protein complex. However, the experiments using both the gel filtration chromatography followed by native-PAGE and co-expression followed by GST pull-down do not experimentally prove interactions between them.

PF0863 has the nucleotidase activity of hydrolyzing nucleosides tri- and diphosphates. PF0864 is a transcriptional regulatory protein, belonging to the Lrp/AsnC family and having the specific DNA-binding activity with the potential promoter region in front of the operon containing PF0865, PF0864, and PF0863. The function of PF0865 is still unclear, although it is annotated as 3-octaprenyl-4-hydroxybenzoate carboxylase in TIGR. How they are functionally related? Why are they conserved in the same operon among *Thermococci* genomes? Does the interaction between PF0863 and PF0864 require PF0865 or other co-factors? These questions need to be addressed and answered in the future.

The protein-protein interaction network is complicated, especially when the functional and physical interactions between protein molecules are sometimes transient. Finding and elucidating the stable functional protein-protein complex is a growing trend in the scientific community. The rapid increasing structural data available from structural genomics is providing

more and more valuable information to facilitate the study of complex from a structural point of view.

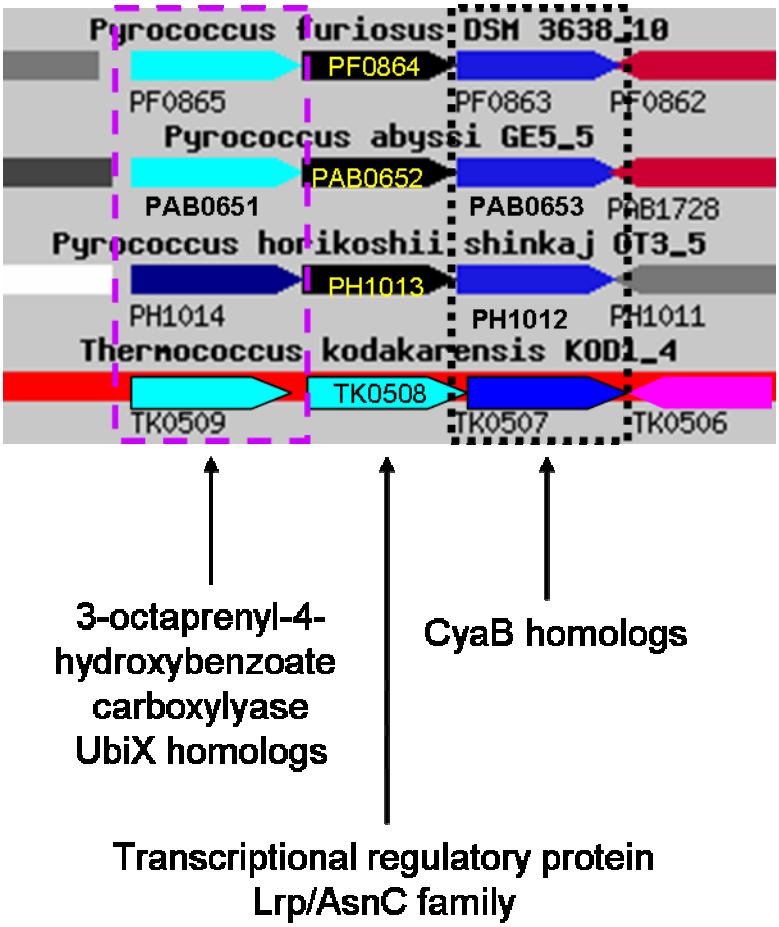


Figure 4.1: The genome region comparison of PF0865-PF0864-PF0863 in *P. furiosus* with *P. abyssi, P. horikoshii shinka.J* OT3, and *Thermococcus kodakarensis KOD1*.

The same pattern of PF0865-PF0864-PF0863 in one operon can be found in *P. horikoshii shinka.J* OT3, *P. abyssi*, and *Thermococcus kodakarensis KOD1*. The first gene is 3-octaprenyl-4-hyfroxybenzoate carboxylase, the second gene is a transcriptional regulatory protein belonging to Lrp/AsnC family, and the third gene is the CyaB homologs. This picture was generated from TIGR-CMR.

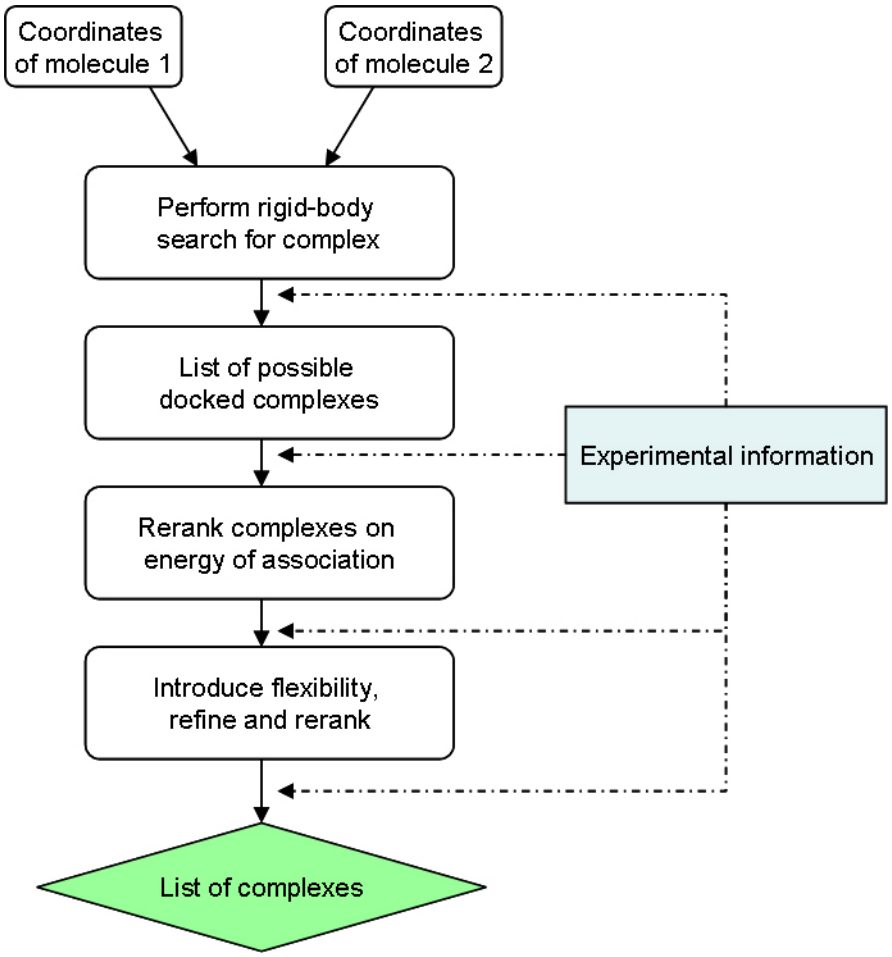


Figure 4.2: The protein-protein docking stages.

The process starts with two known structures. The additional experimental information helps to filter each successive step to the final list of complexes.

Modified from (Smith and Sternberg 2002).

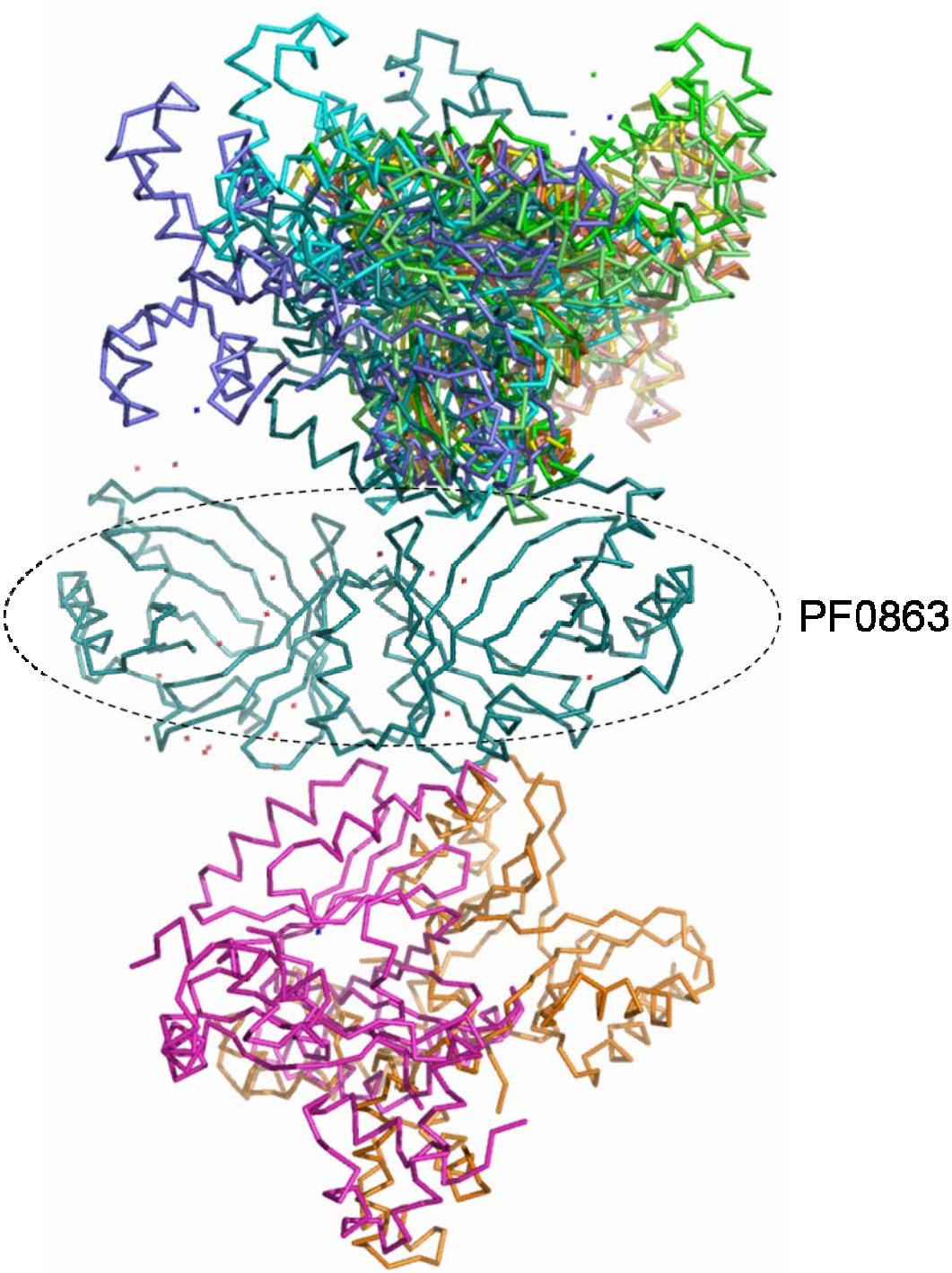


Figure 4.3: The overlap of possible complex models of PF0863 and PF0864 from 3D-Dock.

The PF0863 model is the static one in the global docking search. The predicted interacting region on PF0864 is all in the C-terminal RAM domain. All the ribbon models are generated by PyMOL (DeLano 2002).

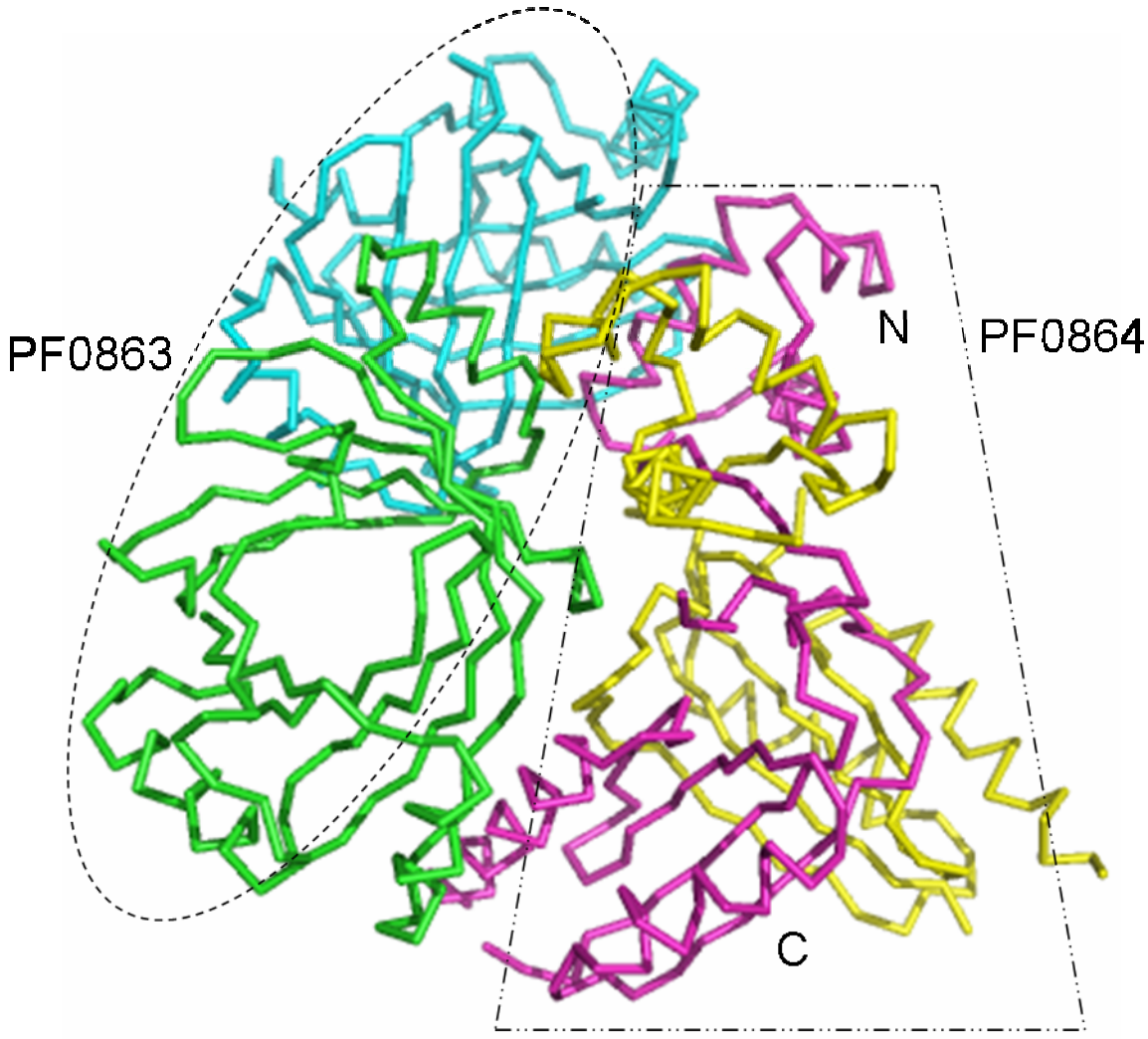


Figure 4.4: The docking result from GRAMM server.

The PF0863 (green & cyan) and PF0864 (magenta & yellow) ribbon structures are made by PyMOL (DeLano 2002; C. elegans Sequencing Consortium 1998). The hinge region between N- and C- terminal domain of PF0864 is interacting with PF0863 near its dimerization interface.

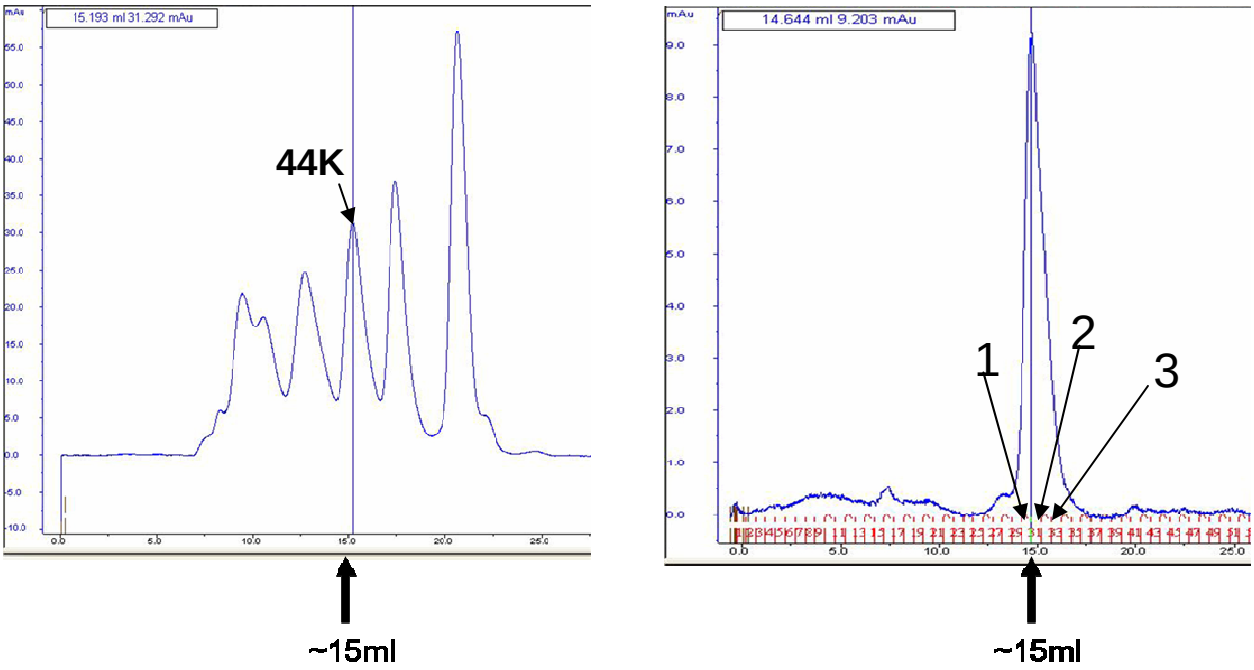


Figure 4.5: The chromatogram of gel filtration.

The left picture shows the elution positions of molecular standard from the gel filtration column. At around 15ml, ovalbumin from chicken (44KDa) is eluted. The right picture shows the PF0863 and PF0864 mixture running through the same gel filtration column. The elution peak is around 15ml position.

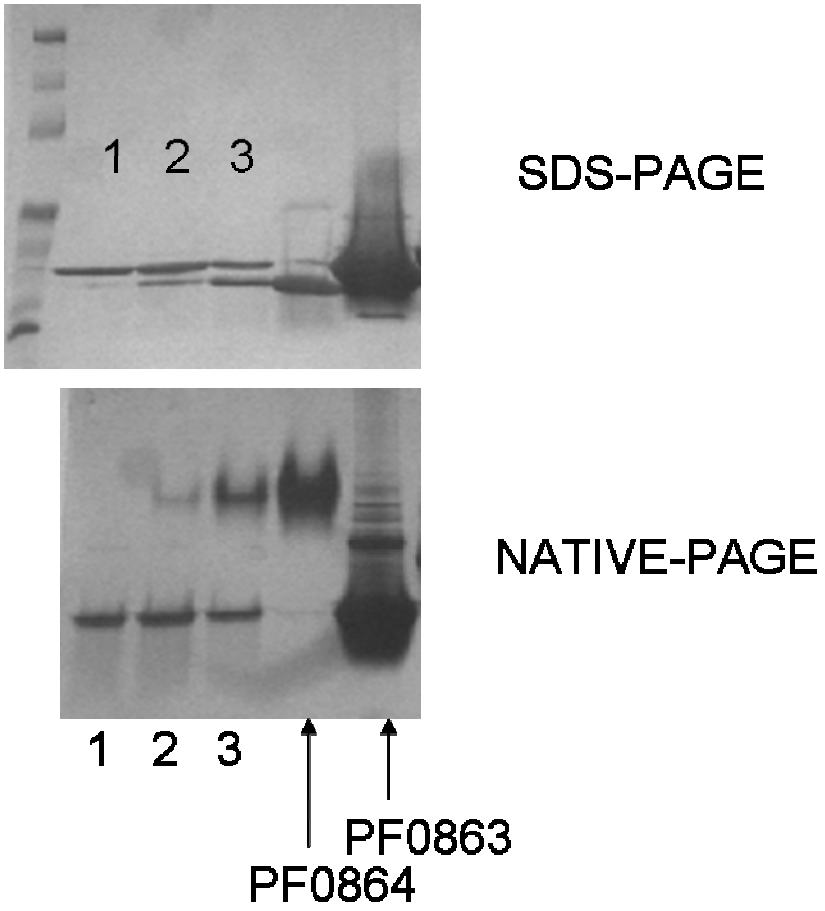


Figure 4.6: The native-PAGE and SDS-PAGE

Lanes 1-3 are peaks from gel filtration, and lanes labeled PF0864 and PF0863 are the two proteins individually. On both gels, two protein bands are observed from the peak fractions that are on the same positions as the bands of individual protein.

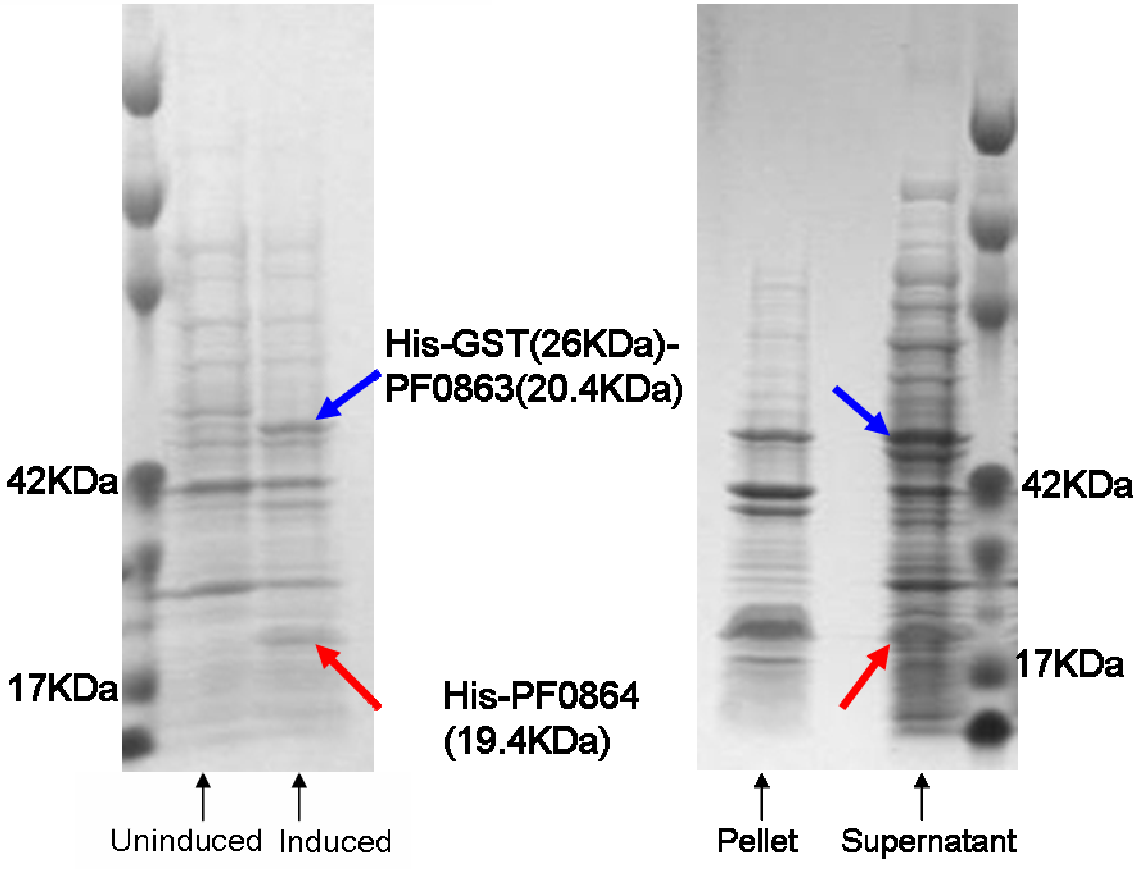


Figure 4.7: The co-expression and solubility test of PF0863 and PF0864.

The left gel shows PF0863 and PF0864 with correct molecular weights co-expressed in the same cell. The right gel shows that both of them are soluble. Blue arrows point the His-GST-PF0863. Red arrows point the His-PF0864.

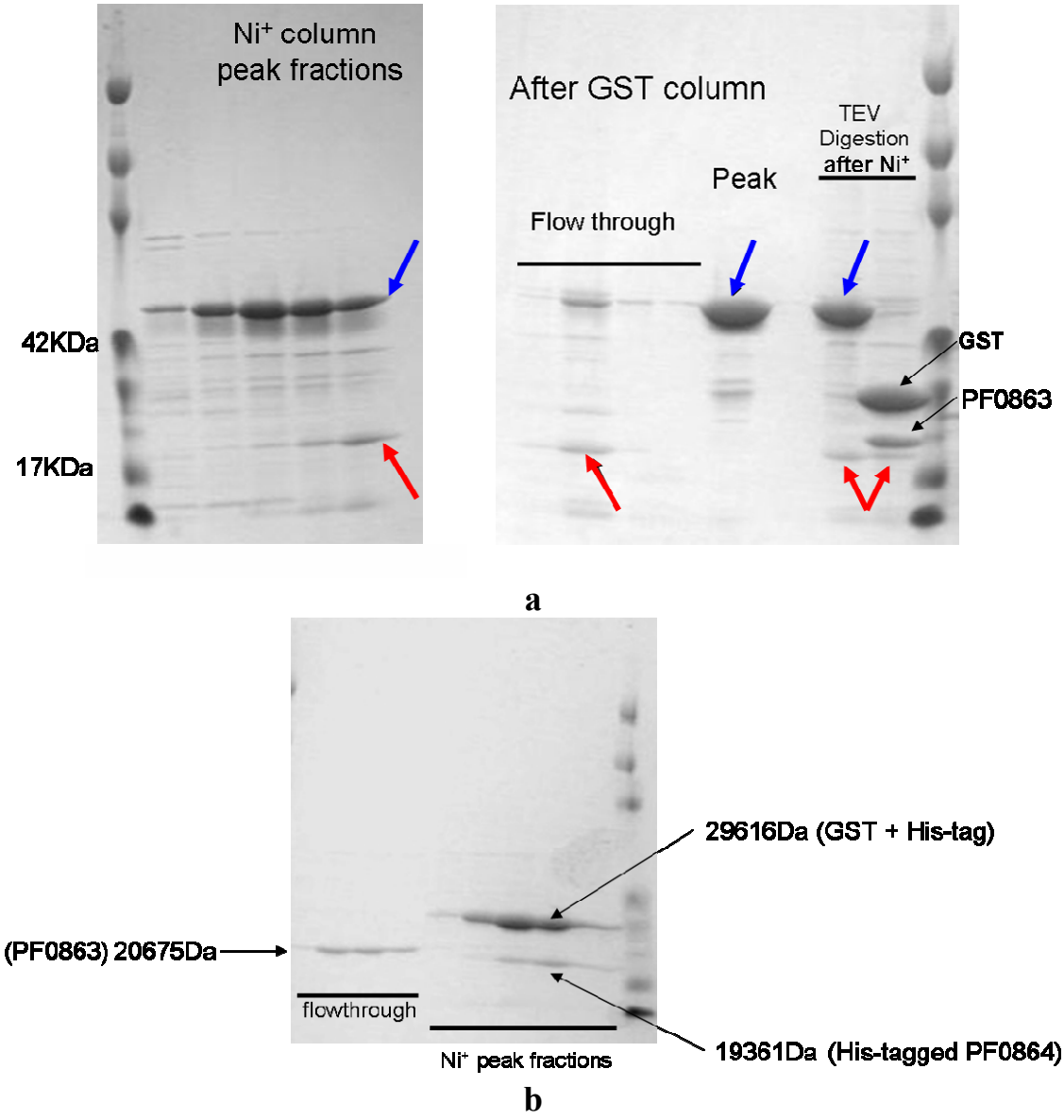


Figure 4.8: The Ni—GST pull-down—Ni analysis of PF0863 and PF0864.

The top gels (a) show the results from Ni and GST affinity columns. The blue arrows point at the His-GST-PF0863 fusion protein. The red arrows point at His-tagged PF0864. Two proteins are observed in the Ni column peak fractions. With GST column, PF0864 is observed in the flow through while PF0863 is in the peak fraction showing that they do not form stable complex in this condition. The TEV digestion can successfully separate GST and PF0863. The bottom gel (b) shows the results of re-run Ni affinity column after TEV digestion. The PF0863 without any tag was in the flowthough. The PF0864 was eluted in the peak fractions along with the His-GST.

Reference

(C. elegans Sequencing Consortium 1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-8.

Adams, M. W., H. A. Dailey, et al. (2003). "The Southeast Collaboratory for Structural Genomics: a high-throughput gene to structure factory." Acc Chem Res **36**(3): 191-8.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anderson, J. J., S. C. Quay, et al. (1976). "Mapping of two loci affecting the regulation of branched-chain amino acid transport in Escherichia coli K-12." J Bacteriol **126**(1): 80-90.

Aravind, L., V. Anantharaman, et al. (2005). "The many faces of the helix-turn-helix domain: transcription regulation and beyond." FEMS Microbiol Rev **29**(2): 231-62.

Aravind, L. and E. V. Koonin (1999). "DNA-binding proteins and evolution of transcription regulation in the archaea." Nucleic Acids Res **27**(23): 4658-70.

Aravind, L. and E. V. Koonin (1999). "Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches." J Mol Biol **287**(5): 1023-40.

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

Bajaj, M. and T. Blundell (1984). "Evolution and the tertiary structure of proteins." Annu Rev Biophys Bioeng **13**: 453-92.

Baker, N. A., D. Sept, et al. (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proc Natl Acad Sci U S A **98**(18): 10037-41.

Baldock, P., V. Mills, et al. (1996). "A comparison of microbatch and vapour diffusion for initial screening of crystallization conditions." J Cryst Growth **168**: 170-174.

Bartlett, M. S., M. Thomm, et al. (2000). "The orientation of DNA in an archaeal transcription initiation complex." Nat Struct Biol **7**(9): 782-5.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-41.

Bell, S. D. and S. P. Jackson (1998). "Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features." Trends Microbiol **6**(6): 222-8.

Bell, S. D. and S. P. Jackson (2001). "Mechanism and regulation of transcription in archaea." Curr Opin Microbiol **4**(2): 208-13.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-42.

Blochl, E., R. Rachel, et al. (1997). "Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C." Extremophiles **1**(1): 14-21.

Blow, D. M. (2003). "How Bijvoet made the difference: the growing power of anomalous scattering." Methods Enzymol **374**: 3-22.

Brennan, R. G. and B. W. Matthews (1989). "The helix-turn-helix DNA binding motif." J Biol Chem **264**(4): 1903-6.

Brenner, S. E. (2001). "A tour of structural genomics." Nat Rev Genet **2**(10): 801-9.

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

Brinkman, A. B., I. Dahlke, et al. (2000). "An Lrp-like transcriptional regulator from the archaeon Pyrococcus furiosus is negatively autoregulated." J Biol Chem **275**(49): 38160-9.

Brinkman, A. B., T. J. Ettema, et al. (2003). "The Lrp family of transcriptional regulators." Mol Microbiol **48**(2): 287-94.

Brunger, A. T. (1992). "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures." Nature (London) **355**: 472-475.

Burley, S. K., S. C. Almo, et al. (1999). "Structural genomics: beyond the human genome project." Nat Genet **23**(2): 151-7.

Calvo, J. M. and R. G. Matthews (1994). "The leucine-responsive regulatory protein, a global regulator of metabolism in Escherichia coli." Microbiol Rev **58**(3): 466-90.

CCP4 (1994). "The CCP4 suite: programs for protein crystallography." Acta Crystallogr D Biol Crystallogr **50**(Pt 5): 760-3.

Chandonia, J. M. and S. E. Brenner (2006). "The impact of structural genomics: expectations and outcomes." Science **311**(5759): 347-51.

Chayen, N. E., P. D. Shaw-Steward, et al. (1992). "Microbatch crystallization under oil -- a new technique allowing many small volume crystallization experiments " J. Cryst. Growth **122**: 176-180.

Chen, R., L. Li, et al. (2003). "ZDOCK: an initial-stage protein-docking algorithm." Proteins **52**(1): 80-7.

Chen, S., M. H. Rosner, et al. (2001). "Leucine-regulated self-association of leucine-responsive regulatory protein (Lrp) from Escherichia coli." J Mol Biol **312**(4): 625-35.

Chipman, D. M. and B. Shaanan (2001). "The ACT domain family." Curr Opin Struct Biol **11**(6): 694-700.

Clausen, M., C. J. Lamb, et al. (1994). "PAD1 encodes phenylacrylic acid decarboxylase which confers resistance to cinnamic acid in Saccharomyces cerevisiae." Gene **142**(1): 107-12.

Cowtan, K. D. and K. Y. Zhang (1999). "Density modification for macromolecular phase improvement." Prog Biophys Mol Biol **72**(3): 245-70.

Crick, F. H. C. and B. S. Magdoff (1956). "MIR." Acta Crystallogr **9**: 901-908.

Cuff, J. A., M. E. Clamp, et al. (1998). "JPred: a consensus secondary structure prediction server." Bioinformatics **14**(10): 892-3.

D'Arcy, A., C. Elmore, et al. (1996). "A novel approach to crystallising proteins under oil." J. Cryst. Growth **168**: 175-180.

Danchin, A. (1993). "Phylogeny of adenylyl cyclases." Adv Second Messenger Phosphoprotein Res **27**: 109-62.

Danchin, A., J. Pidoux, et al. (1993). "The adenylate cyclase catalytic domain of Streptomyces coelicolor is carboxy-terminal." FEMS Microbiol Lett **114**(2): 145-51.

Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci **23**(9): 324-8.

Daniel, R. A. and J. Errington (1993). "Cloning, DNA sequence, functional analysis and transcriptional regulation of the genes encoding dipicolinic acid synthetase required for sporulation in Bacillus subtilis." J Mol Biol **232**(2): 468-83.

Dauter, Z. (2006). "Current state and prospects of macromolecular crystallography." Acta Crystallogr D Biol Crystallogr **62**(Pt 1): 1-11.

Dauter, Z., I. Botos, et al. (2005). "Pathological crystallography: case studies of several unusual macromolecular crystals." Acta Crystallogr D Biol Crystallogr **61**(Pt 7): 967-75.

Davis, D. R. and D. M. Segal (1971). "Vapor diffusion in hanging drop." Methods Enzymol **22**: 266-269.

Deane, C. M., L. Salwinski, et al. (2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." Mol Cell Proteomics **1**(5): 349-56.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. San Carlos, CA, USA, DeLano Scientific.

Drenth, J. (1999). Principles of protein x-ray crystallography. New York, Springer.

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.

Erlandsen, H., E. E. Abola, et al. (2000). "Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites." Curr Opin Struct Biol **10**(6): 719-30.

Ettema, T. J., A. B. Brinkman, et al. (2002). "A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes." J Biol Chem **277**(40): 37464-8.

Ferre-D'Amare, A. R. and S. K. Burley (1994). "Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies." Structure **2**(5): 357-9.

Fiala, G. and K. O. Stetter (1986). "Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C." Archives of Microbiology **145**: 56-61.

Fields, S., Y. Kohara, et al. (1999). "Functional genomics." Proc Natl Acad Sci U S A **96**(16): 8825-6.

Fields, S. and O. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-6.

Fischer, D. (1999). "Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge." Protein Eng **12**(12): 1029-30.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fu, Z. Q., J. Rose, et al. (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.

Gabb, H. A., R. M. Jackson, et al. (1997). "Modelling protein docking using shape complementarity, electrostatics and biochemical information." J Mol Biol **272**(1): 106-20.

Galperin, M. Y. and E. V. Koonin (2000). "Who's your neighbor? New computational approaches for functional genomics." Nat Biotechnol **18**(6): 609-13.

García-Ruiz, J. M. and A. Moreno (1994). "Investigations on protein crystal growth by the gel acupuncture method." Acta Crystallogr D Biol Crystallogr **50**: 484-490.

Geiduschek, E. P. and M. Ouhammouch (2005). "Archaeal transcription and its regulators." Mol Microbiol **56**(6): 1397-407.

Gill, S. C. and P. H. von Hippel (1989). "Calculation of protein extinction coefficients from amino acid sequence data." Anal Biochem **182**(2): 319-26.

Glusker, J. P., M. Lewis, et al. (1994). Crystal structure analysis for chemists and biologists New York, Wiley-VCH, Inc.

Goff, S. A., D. Ricke, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.

Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.

Gong, C., P. Smith, et al. (2006). "Structure-function analysis of Plasmodium RNA triphosphatase and description of a triphosphate tunnel metalloenzyme superfamily that includes Cet1-like RNA triphosphatases and CYTH proteins." Rna **12**(8): 1468-74.

Harker, D. (1956). "MIR." Acta Crystallogr **9**: 1-9.

Hartley, J. L., G. F. Temple, et al. (2000). "DNA cloning using in vitro site-specific recombination." Genome Res **10**(11): 1788-95.

Hendrickson, W. A., J. R. Horton, et al. (1990). "Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure." Embo J **9**(5): 1665-72.

Hendrickson, W. A., J. L. Smith, et al. (1985). "Direct phase determination based on anomalous scattering." Methods Enzymol **115**: 41-55.

Hendrickson, W. A. and M. M. Teeter (1981). "Single wavelength phasing." Nature (London) **290**: 107-113.

Hieter, P. and M. Boguski (1997). "Functional genomics: it's all how you read it." Science **278**(5338): 601-2.

Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." J Mol Biol **233**(1): 123-38.

Hough, D. W. and M. J. Danson (1999). "Extremozymes." Curr Opin Chem Biol **3**(1): 39-46.

Hutchins, A. M., J. F. Holden, et al. (2001). "Phosphoenolpyruvate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus." J Bacteriol **183**(2): 709-15.

Huynen, M., B. Snel, et al. (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." Genome Res **10**(8): 1204-10.

Huynen, M. A., B. Snel, et al. (2003). "Function prediction and protein networks." Curr Opin Cell Biol **15**(2): 191-8.

Ingram, D. J. and J. C. Kendrew (1956). "Orientation of the haem group in myoglobin and its relation to the polypeptide chain direction." Nature **178**(4539): 905-6.

Iyer, L. M. and L. Aravind (2002). "The catalytic domains of thiamine triphosphatase and CyaB-like adenylyl cyclase define a novel superfamily of domains that bind organic phosphates." BMC Genomics **3**(1): 33.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins." J Mol Biol **3**: 318-56.

Jafri, S., S. Evoy, et al. (1999). "An Lrp-type transcriptional regulator from Agrobacterium tumefaciens condenses more than 100 nucleotides of DNA into globular nucleoprotein complexes." J Mol Biol **288**(5): 811-24.

Johnson, M. S., M. J. Sutcliffe, et al. (1990). "Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins." J Mol Evol **30**(1): 43-59.

Kahn, R., R. Fourme, et al. (1985). "Crystal structure study of Opsanus tau parvalbumin by multiwavelength anomalous diffraction." FEBS Lett **179**(1): 133-7.

Katchalski-Katzir, E., I. Shariv, et al. (1992). "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques." Proc Natl Acad Sci U S A **89**(6): 2195-9.

Kim, S. H. (1998). "Shining a light on structural genomics." Nat Struct Biol **5 Suppl**: 643-5.

Koike, H., S. A. Ishijima, et al. (2004). "The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription." Proc Natl Acad Sci U S A **101**(9): 2840-5.

Kolling, R. and H. Lother (1985). "AsnC: an autogenously regulated activator of asparagine synthetase A transcription in Escherichia coli." J Bacteriol **164**(1): 310-5.

Kupke, T., S. Stevanovic, et al. (1992). "Purification and characterization of EpiD, a flavoprotein involved in the biosynthesis of the lantibiotic epidermin." J Bacteriol **174**(16): 5354-61.

Kyrpides, N. C. and C. A. Ouzounis (1995). "The eubacterial transcriptional activator Lrp is present in the archaeon Pyrococcus furiosus." Trends Biochem Sci **20**(4): 140-1.

Kyrpides, N. C. and C. A. Ouzounis (1999). "Transcription in archaea." Proc Natl Acad Sci U S A **96**(15): 8545-50.

Lakaye, B., A. F. Makarchikov, et al. (2002). "Molecular characterization of a specific thiamine triphosphatase widely expressed in mammalian tissues." J Biol Chem **277**(16): 13771-7.

Laskowski, R. A., D. S. Moss, et al. (1993). "Main-chain bond lengths and bond angles in protein structures." J Mol Biol **231**(4): 1049-67.

Legrain, P., J. Wojcik, et al. (2001). "Protein--protein interaction maps: a lead towards cellular functions." Trends Genet **17**(6): 346-52.

Leonard, G. A., G. Sainz, et al. (2005). "Automatic structure determination based on the single-wavelength anomalous diffraction technique away from an absorption edge." Acta Crystallogr D Biol Crystallogr **61**(Pt 4): 388-96.

Leonard, P. M., S. H. Smits, et al. (2001). "Crystal structure of the Lrp-like transcriptional regulator from the archaeon Pyrococcus furiosus." Embo J **20**(5): 990-7.

Lima, C. D., L. K. Wang, et al. (1999). "Structure and mechanism of yeast RNA triphosphatase: an essential component of the mRNA capping apparatus." Cell **99**(5): 533-43.

Linial, M. and G. Yona (2000). "Methodologies for target selection in structural genomics." Prog Biophys Mol Biol **73**(5): 297-320.

Liu, Z. J., D. Lin, et al. (2005). "Parameter-space screening: a powerful tool for high-throughput crystal structure determination." Acta Crystallogr D Biol Crystallogr **61**(Pt 5): 520-7.

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

Lovell, S. C., I. W. Davis, et al. (2003). "Structure validation by Calpha geometry: phi,psi and Cbeta deviation." Proteins **50**(3): 437-50.

Lovenberg, W., B. B. Buchanan, et al. (1963). "Studies on the Chemical Nature of Clostridial Ferredoxin." J Biol Chem **238**: 3899-913.

Lu, L., H. Lu, et al. (2002). "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading." Proteins **49**(3): 350-64.

Madhusudhan, K. T., N. Huang, et al. (1995). "Characterization of BkdR-DNA binding in the expression of the bkd operon of Pseudomonas putida." J Bacteriol **177**(3): 636-41.

Martin, A. C., M. W. MacArthur, et al. (1997). "Assessment of comparative modeling in CASP2." Proteins **Suppl 1**: 14-28.

Martin, D. B. and P. S. Nelson (2001). "From genomics to proteomics: techniques and applications in cancer research." Trends Cell Biol **11**(11): S60-5.

McKusick, V. A. (1997). "Genomics: structural and functional studies of genomes." Genomics **45**(2): 244-9.

McPherson, A. (1999). Crystallization of biological macromolecules. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

McRee, D. E. (1999). "XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density." J Struct Biol **125**(2-3): 156-65.

Moont, G., H. A. Gabb, et al. (1999). "Use of pair potentials across protein interfaces in screening predicted docked complexes." Proteins **35**(3): 364-73.

Mulder, N. J., R. Apweiler, et al. (2003). "The InterPro Database, 2003 brings increased coverage and new features." Nucleic Acids Res **31**(1): 315-8.

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

Newman, E. B. and R. Lin (1995). "Leucine-responsive regulatory protein: a global regulator of gene expression in E. coli." Annu Rev Microbiol **49**: 747-75.

Ng, J. D., J. A. Gavira, et al. (2003). "Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination." J Struct Biol **142**(1): 218-31.

Nowak, J. Z. and J. B. Zawilska (1999). "[Adenylyl cyclase--isoforms, regulation and function]." Postepy Hig Med Dosw **53**(2): 147-72.

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

Overbeek, R., M. Fonstein, et al. (1999). "The use of gene clusters to infer functional coupling." Proc Natl Acad Sci U S A **96**(6): 2896-901.

Panjikar, S., V. Parthasarathy, et al. (2005). "Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment." Acta Crystallogr D Biol Crystallogr **61**(Pt 4): 449-57.

Pazos, F., O. Olmea, et al. (1997). "A graphical interface for correlated mutations and other protein structure prediction methods." Comput Appl Biosci **13**(3): 319-21.

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

Robb, F. T., D. L. Maeder, et al. (2001). "Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology." Methods Enzymol **330**: 134-57.

Rossmann, M. G. (1972). The molecular replacement method; a collection of papers on the use of non-crystallographic symmetry. New York,, Gordon and Breach.

Rossmann, M. G. and D. M. Blow (1962). "The detection of sub-units within the crystallographic asymmetric unit." Acta Cryst. **15**: 24-31.

Rost, B. and C. Sander (1993). "Prediction of protein secondary structure at better than 70% accuracy." J Mol Biol **232**(2): 584-99.

Rothschild, L. J. and R. L. Mancinelli (2001). "Life in extreme environments." Nature **409**(6823): 1092-101.

Russell, R. B., F. Alber, et al. (2004). "A structural perspective on protein-protein interactions." Curr Opin Struct Biol **14**(3): 313-24.

Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-33.

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

Sambrook, J., E. F. Fritsch, et al. (1989). Molecular cloning : a laboratory manual. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory.

Sambrook, J. and D. W. Russell (2001). Molecular Cloning: A Laboratory Manual, 3rd ed. . Cold Spring Harbor, NY., Cold Spring Harbor Laboratory Press.

Schultz, J. E. and S. Klumpp (1994). "Cyclic GMP in lower forms." Adv Pharmacol **26**: 285-303.

Schut, G. J., S. D. Brehm, et al. (2003). "Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: Pyrococcus furiosus grown on carbohydrates or peptides." J Bacteriol **185**(13): 3935-47.

Schut, G. J., J. Zhou, et al. (2001). "DNA microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus: evidence for anNew type of sulfur-reducing enzyme complex." J Bacteriol **183**(24): 7027-36.

Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.

Sismeiro, O., P. Trotot, et al. (1998). "Aeromonas hydrophila adenylyl cyclase 2: a new class of adenylyl cyclases with thermophilic properties and sequence similarities to proteins from hyperthermophilic archaebacteria." J Bacteriol **180**(13): 3339-44.

Smith, G. R. and M. J. Sternberg (2002). "Prediction of protein-protein interactions by docking methods." Curr Opin Struct Biol **12**(1): 28-35.

Sobott, F. and C. V. Robinson (2002). "Protein complexes gain momentum." Curr Opin Struct Biol **12**(6): 729-34.

Sternberg, M. J., H. A. Gabb, et al. (1998). "Predictive docking of protein-protein and protein-DNA complexes." Curr Opin Struct Biol **8**(2): 250-6.

Stetter, K. O. (1996). "Hyperthermophiles in the history of life." Ciba Found Symp **202**: 1-10; discussion 11-8.

Strausberg, R. L., E. A. Feingold, et al. (1999). "The mammalian gene collection." Science **286**(5439): 455-7.

Sugar, F. J., F. E. Jenney, Jr., et al. (2005). "Comparison of small- and large-scale expression of selected Pyrococcus furiosus genes as an aid to high-throughput protein production." J Struct Funct Genomics **6**(2-3): 149-58.

Suzuki, M. (2003). "Stucture and function of the feast/famine regulatory proteins, FFRPs." Proc Jpn Acad **79B**: 274-289.

Szilagyi, A., V. Grimm, et al. (2005). "Prediction of physical protein-protein interactions." Phys Biol **2**(1-2): S1-S16.

Takagi, M., M. Nishioka, et al. (1997). "Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR." Appl Environ Microbiol **63**(11): 4504-10.

Tani, T. H., A. Khodursky, et al. (2002). "Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis." Proc Natl Acad Sci U S A **99**(21): 13471-6.

Taylor, G. (2003). "The phase problem." Acta Crystallogr D Biol Crystallogr **59**(Pt 11): 1881-90.

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

Terwilliger, T. C. (2002). "Automated structure solution, density modification and model building." Acta Crystallogr D Biol Crystallogr **58**(Pt 11): 1937-40.

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

Thaw, P., S. E. Sedelnikova, et al. (2006). "Structural insight into gene transcriptional regulation and effector binding by the Lrp/AsnC family." Nucleic Acids Res **34**(5): 1439-49.

Tuan, L. R., R. D'Ari, et al. (1990). "The leucine regulon of Escherichia coli K-12: a mutation in rblA alters expression of L-leucine-dependent metabolic operons." J Bacteriol **172**(8): 4529-35.

Vakser, I. A. (1995). "Protein docking for low-resolution structures." Protein Eng **8**(4): 371-7.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Venter, J. C., H. O. Smith, et al. (1999). "Microbial genomics: in the beginning." ASM News **65**: 322-327.

Vitkup, D., E. Melamud, et al. (2001). "Completeness in structural genomics." Nat Struct Biol **8**(6): 559-66.

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

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

Weinberg, M. V., G. J. Schut, et al. (2005). "Cold shock of a hyperthermophilic archaeon: Pyrococcus furiosus exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins." J Bacteriol **187**(1): 336-48.

Weselak, M., M. G. Patch, et al. (2003). "Robotics for automated crystal formation and analysis." Methods Enzymol **368**: 45-76.

Willins, D. A., C. W. Ryan, et al. (1991). "Characterization of Lrp, and Escherichia coli regulatory protein that mediates a global response to leucine." J Biol Chem **266**(17): 10768-74.

Winn, M. D., M. N. Isupov, et al. (2001). "Use of TLS parameters to model anisotropic displacements in macromolecular refinement." Acta Crystallogr D Biol Crystallogr **57**(Pt 1): 122-33.

Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-9.

Word, J. M., S. C. Lovell, et al. (1999). "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms." J Mol Biol **285**(4): 1711-33.

Yokoyama, K., S. A. Ishijima, et al. (2006). "Feast/famine regulatory proteins (FFRPs): Escherichia coli Lrp, AsnC and related archaeal transcription factors." FEMS Microbiol Rev **30**(1): 89-108.

Yu, J., S. Hu, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.

Zeppezauer, M., H. Eklund, et al. (1968). "Micro diffusion cells for the growth of single protein crystals by means of equilibrium dialysis." Arch Biochem Biophys **126**(2): 564-73.

Zhou, J. (2004). Microbial functional genomics. Hoboken, N.J., Wiley-Liss.

Zillig, W., K. O. Stetter, et al. (1979). "DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius." Eur J Biochem **96**(3): 597-604.