SIMULATION EXPANSION AND STRUCTURAL REALIZATION OF THE 1985 SULFUR-SAS PHASING INSIGHT

By

JEFFERY EDWARD HABEL

(Under the Direction of Bi-Cheng Wang)

ABSTRACT

The Achilles' heel of a protein crystal diffraction experiment is the loss of phase information necessary in the Fourier transform calculation of electron density. Traditionally, this "phase problem" was overcome by the incorporation of a heavy metal atom into the crystal. In the mid-1980's, phasing proteins from the anomalous scattering of naturally occurring sulfur atoms in amino acids was realized on a small peptide scale (Hendrickson, 1981) and then hypothesized and simulated for a larger protein (Wang, 1985). Technological advances in X-ray generation and detection over the next fifteen years culminated in the *de novo* structure solution of obelin from *Obelia longissima* in 2000 (Liu *et al.*, 2000). This led to the era of "direct crystallography," where anomalous scattering from the atoms inherent in a protein are used to phase the protein structure. A new statistic, R_{as} (Fu *et al.*, 2004), derived solely for measured data was developed to monitor the anomalous signal. In a simulation study of a large protein, R_{as} was used to monitor the incorporation of error and affect of redundancy in rescuing the data to phase the protein, elucidating a minimum R_{as} threshold value of 1.6. Comparison to synchrotron data was unsuccessful, but the value was later validated through the structure solution of a Southeast Collaboratory for Structural Genomics (SECSG) protein, *Pfu*-542154, from multiple

crystals and different X-ray sources. This new structure pushes the current limits of *de novo* phasing with sulfur single-wavelength anomalous scattering (SAS) and verifies the original simulation of phasing approximately 50 amino acids per sulfur atom. At the same time, the simulation presented here raises the bar set by the original 1985 sulfur-SAS simulation to new heights for crystallographers to attain. This research has the possibility of helping every member of crystallographic community and with the continued technological advancement in X-ray sources and detectors; sulfur-SAS will become a more commonplace solution to the phase problem.

INDEX WORDS: Phase problem, direct crystallography, sulfur-SAS, structural genomics

SIMULATION EXPANSION AND STRUCTURAL REALIZATION OF THE 1985 SULFUR-SAS PHASING INSIGHT

by

JEFFERY EDWARD HABEL

B.S., Eastern Kentucky University, 1994

M.S., Georgia Institute of Technology, 1998

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

2005

SIMULATION EXPANSION AND STRUCTURAL REALIZATION OF THE 1985 SULFUR-SAS PHASING INSIGHT

by

JEFFERY EDWARD HABEL

Major Professor: Bi-Cheng Wang

Committee: John Rose
Harry Dailey
William Lanzilotta
James Omichinski

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

DEDICATION

For Paula

No matter the time, place, or distance between us, you have always been my best friend and for that I will love you the rest of my life.

ACKNOWLEDGEMENTS

The past five years have been the most exhilarating and scientifically frustrating times of my life. I have been privileged to work with amazing people and experience more than most graduate students ever dream about. This environment of scientific and technological discovery was all made possible by one man, Dr. B.C. Wang. As a mentor, he is unsurpassed in his drive and enthusiasm for all aspects of science. These qualities coupled with his patience and vast knowledge make learning from him a once in a lifetime experience. It is the endurance needed for a successful outcome that I take with me as I advance my scientific career.

I would also like to thank all of the members of my committee, Dr. John Rose, Dr. Harry Dailey, Dr. Bill Lanzilotta, and Dr. Jim Omichinski for their guidance and constant direction in making me as well rounded a biochemist as possible. My sincerest thanks go to Dr. James Liu whose door continues to always be open for discussion and problem solving. Without his help, I would not have been able to carry out the research presented here. Thank you to Dr. Wolfram Tempel for help with the final stages of structure refinement and deposition, and allowing me to focus on my research without worrying about systems administration over the years.

I am eternally grateful for all of my friends and co-workers in the Wang Lab and the biochemistry department these past five years. Sharing the good and bad while laughing all the time makes graduate school one of the best experiences of my life and I wouldn't trade it for anything. Finally, thank you to my family and loved ones for their unending and unwavering support and love throughout my entire scholastic career.

You go through graduate school with “your” project and focus on “your” research. At times it seems as if you are alone. Only when you look back do you realize that there was never a time when you weren’t surrounded by people who supported and cared about you and your work. For that, I am truly thankful.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

Introduction

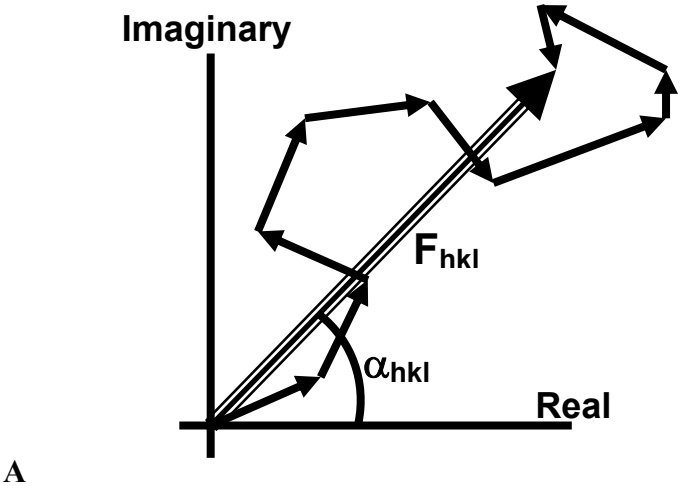
The primary equation used by all protein crystallographers in structure solution is the electron density Fourier transform equation (Figure 1.1). This equation requires four components: volume of a unit cell, a position value, an intensity measurement, and a phase difference. In a single crystal diffraction experiment, the diffraction data on the detector's surface is assigned a positional grid system (h,k,l), and the intensity (I_{hkl}) of each diffraction spot can be measured. The volume of the unit cell can be calculated from the angles and distances between the measured diffraction spots. The F_{hkl} value, known as the structure factor, in the equation is proportional to the square root of I_{hkl} ($I_{hkl} \approx F_{hkl}^2$). In addition to being a calculated value, F_{hkl} is also the sum of all the individual atomic scattering factors in the unit cell, smallest building block, of the crystal. Therefore, any change in the contents of the unit cell or individual atomic scattering intensity is manifested as change in every I_{hkl} and F_{hkl}. Unfortunately, the phase information (α_{hkl}) can not be measured directly from the diffraction data and is lost. This loss of phase information is classically referred to as the phase problem.

1.1 The phase problem

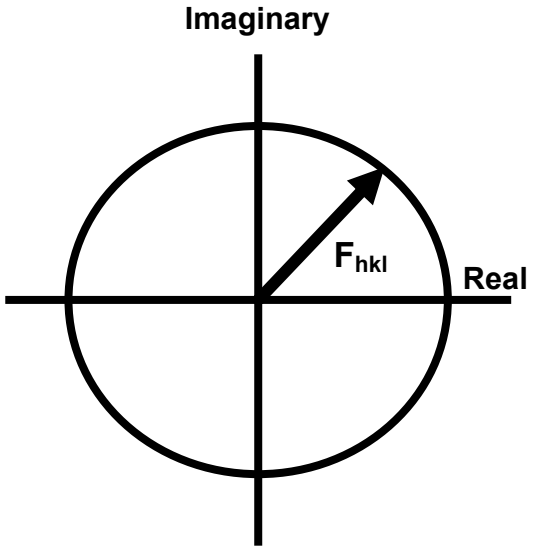
The structure factor can be represented as a vector on an Argand diagram with a horizontal "real" axis and a vertical "imaginary" axis (Figure 1.2a). The length of the arrow represents the magnitude of F_{hkl} and the phase is the angle between the vector and 0°. Since the structure factor is the sum of all the atomic scattering factors, in vector notation it is the sum of

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

Figure 1.1: The electron density equation
The Fourier transform used for calculating electron density (ρ) at any point (x,y,z). This equation requires three components: a position (an x,y,z calculated over all h,k,l coordinates), an intensity ($I_{hkl}=F_{hkl}^2$), and a phase (α_{hkl}). V is the volume of the unit cell. Two of the three can be calculated directly from a single crystal X-ray diffraction experiment.



A



B

Figure 1.2: The structure factor and phase problem
(A) The structure factor, F_{hkl}, shown in vector notation (triple lined arrow) on an Argand diagram, where the horizontal axis is the real axis and vertical the imaginary, as the sum of all the atomic scattering vectors in the unit cell. **(B)** In a single crystal X-ray diffraction experiment, the length of F_{hkl} is calculated ($I_{hkl}=F_{hkl}^2$) but the phase information is lost. This loss of phase information, known as the phase problem, is graphically represented as a circle on the Argand diagram with radius F_{hkl}.

all the individual atomic scattering vectors (Figure 1.2a). However, since the phase information is lost, only the length of the vector is determined and an infinite number of phase solutions exist (represented graphically as a circle with radius F_{hkl})(Figure 1.2b). Herein lays the inherent difficulty of every single crystal X-ray diffraction experiment. The crystallographer must discover a way to find the true phase from the infinite number of solutions. Several ways exist to accomplish this goal, but two of the most popular methods are multiple isomorphous replacement (MIR) (Crick, 1956, Harker, 1956) and multi-wavelength anomalous dispersion (MAD) (Hendrickson, 1985, Kahn *et al.*, 1985). Both involve incorporating a heavy atom, higher Z-number than carbon, nitrogen, oxygen, and sulfur, into the unit cell and measuring the resulting change of diffraction spot intensities.

1.2 Multiple isomorphous replacement

Until the 1990's, MIR was the most popular way to solve the phase problem. MIR involves taking native protein crystals and placing a crystal into a stabilizing solution in the presence of a heavy metal salt with the hope that the heavy metal would incorporate itself in an isomorphous manner, through coordination with solvent accessible amino acids and organized water, into the crystal via the solvent channels. Some of the most popular heavy metals used are salts of mercury, lead, gold, and platinum. A diffraction dataset of native protein in the absence of any heavy metal is collected and structure factors calculated, $F_{P(rotein)}$. Next, a heavy metal is soaked into the crystal and dataset collected. The structure factors from this dataset, $F_{PH(eavy\ atom)}$, are calculated and relate to the native dataset by the equation $F_{PH} = F_P + F_H$. By drawing the relationship in vector notation (Figure 1.3), we see that the native dataset circle with radius F_P, intersects the heavy atom dataset circle with radius F_{PH1} at two points. By collecting a second set

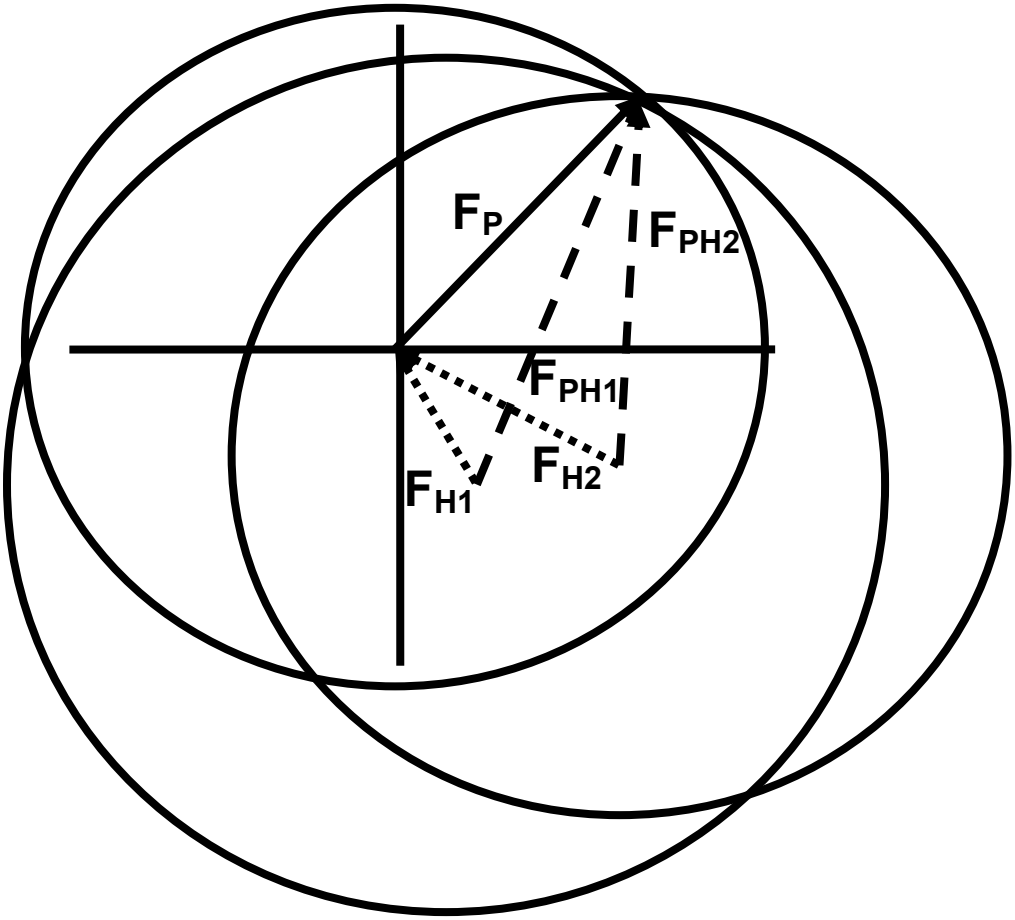


Figure 1.3: Multiple isomorphous replacement
Vector notation showing how MIR solves the phase problem. The heavy atom derivatives relate to the native dataset through the equation $F_{PH} = F_P + F_H$. The convergence of the native dataset, circle with radius F_P, with two separate heavy atom derivatives, circles with radius F_{PH1} and F_{PH2}, points to the correct phase solution. This phase is then used to calculate the electron density for model building.

of data, the phase problem has been reduced from an infinite number of solutions to two possible solutions. In order to distinguish which of these two phases is the correct one, a second heavy metal, binding to a different location, is soaked into the crystal. Just as before, the structure factors are calculated, F_{PH2}, and the native dataset intersects with the second derivative dataset in two places. Typically, the three circles converge at a single location revealing the correct phase value (Figure 1.3), solving the phase problem! Sometimes, the three circles do not converge to a single value, known as lack of closure error, and then it becomes a statistical probability function where the "most likely" phase is used to solve the phase problem. Now, an electron density map can be calculated and the protein model traced.

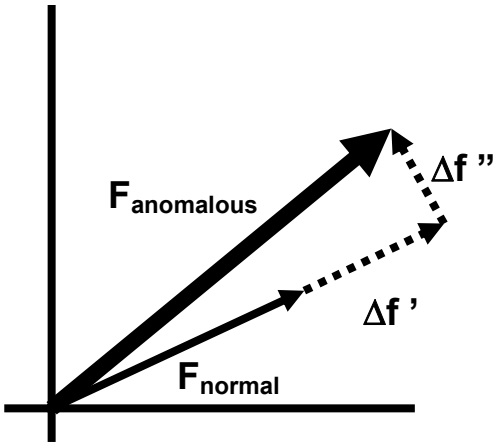
There are several concerns associated with MIR. In order for this technique to be successful, when the heavy atom soaks into the crystal the packing of the crystal can not be disrupted significantly. The unit cell is typically stretched slightly to account for the much larger heavy metal atom and often a small increase in the unit cell dimensions, an angstrom or two, is a quick way to check for possible heavy metal incorporation when screening crystals for diffraction. If the crystal lattice expands too much, the two crystals are no longer isomorphous with respect to each other and that derivative crystal can not be used in solving the phase problem. Getting the heavy metal atom into the crystal itself is another large concern because incorporation has a low success rate and even when successful often results in lower diffraction resolution. Many times, placing your crystal into a solution containing the heavy metal salt will cause the crystal to crack or even dissolve away completely. Just physically handling the crystal and moving it from the experiment in which it was grown to another drop with the heavy metal salt can damage the crystalline nature causing a loss of diffraction. Also, many of these salts are expensive and hazardous if not toxic. So not only is there a concern about even getting a heavy

atom into the crystal, there is a problem of the heavy metal incorporation changing the crystal to such an extent that the crystal is no longer usable. An ideal situation would be where the crystal remains 100% isomorphous but somehow the intensities could still be altered.

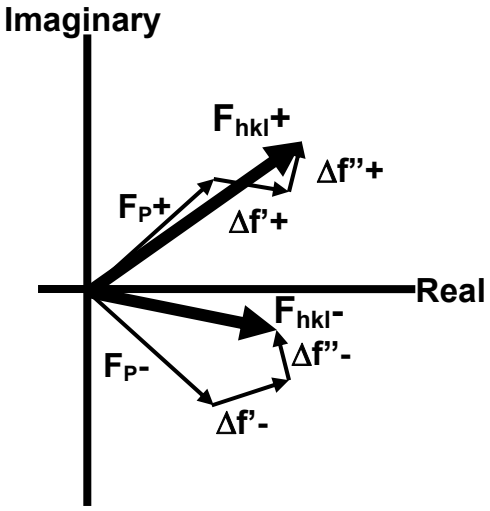
1.3 Anomalous scattering and multi-wavelength anomalous dispersion

In 1913, Friedel put forth the research 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 (Friedel, 1913). In 1949, Bivjoet used 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 (Bijvoet, 1949) and later envisioned this difference being used in solving isomorphous replacement for protein crystallography (Bijvoet, 1954). A vision that would be observed in diffraction patterns by Wyckoff and colleagues a few years later, from the iron atom associated with myoglobin, and referred to as Bijvoet differences (Kendrew, 1956). At the time, this deviation from Friedel's law was not expected and referred to as "anomalous" scattering. Anomalous scattering is a misnomer for a naturally occurring resonance phenomenon seen when the wavelength of the X-rays approaches the absorbance edge of an electron. Compared to fluorescence, where a photon is absorbed and re-emitted at a lower energy level, in anomalous scattering a photon is absorbed and instantly re-emitted at the same energy level, gaining an added real and imaginary component to its phase. The equation in Figure 1.4a 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, graphically illustrated in Figure 1.4b. It is the phase shift of the

A $\mathbf{F_{anomalous} = F_{normal} + \Delta f' + i\Delta f''}$



B



C

Figure 1.4: 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. Note that the imaginary component is always a positive 90° from the real component. **(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.

imaginary component that disrupts Friedel's law when dealing with two types of atoms; F^+ no longer has the same intensity, vector length, or phase, relative angle to the real axis, as F^- (Figure 1.4c). Since this affect is wavelength dependent, with a tunable X-ray radiation source it would be possible to alter the diffraction spot intensities without changing the contents of the unit cell, making the datasets 100% isomorphous.

In 1985, Hendrickson and Kahn *et al.* took this idea from the theoretical realm and successfully implemented it for solving proteins; paving the way for the next generation of phase solution, multi-wavelength anomalous dispersion (MAD) (Hendrickson, 1985) (Kahn *et al.*, 1985). However, the technique was limited to proteins naturally containing a metal atom whose edge was within tunable synchrotron X-ray radiation. Five years later, the first protein phased by selenium, a selenobiotinyl derivative, and the first selenomethionine derivative protein were both phased using MAD (Hendrickson *et al.*, 1990, Yang *et al.*, 1990, Yang *et al.*, 1990). Typically, in a MAD experiment, a crystal dependent graphs of $\Delta f'$ and $\Delta f''$ are generated (Figure 1.5) to determine the optimal wavelengths to use in the experiment. A minimum of three wavelengths are used: the inflection (greatest difference of $\Delta f'$ and $\Delta f''$), the peak (greatest value of $\Delta f''$), and a remote (minimal anomalous difference). In a situation analogous to MIR, the remote data is used as a "native" dataset and the inflection and peak datasets are "derivatives." The intensity values between the derivative and native are related through the anomalous scattering summation where F_{remote} is substituted for F_{normal} (since there is "no" anomalous signal in the data). By calculating the derivative structure factors and combing them with the calculated values of $\Delta f'$ and $\Delta f''$, we can draw a vector diagram similar to the MIR case (Figure 1.6). The remote structure factor circle intersects with the inflection and peak circles at a consensus phase solving the phase problem. The best part of this solution is that the difference in intensities came from

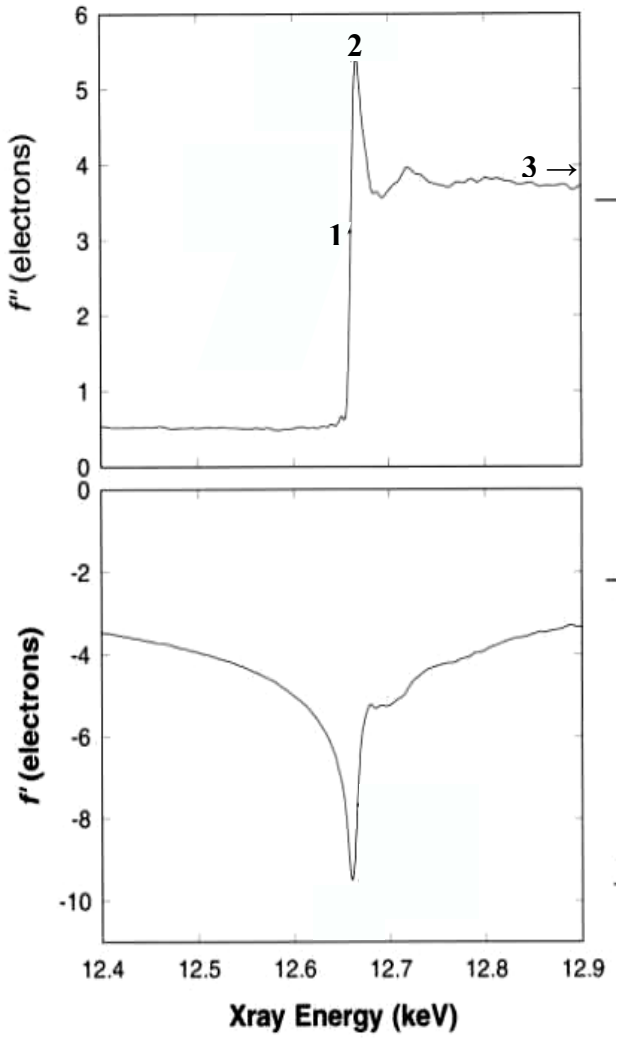


Figure 1.5: Graph of Δf' and Δf"
Graph of electrons versus energy showing the wavelength dependence of anomalous scattering for selenium. Point 1 represents the inflection point where the difference between Δf' and Δf" are the greatest, 2 the peak having the highest value of Δf", and 3 represents a remote peak at an energy level further upstream with "no" anomalous scattering. These graphs should be calculated on a crystal to crystal basis since the local chemical environment, while not altering the shape of the graph, will shift it either to the right or left.

Adapted from: Ramakrishnan, V and Biou, V. Methods in Enzymology Vol. 276 New York, Academic Press 1997

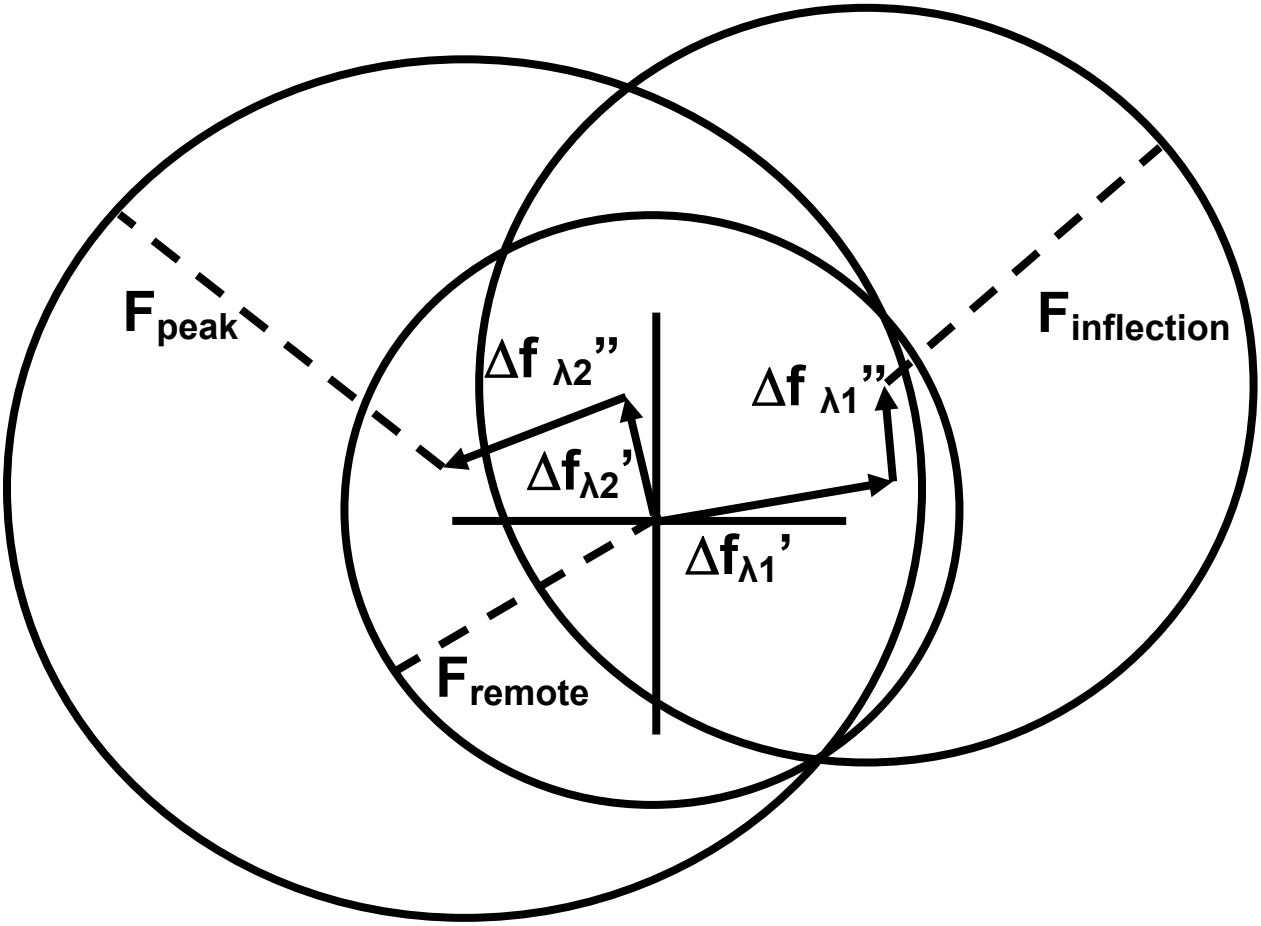


Figure 1.6: Phase solution by MAD:
Analogous to MIR, the solution of the phase problem involves the use of a "native" dataset, the remote, and data from two "derivatives," the inflection and peak. In a MAD experiment, the remote dataset, with "no" anomalous scattering serves as the native data and can be related to the peak and inflection datasets through the calculated Δf' and Δf" values by the anomalous scattering summation equation in Figure 1.4a. As in MIR, the three circles converge onto the true phase solution of the data.

Adapted from G. Rhodes Crystallography Made Crystal Clear 2nd ed. (2000) San Diego, Academic Press: p.122

substructures of heavy atoms that are 100% isomorphous with the "native" dataset typically with no loss of diffraction resolution.

There are some concerns associated with MAD and specifically selenomethionine derivative proteins. A MAD experiment requires a trip to the synchrotron and that can be costly. Selenomethionine derivatives are not straight forward and selnomethionine can be quite costly depending on the *E. coli* preparation volume needed for adequate expression. Often selenomethionine incorporation makes the protein less soluble, alters the crystallization conditions and diffraction limit of the crystals, and mixed oxidation states of the selenomethionine results in a reduction of anomalous signal.

The detractions associated with both MIR and MAD are not trivial and represent a significant investment of money and time altering either the contents of the unit cell or re-engineering the recombinantly expressed protein itself. An optimum solution would be using atoms inherent in all proteins with a single wavelength data collection strategy to overcome the phase problem. The "heaviest" of the naturally occurring atoms in the amino acids is sulfur, occurring in cysteine and methionine. Unfortunately, the absorption edge of sulfur is around 2475eV (wavelength of 5.02Å) (Table 1) which, while attainable at a synchrotron, is not practical due to the dramatic loss of beam intensity, severe absorption affects, and air scattering. However, this is not to say that phasing a protein with the anomalous signal from sulfur using a single wavelength X-ray diffraction experiment is impossible.

1.4 History of sulfur phasing

The first macromolecule phased with the anomalous signal from sulfur was the 45 amino acid polypeptide crambin (Hendrickson, 1981). Hendrickson and Teeter introduced the concept

Element	Absorption Edge Wavelength (Å)	Δf" at λ=1.54Å (e^-)	Δf" at λ=2.29Å (e^-)
S	5.02	0.54	1.12
Se	0.97	1.14	2.52
Ca	3.07	1.2	2.95
Zn	1.28	0.74	1.44
Fe	1.74	3.33	0.75

Table 1: Anomalous scattering comparison of selected elements

of resolved anomalous phasing, a statistical method that owes its origins to Ramachandran's quasi-anomalous phasing a decade earlier (Ramachandran, 1970). Both of these methods are statistical measures of the diffraction intensities where the anomalous scattering must contribute significantly to the magnitude of the overall structure factor so that the phase of the anomalous scatterers is very close to the resulting phase of the structure factor. In the case of crambin, there were 6 sulfur atoms per 45 amino acids and a measurable contribution was obtained. However, this ratio of sulfur to amino acids is an unlikely situation for larger proteins, and in fact the statistical method will only resolve a portion of the structure. This method is the basis for Hendrickson's MAD experiment but, since the anomalous scattering of selenium is much higher than that of sulfur, the anomalous scatter substructure allows for larger proteins to be phased accurately. Four years later, B. C. Wang developed a method where solving the phase problem with anomalous scattering became independent of the percent contribution of the anomalous scatterers to the overall intensity value (Wang, 1985).

1.5 Iterative single-wavelength anomalous scattering

The innovation in Wang's design was using a low resolution image of the entire protein molecule to improve the initial phases calculated by the anomalous scatterer substructure. The only improvement in Hendrickson's method came from refining the heavy atom positions, whereas Wang's method uses the inherent difference of protein and solvent density in reciprocal space to then inverse Fourier transforms back to generate calculated phases from the model. Figure 1.7 show a simplified flowchart of the Wang method. Initial phases from the data are calculated and Fourier transformed to produce electron density. A summation of density around a point is calculated to outline the boundary between protein density, large continuous positive values, and solvent, small sporadic positive and negative values. This summation is used as a

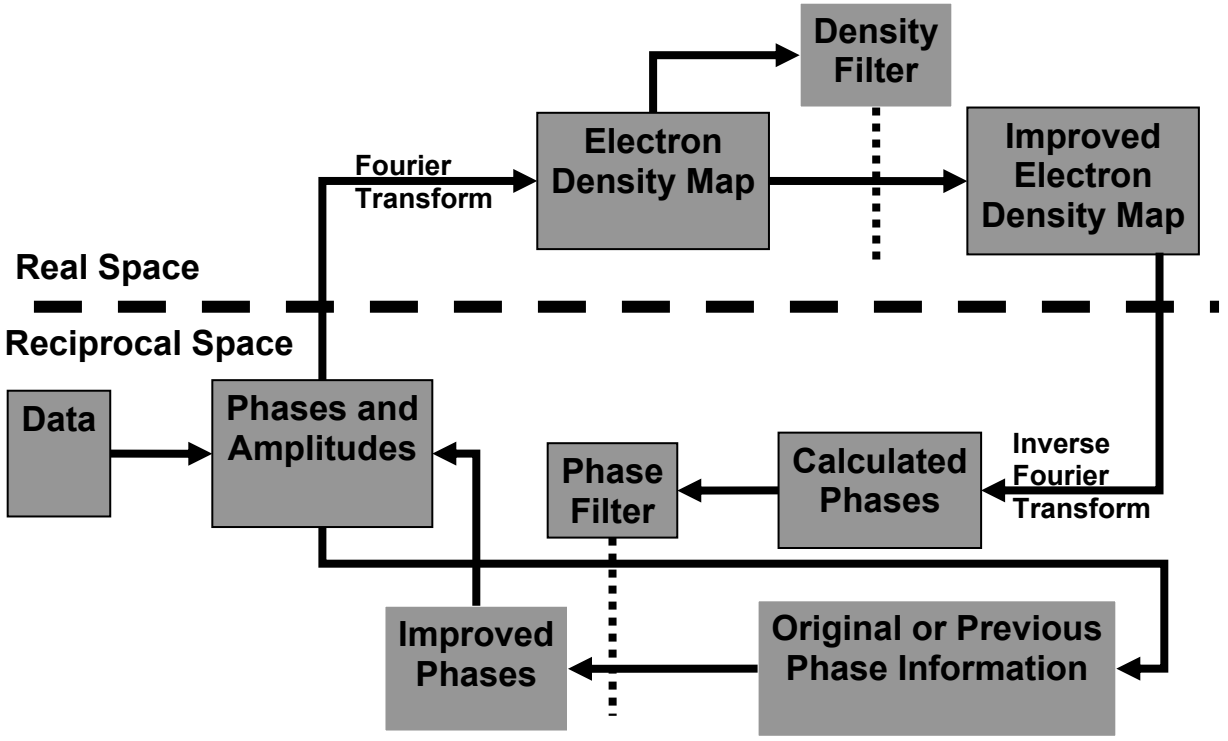


Figure 1.7 I-SAS Flowchart
Phases are calculated from the original data and Fourier transformed into electron density. A density filter is employed to improve the electron density which is then inverse Fourier transformed to calculated phases. The phases are combined with the starting value and the improved phases are used to calculate electron density for the next cycle and so on. Phase improvement continues for 4 rounds and the final density filter is then used with the original phases from the data, as to not bias the phase result, and the process starts up again for another 4 rounds. This continues for 20 rounds total when the final improved phase is used to calculate the final electron density map.

Adapted from: Wang, B. C. (1985). Methods Enzymol 115, 90-112

mask and then the electron density outside of the mask, presumably only solvent, is flattened to 0 leaving only protein electron density. This improved electron density map is then inverse Fourier transformed back to calculated phases. A phase filter then combines the original phases and the calculated phases through simple phase averaging to give improved phases. The improved phases are then cycled back with the structure factors from the data and the process cycles through again. This iterative cycling between real and reciprocal space occurs four times, and at the end of the fourth cycle, the electron density filter is saved and the process starts back up with the original phases and original amplitudes, as to not over bias the phases towards the inverse Fourier calculated phase value. This continues on for another four rounds and then that final density filter is used with the original phases and amplitudes. The density filter from the second cycling is used for a final four iterations to optimize the phase and the last step is three rounds of phase extension. The final phases, after phase extension, are used to calculate the final electron density map. It is this simple cycling between real and reciprocal space that give iterative single-wavelength anomalous scattering (I-SAS) its name.

At the time, this method was designed to phase a protein using the anomalous scattering from any atom, but within the same work, Wang presented a computer simulation of Bence Jones protein Rhe, a 12 kilodalton (KDa) protein, to show that the structure could be phased using just the sulfur anomalous signal from the 2 cysteine amino acids in the protein using 8067eV X-rays (1.54Å λ). While Wang hypothesized that it would theoretically be possible, he also realized that the equipment of the time was not ready to accurately measure the extremely small, 0.54e^-, difference. He prophesized that one day, the use of the anomalous signal in sulfur would be used to routinely solve the phase problem. For fifteen years, this prophecy lay dormant until Liu *et al.* solved the *de novo* structure of the 22KDa protein Obelin from the 8 sulfur atoms

using 8067eV X-rays from a synchrotron (2000). To date, 10 more *de novo* structures of moderate size have been solved using sulfur's anomalous signal from different X-ray sources (Table 1.2). While not a routine phasing method for every protein crystallography project yet, recent advances in technology and method have now allowed the crystallographer to seriously consider sulfur-SAS as a viable choice for solving the phase problem. There are two ways to better your chances for a successful outcome with sulfur-SAS, increase the signal and reduce the noise in the dataset. Since sulfur's Δf" is so small, the crystallographer really needs a way to accurate assess the amount of anomalous signal within a dataset and, if possible, know when that signal has disappeared.

1.6 Maximizing and monitoring sulfur's anomalous signal

The development of a new anomalous signal statistical measure was put forth by Fu ***et al.*** called R_{as} (2004). As we saw before, the anomalous signal is measured as the difference between the intensities of positive and negative (h,k,l) breaking Friedel's law. However, there is a spacegroup specific qualifier to that. Depending on the spacegroup, certain reflections will always have a phase of either 0° or 180°, meaning that the structure factor will always lie on the real axis and contain no imaginary component. These reflections, known as centric reflections, can not have any anomalous signal because, by definition, anomalous scattering adds an extra imaginary component. Therefore, the anomalous signal must come from reflections whose structure factors contain an imaginary component and do not lie on the real axis, acentric reflections. Even in the presence of anomalous scattering, the intensities of the Friedel pairs of centric reflections should always be the same. However, in a real world situation, there are always small differences between the Friedel pairs of centric reflections, and that difference represents the baseline systematic error associated with that particular experiment. So, in an

Protein	Mol. Weight (KDa)	# of sulfurs	X-ray source	PDB ID	Reference
Obelin	22	8	Synchrotron	1EL4	(Liu *et al.*, 2000)
Apocrustacyanin	20.7	8	Synchrotron	1I4U	(Gordon *et al.*, 2001)
C. elegans 11C1	14	4	Synchrotron	1LPL	(Li *et al.*, 2002)
Tryparedoxin	17.1	7	Synchrotron	1O6J	(Micossi *et al.*, 2002)
E. coli Argininosuccinate Synthetase	50.9	19	Cu-rotating Anode	1K92	(Lemke *et al.*, 2002)
Lima Bean Trypsin Inhibitor	8.8	14	Cu-rotating Anode	1H34	(Debreczeni *et al.*, 2003)
Hum-IGF2R Domain 11	15.5	11	Synchrotron	1GP0	(Brown *et al.*, 2002)
Pfu-1801964	34	9	Cr-rotating Anode	1NNH	To be published
Hum-15691	30	9	Cr-rotating Anode	1VKA	To be published
Sso-10a	11.1	6	Cr-rotating Anode	1R7J	(Chen *et al.*, 2004)
S. cerevisiae Oxygen-dependent CPO	37.5	10	Cr-rotating Anode	1TKL	(Phillips *et al.*, 2004)
Pfu-542154	16.8	3	Cr-rotating Anode	1ZD0	To be published

Table 1.2: *De novo* structures solved using sulfur anomalous scattering

anomalous scattering situation, the Bijvoet differences represent the actual anomalous signal, and the difference in the centric Friedel pairs reflects the noise. R_{as} equals the difference between the positive and negative (h,k,l)'s of the acentric reflections divided by the positive and negative (h,k,l)'s of the centric reflections (signal divided by noise). Since all current scaling programs in data processing force the centric intensities to be identical (which they should be in an ideal world), Fu has developed a program, 3DSCALE, that keeps all intensities separate and reports the R_{as} statistic. With this statistic, it would be possible to process and scale a portion of your data, with the crystal still mounted and collecting, to see if you are gaining anything in terms of anomalous signal, or if enough radiation damage or other error has washed out the anomalous signal to the extent that all you are doing by adding more data is adding noise.

While this new statistical approach will help reduce the noise introduced into the dataset, there are simple things the crystallographer can do to get as much anomalous signal out of the crystal as possible. The working energy ranges of X-rays generated from a copper rotating anode generator, 8067eV (1.54Å wavelength), and synchrotrons, many optimized for about 12424eV (1.0Å wavelength), are very much below the absorption edge of sulfur at 2484eV (5Å wavelength). Currently, it's not practical to tune a synchrotron beamline to 2484eV because of the dramatic loss of beam intensity and air absorption associated with the wavelength. However, it is in the best interest of the crystallographer to go to longer wavelengths to increase the sulfur anomalous signal as much as possible. Home rotating anode generators can now be fitted with a chromium rotating anode that produces 5425eV X-rays (2.29Å wavelength). However, the researcher must overcome lower beam intensity compared to a synchrotron and the tremendous amount of air scattering at this wavelength. More intense generators and better optics are constantly being produced to maximize the X-ray beam intensity of a home generator. For a

synchrotron, research by Liu *et al.* indicates that a wavelength of 1.74-1.8Å (7140-6902eV) is the best compromise of synchrotron beam intensity and sulfur anomalous signal (Liu, 2004). That same research also suggested a better data collection strategy for SAS experiments at the synchrotron. When using a CCD detector, the norm at a synchrotron, it is common to overload the detector, above ~65,000 photons and displayed as a red pixel, for one or two diffraction spots. This is a good indicator that the correct exposure time and beam intensity combination for that particular oscillation range has been achieved. However, Liu *et al.* suggest that it is better to cut the most intense spot value in half (~32,000-35,000 photons) and instead of collecting the highest resolution data possible for that crystal, collect more data while incorporating less radiation damage. Traditionally, we have always tried for the highest intensity data possible and in fact more intense X-ray sources give better intensity statistics (I_{hkl}/σ, the signal to noise ratio). This is one of the pros with using very intense X-ray beams at synchrotrons versus less intense rotating anode X-ray generators. Though, by using more intense X-ray beams, we introduce radiation damage to the crystal more quickly, but this has never been a problem before because phasing was never based off of such a small anomalous signal. In many cases, phasing a structure with sulfur-SAS requires re-learning every step of the structure solution process because the anomalous signal is so small, that even small introductions of error from a source as mundane as having too much liquid in the nylon loop, when mounting your crystal from the crystallization condition, can cause your experiment to be unsuccessful. In the end, refining and perfecting techniques to minimize systematic error is inevitable even with maximizing the anomalous signal of sulfur by going to longer wavelengths or more intense X-ray sources.

1.7: Questions to be addressed

Using R_{as} as a guideline, this research hopes to examine the minimum threshold value needed for successful phasing of a protein structure. In this manner, a crystallographer could use this benchmark and constantly monitor the data and its approach to that value. Along the same lines, structure solution from a single crystal may not be feasible in every case. We also want to verify that merging together data from multiple crystals has an additive affect on R_{as}. Again, the minimum value of R_{as} can be used to verify a successful phasing outcome even in te merged case. Finally, sulfur-SAS has been successful on smaller proteins with a handful of sulfur sites. We'd like to expand the current applicability of sulfur-SAS by showing that it can phase a large (>100KDa) protein with many (>20) sulfur sites. By finding the current limitations of sulfur-SAS, we can see the advancements and improvements in the method as better X-ray sources and detectors come about.

1.8: Significance of this work

The upside of this research is its far reaching ability in the field of protein crystallography. The results do not affect or enhance the research of a small sector or subset of the community, but apply to every member. Every crystallographer will benefit as sulfur-SAS phasing matures from a rare exception to common place solution. Sulfur-SAS represents the easiest method of *de novo* phasing in that only data from a native crystal(s) needs to be collected. There is minimal physical handling of the protein crystal, no need to soak in heavy metal salts, and no need to re-express the protein in an engineered form. All you have to do is mount and shoot. That's it.

More importantly, this research shows the value of the oft neglected home source X-ray generators. For many, their rotating anode generator is just a tool for screening crystals before

sending them to the synchrotron. By looking at the most recent structures solved from Table 2, we start to see the advantages of the rotating anode for structure solution over a synchrotron in the area of sulfur-SAS. The Liu *et al.* research shows that intensity and diffraction limit are not the most important factors in a data collection, signal is. In this work we help refine how we look at data for anomalous signal by using a new statistical measure. A measure based solely on the data itself; that is detectable at any point in a single data collection or as the end result of merging data collections together.

The first portion of this research focuses on expanding the current understanding of the limitations of sulfur-SAS through computer simulation. Just as the first protein to be "solved" with sulfur-SAS was via simulation, almost twenty years later simulation is being used again to show the value and power of this phasing technique. Also, this simulation helps establish a base level of sulfur signal that much be achieved before realistically attempting structure solution through the use of a new statistic, R_{as}. This work was separately and simultaneously carried out at the same time as the Fu *et al.* research and independently came to a similar conclusion.

The second portion of this research expands the real world application of sulfur-SAS to the doorstep set forth by the original Wang simulation from 1985. Now that this level of sulfur to amino acid ratio has been successfully shown, further real world expansion towards the large protein simulation presented in this work can be started with hope of completion. We are now stepping into the realm of possibility only dreamt about a mere twenty years ago. We have reached this goal and reset the bar at a much higher level. Sulfur-SAS method development is an ongoing process, but with each successive generation, greater and greater accomplishments await.

Chapter 2

Sulfur-SAS Simulation: *Clostridium botulinum* Neurotoxin Type B

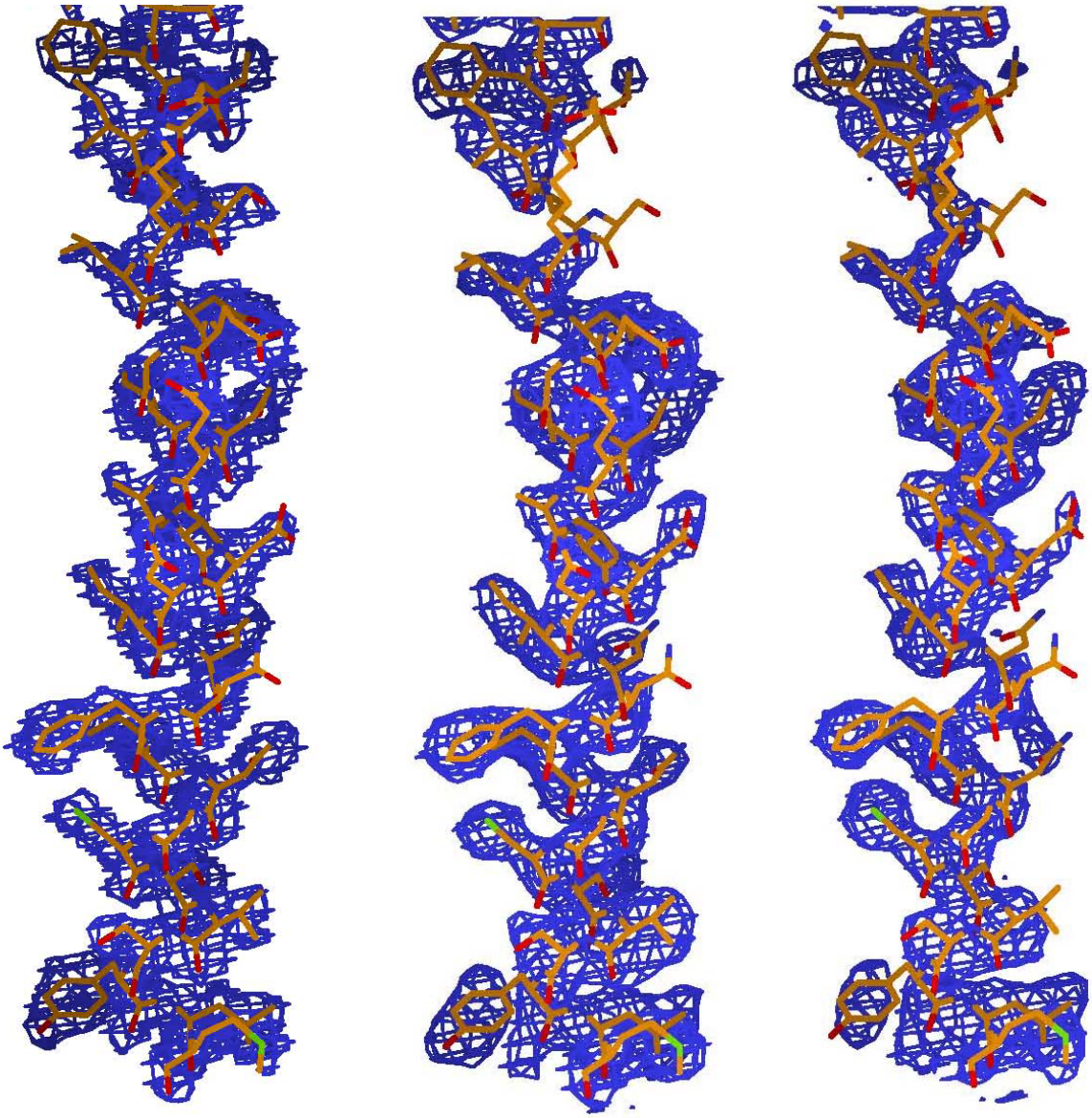
While the peptide crambin was originally solved with sulfur-SAS in the early 1980's, the idea of protein phasing with this method did not come about until a simulation study by Wang in 1985. The power of this simulation showed an idea that no one thought possible at the time. Simulations have the advantage of being as easy or as complex as we make them. The more we make them look like a real life situation, the more complex they become. At the time of this research in 2002, *de novo* sulfur-SAS had been realized but only on a small protein scale (<30KDa) and for the situation where a handful of sulfur positions existed. The question arose as to whether sulfur-SAS could indeed be used on any protein large or small and with a large number of sulfur positions. The easiest answer of this question came about in the re-birth of the sulfur-SAS simulation. This simulation also identifies the percentage of sulfur positions needed for an interpretable electron density map, and demonstrates the power of redundancy to rescue sulfur sites from data with introduced random error.

2.1: Error-free data simulation

Target selection for the simulation was centered on a commercially available protein whose structure was already in the PDB. By paging through the Sigma-Aldrich catalog online and pulling out every protein available, and then cross referencing each of those with a PDB search, *Clostridium botulinum* neurotoxin Type B (PDB ID 1EPW), solved by Swaminathan and Eswaramoorthy (2000), was clearly shown to be the best target. *C. bot.* neurotoxin type B is a single chain protein with a molecular weight of 150KDa and 32 sulfur atoms (10 cysteines and

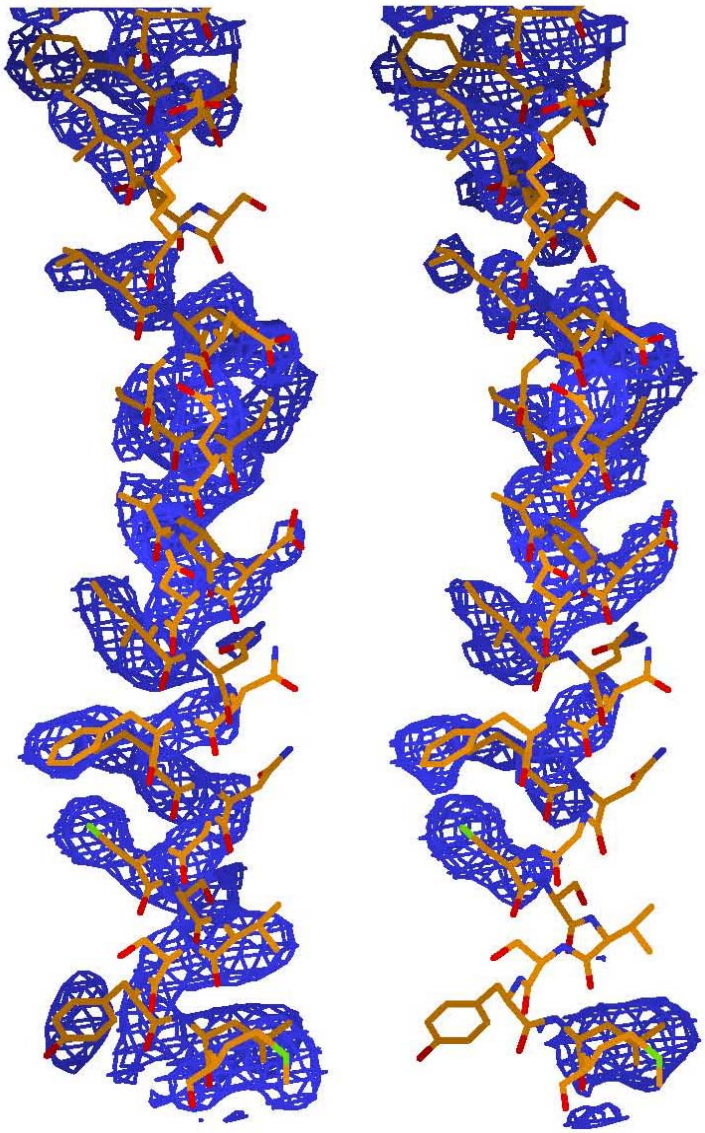
22 methionines), and represents one of the largest single chain proteins structures solved in the PDB. More importantly, the structure contains a single zinc atom. Since zinc's anomalous scattering is only slightly larger than that of sulfurs (0.74 electrons at 8067eV), it allows for a direct comparison between the simulation of calculated ideal structure factors and data collected from a diffraction experiment.

The file 1EPW was downloaded from the Protein Data Bank. All comments, water, and other hetero atoms were removed from the file (from here on referred to as 1epw). Ideal data was calculated with FCAL (Wang, 1983) using 1epw. The resulting calculated structure factors from FCAL were reformatted to a DENZO/HKL style (.sca) file and read into XPREP (Schneider & Sheldrick, 2002). A 3.0Å resolution single wavelength anomalous scattering ΔF (.hkl) file and instruction (.ins) file for XM (Otwinowski, 1997, Schneider & Sheldrick, 2002) were generated searching for 32 sites. XM was run at a resolution range of 99.0 to 3.0Å and the number of tries set to 50. XM successfully located 31 heavy atom sites; a single disulfide bond is seen as a single peak at a resolution of 3.0 Å. The resulting peak list (.lst) file and the sulfur positions from 1epw were compared using MOLEMAN (Kleywegt, 1992-2001) in order to assign individual sulfur positions from 1epw with the peaks in the .lst file. A sulfur position list file (.xyz) for ISAS2001 was created using the sulfur locations from 1epw (the average x,y,z of cys 436 and 445 were used to model the center of the disulfide bond). Electron density maps were calculated using ISAS2001 and visualized with XFIT (McRee, 1999), with each map loaded in as F_o*f.o.m. The initial question of whether sulfur-ISAS would even work on a large protein was answered immediately with the production of an interpretable electron density map (Figure 2.1a 0% removed). Not being able to find every sulfur position the first time through is a



A

0% Removed **10% Removed** **20% Removed**



30% Removed 40% Removed

B

Sulfur Sites Removed	Map Correlation Coefficient
10%	0.895
20%	0.819
30%	0.792
40%	0.666
50%	0.551
60%	0.494
70%	0.347
80%	0.282
90%	0.186

Figure 2.1: Sulfur site removal and its affect on electron density
(A) With ideal data, the electron density maps from ISAS2001 show connected density throughout the structure with all sulfur sites located (0% Removed). In order to assess what percentage of sulfur sites are necessary for an interpretable electron density map, ten percent of the weakest sulfur sites (3 sites) were removed from the bottom of the list file used in ISAS2001. Between removal of 30 and 40% of the weakest sulfur sites the electron density can no longer be interpreted. Images created with XFIT. **(B)** The map files were reformatted, read into MAPMAN, and map correlation coefficients compared to the "0% Removed" map were calculated for all of the maps where sulfur sites had been removed. Note the large decrease in the coefficient between removal of 30 and 40% of the weakest sulfur sites verifying what was seen visually.

very common occurrence and that's why it is important to identify the amount of total sulfur positions needed for interpretable maps. Since the ideal case of finding every sulfur site was successful, how many sulfur sites do you need to find in order to have an interpretable map?

The corresponding three weakest (lowest three) sulfur positions from the XM .lst file (10% of the total number) were removed from the .xyz file and the file was renamed and saved. The three next weakest positions were removed and the file was renamed and saved and so on until only three positions remained. All of the .xyz files were run through ISAS2001 and the resulting log and map files were renamed and saved. Electron density map correlation coefficients compared to the "0% Removed" map were calculated using MAPMAN (Kleywegt, 1996). As would be expected, as sulfur sites are removed the electron density degrades. The transition to a map that is not interpretable visually occurs between 30% and 40% removed (Figure 2.2a.) and is verified statistically by the large decrease in map correlation coefficient (Figure 2.2b.). This simulation with ideal data suggests that in order to be able to trace the main chain of your model into the electron density, the crystallographer must find positions of at least 70% of the sulfur sites; which is a good starting point for sulfur-ISAS phasing since the sequence is known in almost all cases. Though, this only takes into account location of sites by heavy atom searches and not locating more sites by anomalous difference Patterson search or anomalous difference Fourier. This does give the crystallographer an idea of how many of the sulfur positions within their own protein must be identified in order for interpretable electron density map calculation. However, the data collected in a single crystal diffraction experiment is never error-free. The next step in making this simulation more like its real life counterpart is to incorporate error into the ideal structure factor calculations and see if we can overcome this error by "collecting" more data.

2.2: Introducing Gaussian error

If you try to incorporate true random error into your structure factor calculations, what you typically end up with is a flat and even distribution in the error percentage around the ideal number. While this does introduce error into the calculations, it does not truly reflect the situation of a diffraction spot. In an error-free situation, a diffraction spot would be a single point spike. But, with error the spike widens in all three dimensions and when cross-sectioned looks like a Gaussian curve. Therefore, in order to make the simulation more realistic, it would be better for the error introduced into the structure factor calculations to be Gaussian. An extensive search of the literature finally revealed the Box-Muller transformation (Figure 2.2) (Box, 1958) which was incorporated into the program RNDME (Liu, 2002). The ideal data file was first run through TAB (Liu, 2002) to generate a protein specific scale factor table needed for input into RNDME. Figure 2.3 shows 4700 independent ran_{gauss} calculations (F_{cal} =1 and σ =1) from RNDME verifying the Gaussian distribution of error. The ideal data file, scale factor, and error percentage were fed into RNDME and an erred data file produced. As a control, the amount of random error introduced into the ideal data set was independently evaluated using the data processing program SCALEPACK (Otwinowski, 1997)by an R_{merge} comparison between the erred and error-free data sets ($R_{merge} = \Sigma_{all\ hkl} \mid |F(hkl)|_{erred} - |F(hkl)|_{ideal}| / \Sigma_{all\ hkl} |F(hkl)|_{erred}$).

2.3: The effects of redundancy overcoming error

Data sets with 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0% error were generated with redundancy values of 1, 2, 4, 8, 16, 32, 64, and 128. The redundancy represents that number of

μ_1= ran (c) n^{th}
μ_2= ran (c) $(n+1)^{th}$

$$ran_{gaus} = \sigma \cdot (-2\ln\mu_1)^{1/2} \cdot (\cos(2\pi\mu_2))$$

$$F_{error} = F_{cal} + ran_{gaus}$$

Figure 2.2: The Box-Muller transformation
Random error is added to the ideal structure factors using the Box-Muller transformation (Box, 1958). In this equation, μ_1 and μ_2 are randomly generated numbers, independent of each other, and σ is one standard deviation from the mean of the Gaussian peak.

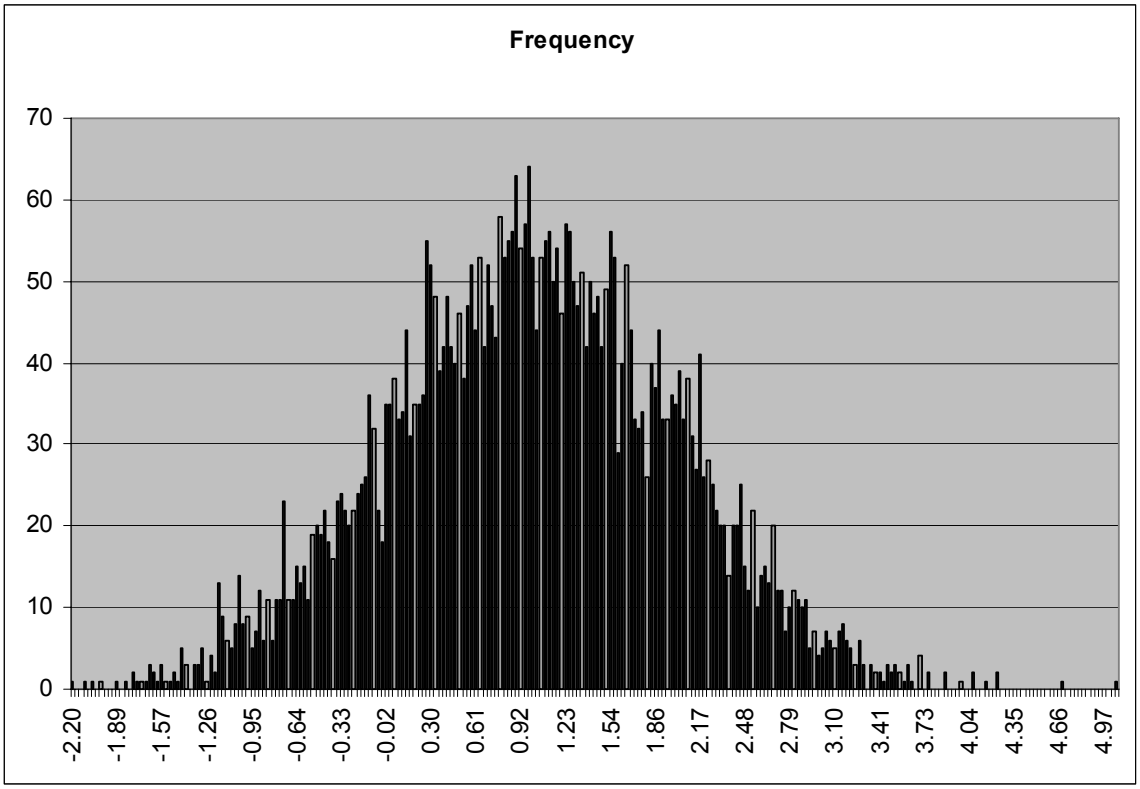
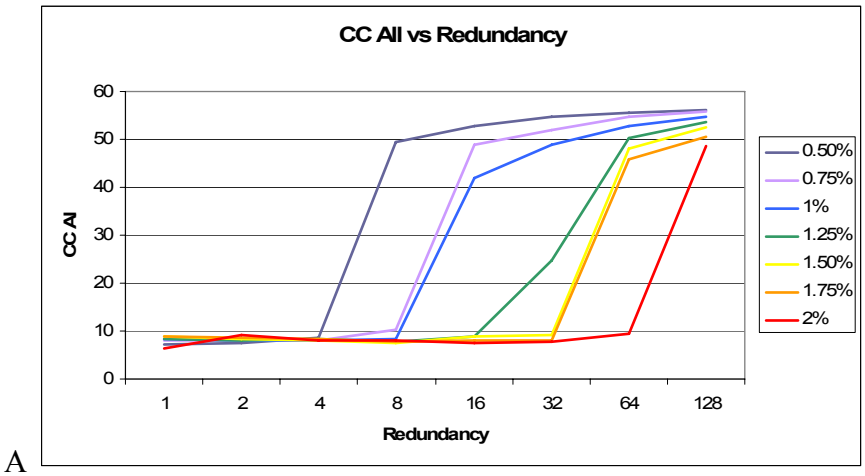


Figure 2.3: Gaussian error from RNDME
Distribution of 4700 independent ran_{gauss} calculations where $F_{cal}=1$ and $\sigma=1$ confirming Gaussian distribution.

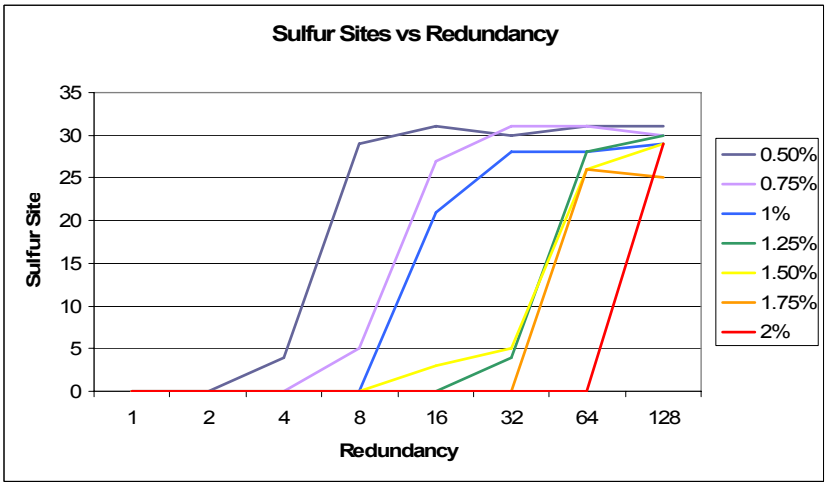
copies of the ideal structure factors of an asymmetric unit, with the appropriate error applied, scaled down to a single copy of the asymmetric unit. Each combination of error and redundancy was run through XPREP and peak searching was performed by XM. In evaluating the resulting XM list files, a sulfur site was thrown out if it contained a Patterson minimum function of 0.0. Graphs of the XM correlation coefficient for all peaks, the number of sulfur sites located, and electron density map correlation coefficients calculated using MAPMAN graphed versus redundancy (Figure 2.4a-c) demonstrate, as would be expected, that as error increases more data, in the form of redundancy, is needed in order to locate the sulfur sites and improve the resulting electron density maps. For electron density map calculation by ISAS2001, the error introduced data was paired with all 31 sulfur sites and, maps generated and renamed. Figure 2.5 shows the improvement of the electron density maps of a 1% random error dataset as more data is added through redundancy. While no initial threshold values can be directly extrapolated from the peak searching data in the random error case, this data provides a valuable insight if we combine the peak searching data with the R_{as} statistic.

2.4: Minimum R_{as} threshold for structure solution

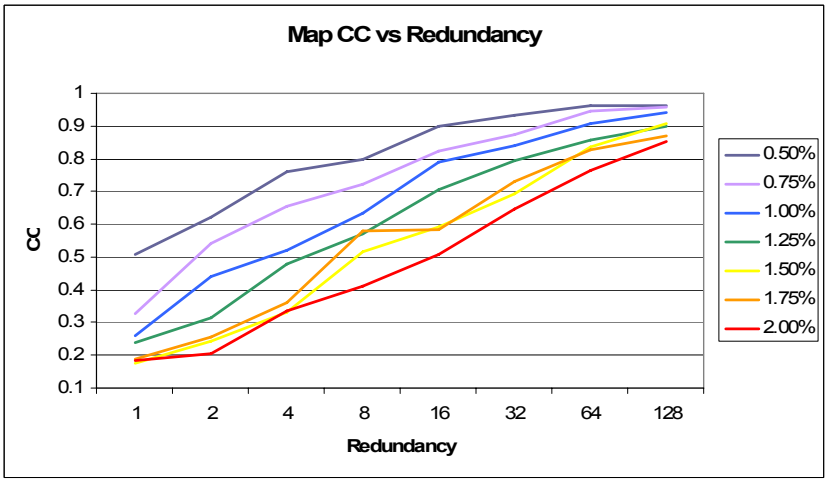
As discussed in section 1.6, R_{as} measures the anomalous signal to noise ratio by taking the F_{hkl}+ and F_{hkl}- difference in the acentric reflections of a dataset divided by difference in the centric reflections. The program RNDME was used to calculate the R_{as} of the ideal data with random error added. Figure 2.6a shows that, by graphing R_{as} versus redundancy, taking the redundancy value from Figure 2.4b necessary to locate ≥70% of the sulfur sites, and dropping a line from that redundancy point to the y-axis (R_{as}), a clustering of lines can be seen around the R_{as} value equal to about 1.6. Fu and co-workers have demonstrated similar results obtained from in-house data collected on cubic insulin. The same thing can be done with the graph of electron



A

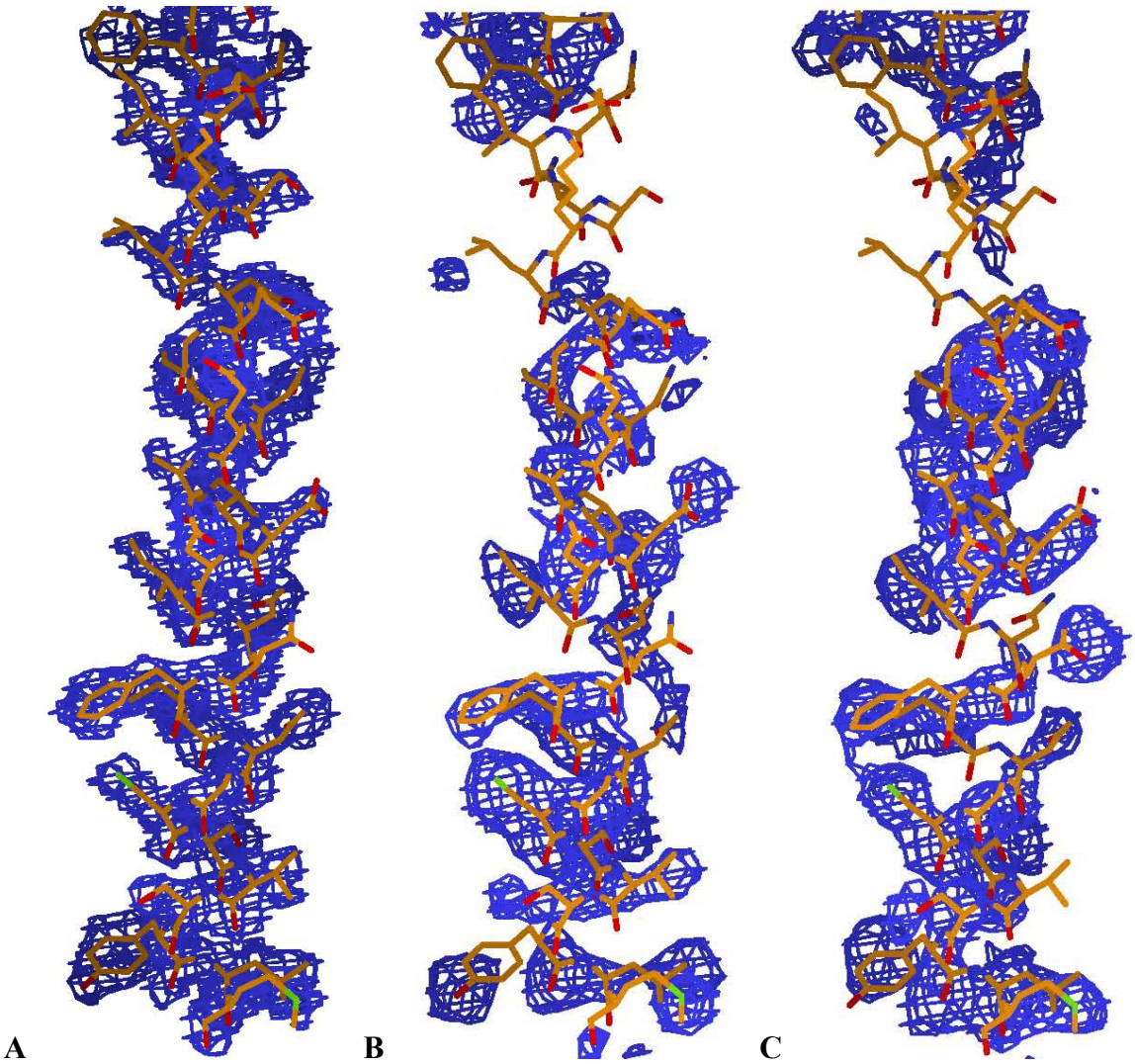


B



C

Figure 2.4: The affect of redundancy on rescuing erred data
(A) Graph of XM correlation coefficient for all peaks versus redundancy. **(B)** Graph of sulfur positions located versus redundancy. **(C)** Graph of MAPMAN electron density map correlation coefficient versus redundancy. As would be expected, more redundancy is needed as more error is introduced to determine sulfur site location and improve correlation statistics.



A **B** **C**

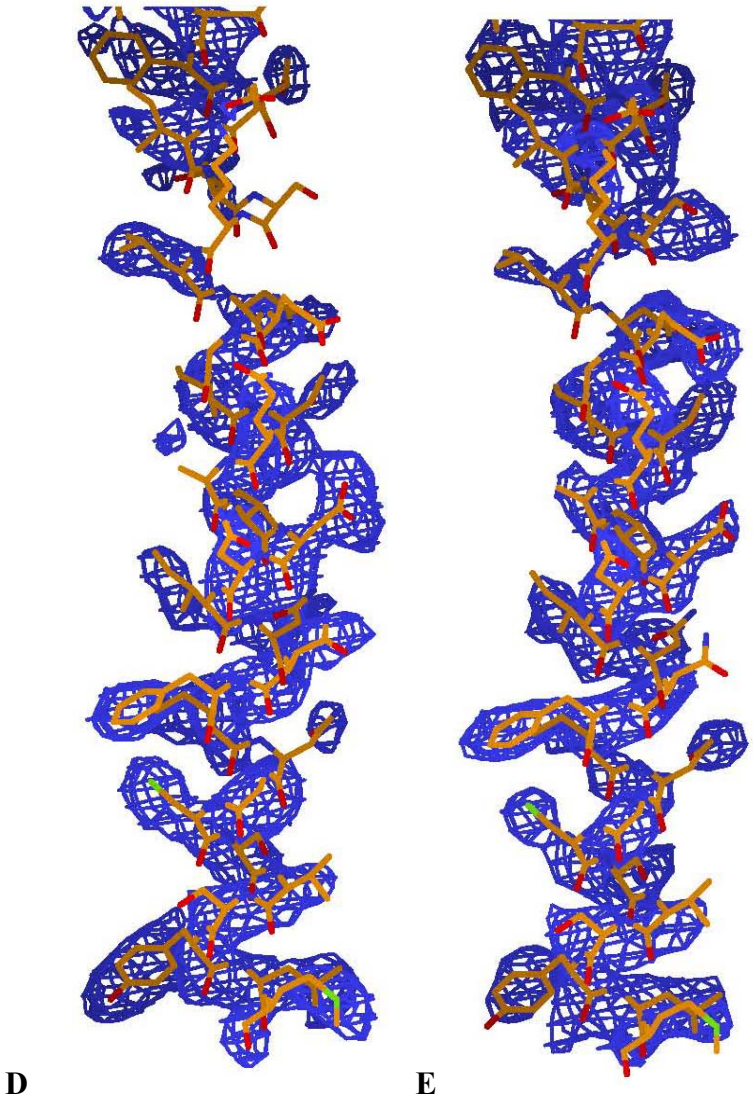
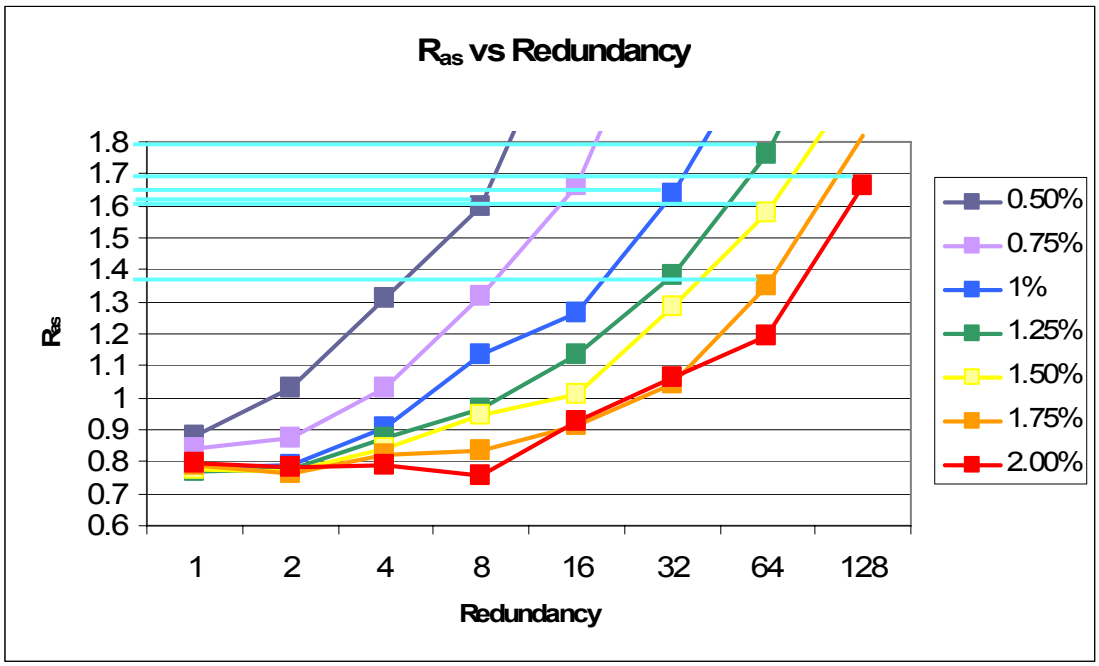
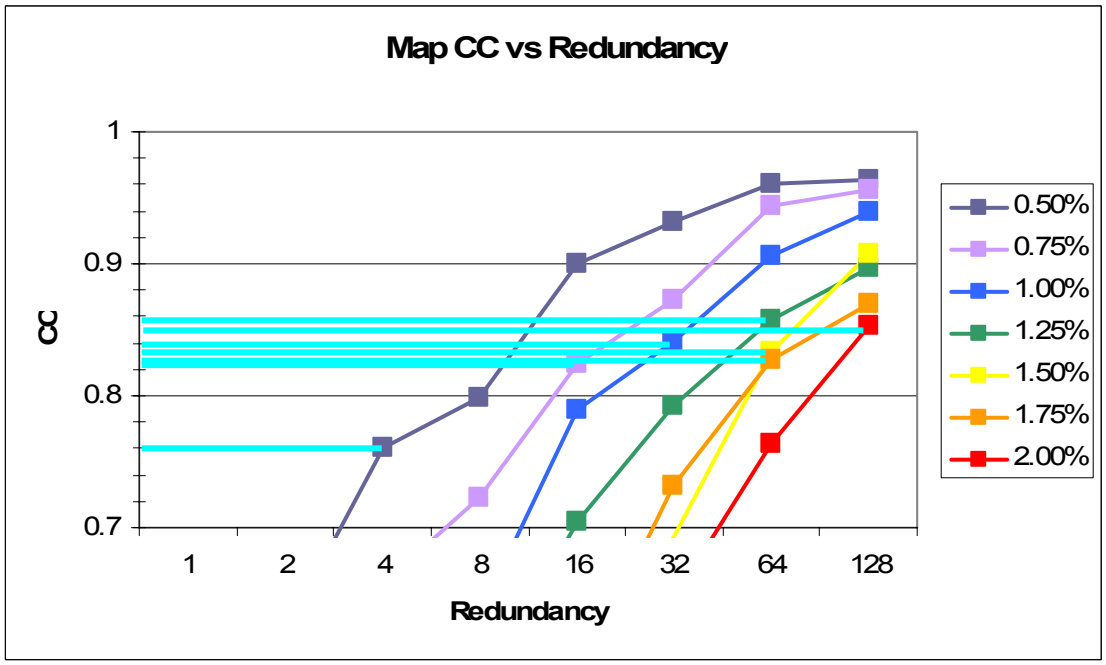


Figure 2.5: Electron density of data with 1% error and the affect of redundancy
Electron density maps of residues 749-783 (helix 19) visualized with XFIT. **(A)** Error free data. **(B)** 1% Error with a redundancy of 4. **(C)** 1% error with a redundancy of 8. **(D)** 1% error with a redundancy of 16. **(E)** 1% error with a redundancy of 32. At a redundancy of 32, the electron density is completely restored back to the error free control in (A).



A



B

Figure 2.6: Determination of R_{as} threshold value
(A) Graph of R_{as} versus redundancy. The redundancy values corresponding to locating >70% of the sulfur sites (Figure 2.4b) have lines drawn to the y-axis resulting in a clustering around 1.6. **(B)** Magnified portion of the electron density map correlation coefficient versus redundancy graph (Figure 2.4c). The same redundancy values from 2.6a have lines drawn to the y-axis resulting in a clustering around 0.83.

density map correlation coefficient (Figure 2.6B). Again, by graphing map correlation coefficients versus redundancy and dropping lines from the same redundancies to the y-axis (map correlation coefficient) we see a clustering around the value of 0.83. Recall, that the map correlation coefficient where 20% of the sulfur positions had been removed was 0.819. The correlation coefficient comparison also supports the idea that a R_{as} of 1.6 is enough to generate interpretable electron density maps to solve a protein's structure from a sulfur-SAS diffraction experiment. With this information in hand, the next logical step was to take the simulation results and use them as a guide to solve the *C. bot.* neurotoxin type B structure with an actual single crystal X-ray diffraction experiment using sulfur-SAS.

2.5: Comparison to synchrotron data

The original *C. bot.* neurotoxin type B structure was solved by Swaminathan's group at Brookhaven National Lab using the synchrotron housed on the grounds. This present and excellent opportunity with the synchrotron in such close proximity that a collaboration with the Swaminathan lab was established. Crystals were grown and mounted at Brookhaven National Lab. Since *C. bot.* neurotoxin Type B is known as the world's most poisonous poison, handling of the crystals had to be done by trained staff at Brookhaven. A 1.7Å wavelength (7140eV) x-ray diffraction experiment occurred on two crystals at beamline X-25 at the National Synchrotron Light Source (NSLS). Two 360° data sets were collected on two separate crystals. The diffraction data were integrated and scaled with HKL2000 (Otwinowski, 1997). Merging the raw images of two datasets together was unsuccessful (R_{merge}= 11.9%) and each dataset was treated separately in peak searching and phasing. This inability to merge the data together resulted in a lower then anticipated redundancy value for the data. The scaled data file (.sca) was loaded into XPREP and a single wavelength anomalous scattering ΔF file and instruction file

were generated for XM. Peak searching was run through XM but no peaks were found for either dataset. Both scaled data files were loaded into ISAS2001 with the entire list of sulfur sites (from 1epw), but the resulting maps were not interpretable. While, this portion of the research was not successful, it was not due to the methodology. Several factors contributed to the overall inability to successfully phase this structure. NSLS is a 2nd generation synchrotron source and in as much does not produce a very brilliant X-ray beam. Mounting of the crystals was carried out by members of the Swaminathan lab prior to my arrival. This can contribute a significant amount of background scattering which increases the level of noise in the dataset. Due to the poisonous and biohazardous nature of *C. bot.* neurotoxin type B, transporting the crystals to a 3rd generation synchrotron source is not possible. The establishment of a lab for the handling of that type of biological agent was not feasible at the time of the research. The greatest limiting factor for this research was the inability to merge together any of the data collected on different crystals. As the simulation showed, increased data redundancy resulted in a rescuing of the anomalous signal.

To summarize this experiment, these computer simulations provide support for the ability of sulfur-SAS to phase a protein greater than 100KDa. Initial results suggest that 70% of the total sulfur positions are necessary for an interpretable electron density map. Also, random error in the dataset can be overcome with addition of more diffraction data in the form of redundancy. The R_{as} ratio represents a measure of anomalous signal to noise and can be calculated as the data is being collected. The simulation suggests a threshold R_{as} value of 1.6 in order to locate the sulfur positions for phasing and generating interpretable electron density maps. Since this ratio can be calculated while the data is collecting, it can be used as a qualitative indicator of anomalous signal. The R_{as} value of 1.6 has been independently calculated by other members of

our lab group on cubic zinc-free insulin with a copper rotating anode X-ray generator (Fu *et al.*, 2004) and represents an achievable goal in data collection from both home source and synchrotron X-ray radiation.

Since the carryover from simulation to synchrotron data was not successful, questions still remain about the additive affects of merging R_{as} from two or more separate crystals. Was the inability to phase *C. bot.* neurotoxin type B from not being able to merge the data together or is there an inherent inability that R_{as} can not be increased without collecting data on the same crystal? Analysis of a new *de novo* structural genomics target helps answer this aspect of sulfur-SAS with respect to R_{as}.

Chapter 3

Pfu-542154: Crystallization, phasing, and structure

The National Institutes of Health (NIH) started the first phase of the Protein Structure Initiative (PSI1) in 2000 by establishing pilot structural genomics centers. These centers were to develop high throughput pipelines for all aspects of structure determination from recombinant expression to model building. The ultimate structural goal was develop methods and technology to reach a lofty 100 new *de novo* structures a year per center. The end result would be to establish a "catalog" of the major families of protein structure that could be utilized by any researcher later on to help solve their individual structure either through homology modeling or as simple as molecular replacement when using X-ray crystallography. At the Southeast Collaboratory for Structural Genomics (SECSG), one of the model organisms chosen for structural genomics study was the hyperthermophillic archaeon *Pyrococcus furiosus. P. furiosus*, an obligate anaerobe, was discovered in 1986 on the ocean floor off of the coast of Italy in a shallow water hydrothermal vent (Fiala, 1986) and complete genome sequencing was completed in 2001 (Robb *et al.*, 2001). This archaeon has an optimal growth temperature at 100°C which makes it an interesting model organism for structural genomics. Interesting in terms of the transition from prokaryote to eukaryote and how the archaeal proteins within the cell remain stable and soluble at a temperature that typically denatures the average protein. This thermostability poses a side benefit in recombinant expression and purification from *E. coli*, because of the addition of a heat treatment step that should denature the native *E. coli* proteins and leave the *P. furiosus* protein intact.

3.1: Purification and initial crystallization

Pfu-542154, a 150 amino acid (16.8 KDa) protein, with 3 cysteines and 0 methionines (the N-terminal methionine was lost during fusion of the 6xHis tag), was expressed and purified by the Crystallomics division of SECSG following the standard high throughput SECSG protocol (Jenney *et al.*, 2005). The DNA was cloned into the SECSG modified pET vector, pET24 dBam, and expressed in BL21DE3 *E. coli* cells. The cells were sonicated and the resulting lysate was placed into an 80°C water bath for 60 minutes. After centrifugation, the soluble protein fraction is loaded onto a 5mL NiNTA Ni-affinity chelating column and eluted with a step imidazole gradient. Fractions were collected and analyzed using denaturing sodium dodecyl sulfate poly-acrylamide gel electrophoresis (SDS-PAGE) and visualized with coomassie blue staining. Fractions containing the correct molecular weight protein are loaded onto a Supredex 75 gel filtration column. Peaks from the column are analyzed by SDS-PAGE and the protein fractions pooled and concentrated into buffer containing 20mM HEPES pH 7.6 and 100mM potassium chloride.

The purified product was given to the SECSG crystallization core for screening and optimization. Crystallization screening uses the modified microbatch crystallization experiment consisting of 0.5µL of protein and 0.5µL of precipitating solution with 4 milliliters of 80:20 paraffin to silicon oil layered on top. This layer allows water to slowly evaporate through the oil, concentrating the contents of the drop. The experiment will completely evaporate and go to dryness in about 4-5 weeks. Screening consists of eight commercially available screens; Hampton Research's Crystal Screen I & II, Peg/Ion, Cryo, and MembFac, Emerald Biosystem's Wizard I & II, and Memsys from Molecular Dimensions Ltd. totaling 384 crystallization conditions. *Pfu*-542154 showed a positive initial crystallization hit in PegIon screen condition

25, 200mM magnesium acetate and 20% (w/v) poly-ethylene glycol (PEG) 3350. Grid screen optimization around initial PEG and magnesium acetate concentrations did not yield any crystals. An additive screen of 96 different small molecules was used in combination with the original screen condition. Small, ~50μm, crystals grew in the screening condition with a 12% final concentration methanol additive. A crystal was mounted directly from the crystallization condition and diffracted to a resolution of 1.7Å. A hexagonal dataset was collected with 0.97Å wavelength X-rays (12807eV) at beamline 22-ID, Southeast Regional Collaborative Access Team (SER-CAT), in the Advanced Photon Source on Argonne National Lab in Argonne, Illinois (use of the Advanced Photon Source was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38) and, based on systematic absences, determined to be either $P3_121$ or $P3_221$. Attempts to reproduce these crystals were never successful. Due to the high throughput nature of SECSG and PSI1, structural solution of this target through SECSG crystallization core was cancelled.

3.2: Crystal optimization

The lack of methionine, meaning traditional MAD could never be successful and the lack of reproducibility in the crystals hinders the generation of heavy atom derivatives. The sequence information combined with the diffraction limit and spacegroup of the data, make it an excellent candidate for sulfur-SAS phasing. However, the reproducibility of this target makes even sulfur-SAS a difficult proposition. The first hurdle was making *Pfu*-542154 crystallization reliable and predictable.

Two trays of the 36 condition optimization grid screen with 15% methanol additive were setup as modified microbatch experiments using a Douglas Instruments Oryx-6 crystallization

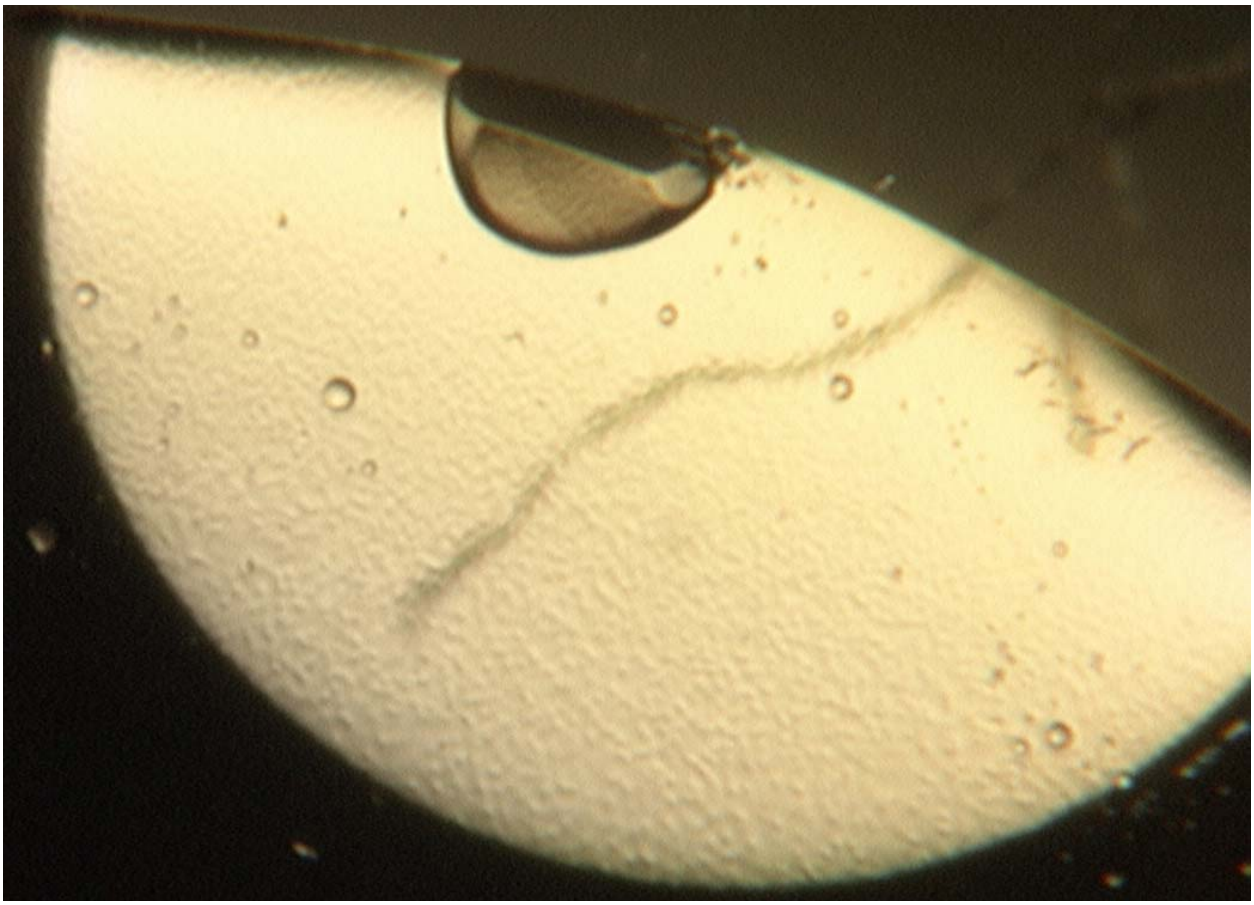


Figure 3.1: Crystal of *Pfu*-542154
In an attempt to crystallize *Pfu*-542154, 12 copies of the 36 condition optimization screen, 432 experiments, with a final concentration of 12% methanol in the well were setup as modified microbatch experiments using the Douglas Instruments Oryx-6 crystallization robot. 0.5µL of protein and 0.5µL of precipitating solution are mixed and then a 4mL layer of 80:20 mixture of paraffin to silicon oil is placed on top. A single large, 200µm in length, crystal formed within 3 days in condition 10, 25% (w/v) PEG 3350 and 150mM magnesium acetate, on one of the optimization screens.

robot but no crystals observed. After reassurance about the difficult crystallization, two more trays were setup and again no crystals formed. Out of frustration, I setup 8 trays of the screen with 15% methanol additive. A single large (200μm in length) crystal (Figure 3.1) formed within 3 days in condition 10, 25% (w/v) PEG 3350 and 150mM magnesium acetate, on one of the trays of the optimization screen. The crystal was firmly cemented to the bottom of the well. During mounting, the crystal had to be broken and pieces of sufficient size were mounted straight from the crystallization experiment well without an added cryoprotectant. One of the shards of the crystal diffracted to a resolution of 2.2Å during screening on the Cu-rotating anode X-ray generator and was saved for a later data collection.

Since smaller unmountable crystal shards remained and no other crystals had formed by day 7, streak seeding was attempted in all wells of the tray that produced the crystal. In streak seeding (Stura, 1991), a whisker or wire on the end of a dowel is touched to the remnants or unusable clusters of a crystal in order to pick up small microscopic pieces of the crystal on the tip of the whisker. The whisker is then drug through a crystallization experiment where no crystals have formed. Crystals form when an experiment reaches the nucleation zone on the solubility diagram and then fall down into the metastable zone where crystals grow. If the experiment enters nucleation and stalls there, crystals will never form; only microscopic crystal nuclei. In the same way, if an experiment stalls in the metastable zone, without reaching nucleation first, no crystals form because a crystal nucleus must be present first and then a crystal can grow from that. By streak seeding an experiment that is stalled in the metastable zone, nuclei, the small bits of crystal stuck to the end of the whisker, are deposited into the experiment and crystals grow along the streak line, where the whisker was drug through the well. Streak seeding was successful for the E and F conditions (36 conditions of 6 rows with columns A-F) in every row

signifying that these experiments were in the metastable zone. Since these were modified microbatch experiments, there was no way of knowing how much water had already evaporated out of the experiment and therefore no way of knowing the concentrations of the chemicals inside of the well at the time of crystallization. Another optimization screen tray was setup but layered with pure paraffin oil, changing from a modified microbatch experiment to a true batch experiment, where no water is allowed to evaporate out of the drop. If streak seeding is successful in this setup then reproducibility is assured as long as there is a source of nuclei to deposit. Again, streak seeding proved successful for the E, 30% PEG 3350 with 100-350mM magnesium acetate, and F, 35% PEG 3350 with 100-350mM magnesium acetate, columns of every row but the crystal morphology, edges and facets, were not as sharp as the previous streak seeding experiment when some water and presumably methanol, since it is so volatile, had evaporated. A series of methanol dilutions, from 0% to 10% at every 2%, was setup as batch experiments and streak seeded as before. Well defined crystals with sharp edges were seen in experiments with a final concentration of methanol as high as 4%. Transposed plates, where the vertical column E in the 6 rows of the optimization screen is transposed to the horizontal 6 welled individual row (well A1=condition 5, A2=condition 11, etc.), were set up as a72 well experiment with the odd rows being transposed column E and the even rows transposed column F. All experiments contained methanol at a 4% final volume in the experiment and layered with pure paraffin oil. The tray was allowed to sit overnight for complete mixing to occur and then streak seeded. Figure 3.2 represents a typical streak seeding result in the odd numbered rows. Crystals grow along the streak line as a conglomerate, but also can grow as single crystals. In the image, the single crystal is 75μm on edge and eventually grew to 125μm before being mounted. Unfortunately, these crystals are also attached to the bottom of the well very tightly.

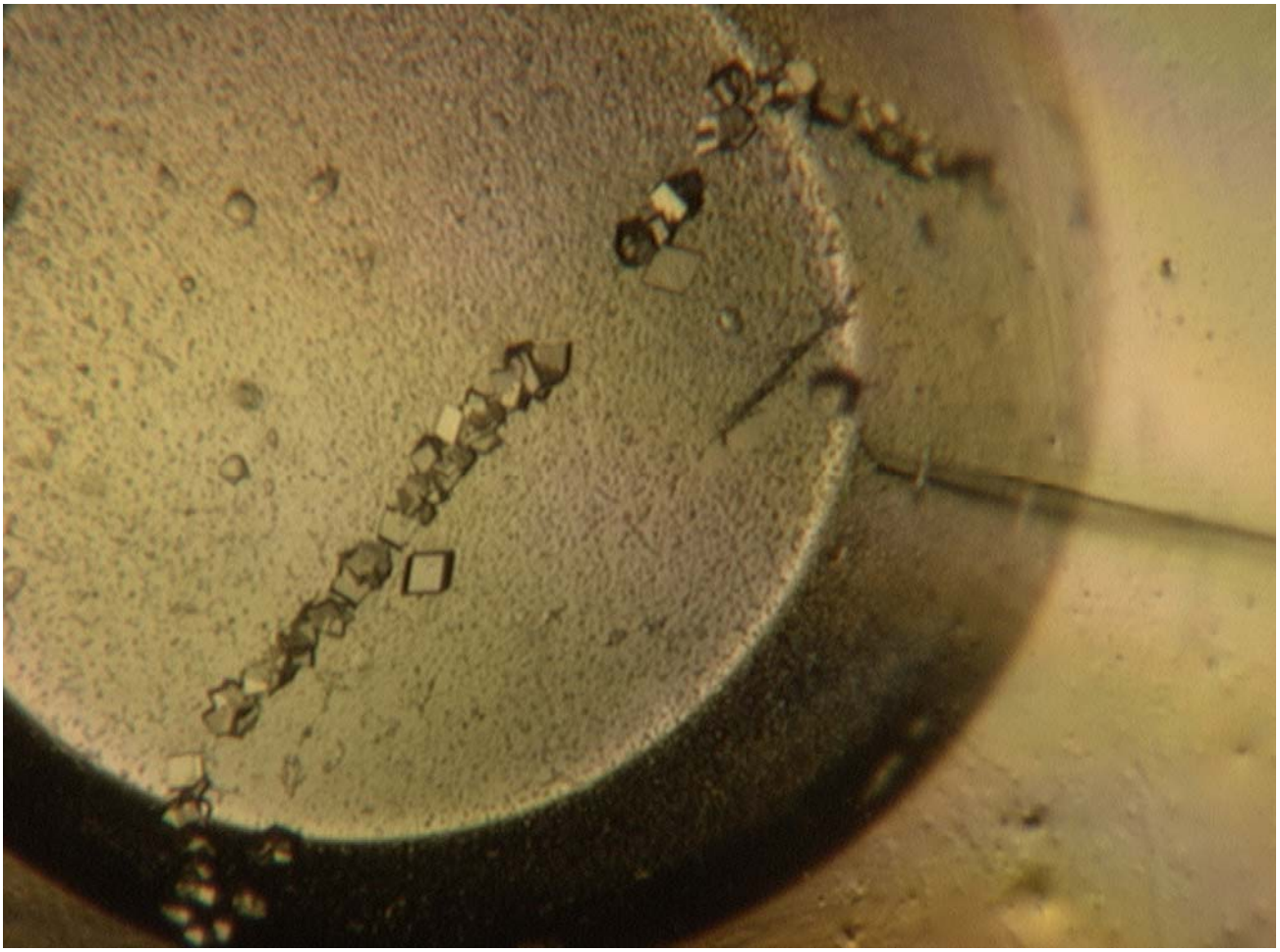


Figure 3.2: Streak seeding of *Pfu*-542154
Typical streak seeding result in the odd numbered rows of the column E and F transposed plate. All wells contained methanol at a 4% final volume and are layered with pure paraffin oil, forming a batch experiment with no water evaporation. The tray was allowed to sit overnight for complete mixing to occur and then streak seeded with a cat whisker. Nuclei are deposited and, if the well is in the metastable zone, crystals grow along the streak line as a conglomerate, but also can grow as single crystals. In the image, the single crystal to the side of the streak line is 75μm on edge and eventually grew to 125μm before being mounted.

This presents a significant problem because, unlike the larger "bubble" shaped crystal, breaking these off the bottom of the well will not result in pieces large enough to mount and test for diffraction. After several failed attempts at trying to pry the crystal off the bottom of the well, a discussion with Kris Tesh from Rigaku-MSC proved very beneficial. Kris advised taking a hard needle or pointed microtool and marring the plastic in the area next to the crystal. The idea being that marring the plastic would cause the surface under the crystal to become slightly convex and pop the crystal away from the plastic surface. This technique proved instantly successful and resulted in single crystals for mounting. Crystals were mounted directly from the crystallization experiment and screened with the Cu-rotating anode X-ray generator. However, the screening images all contained significant ice rings in the diffraction pattern.

The ice rings could be contributed to a couple of factors but the likely cause was preventing water from evaporating out of the experiment. In the previous experiments, water was allowed to evaporate out concentrating the chemicals in the drop, including the PEG 3350. PEG's are excellent cryoprotectants at sufficient concentration and allow for direct mounting of crystals without organized ice formation upon cryocooling. Since these crystals were grown under pure batch conditions, a cryoprotectant was needed. 8μL of the crystallization condition, without methanol, was mixed with 2μL of 100% glycerol, resulting in the crystallization condition with 20% glycerol. 0.2μL of this solution was injected onto the side of the well using a Hamilton 1μL syringe. This droplet was not injected directly into the crystallization experiment because locally high concentrations of PEG and glycerol could develop and possibly crack or destroy the crystals in the well. Instead, the droplet is placed on the side wall off the well and then a small liquid bridge is established under the oil using the cat whisker from streak seeding. This allows slow diffusion of the glycerol into the crystallization experiment. Once the

bridge was established, 10 minutes is allowed to pass to ensure mixing has occurred and to verify that the final concentration of cryoprotectant in the drop has not damaged the crystals. Single crystals of sufficient size were mounted and excess liquid from the drops was removed by touching the liquid portion to the side wall of the well above the drop. This seemingly insignificant procedure removes excess liquid from around the crystal that may increase the brackground scattering during a single crystal diffraction experiment to the point of drowning out the sulfur anomalous signal. Crystals were screened using the Cu-rotating anode X-ray generator and those crystals with diffraction above 2.5Å resolution were saved for data collection.

3.3: Data collection and processing

Initial data collection started prior to the success of streak seeding. At that time, the piece of the "bubble" crystal was the only mounted diffraction quality crystal of *Pfu*-542154. Unsure if more crystals would be forthcoming, plans were made to send the crystal to the Rigaku-MSC North American headquarters in The Woodlands, Texas. The demo floor at Rigaku-MSC contains a higher intensity Cr-rotating anode X-ray generator and second generation Cr-optics compared to the current setup in the UGA X-ray lab. The overall result is an X-ray beam about 5 times more intense than we can generate. Using a Micromax-007 Cr-rotating anode X-ray generator with the Varimax-Cr optics and a Raxis-IV++ image plate detector, the crystal was mounted, centered, and two 360° spheres of 100°K data were collected with 1° oscillation images at an exposure of 5 minutes per image with a crystal to detector distance of 100mm. The face of the detector has been modified to include a helium filled square pyramidal "cone" extending from the face of the detector to almost touching the beamstop; used to minimize air absorption of the scattered X-rays between crystal and detector. The crystal diffracted to 2.4Å

Diffraction Statistics

	360° MSC	360° UGA	720° UGA-MSC Merged
Crystal Size (μm)	100×100×75	125×125×100	
Crystal to Detector Distance (mm)	100	100	
Frame Width (°)	1	1	
Exposure (s)	300	300	
2θ (°)	0	0	
X-ray Wavelength (eV)	5425	5425	
Spacegroup	P3121	P3121	P3121
Unit Cell			
a = b (Å)	53.71	53.66	53.69
c (Å)	86.88	86.66	86.66
γ (°)	120	120	120
Resolution (Å)	2.4	2.4	2.4
Redundancy	18.8 (12.8)	19.9 (18.6)	37.9 (31.7)
Completeness	99.4 (93.9)	99.8 (97.8)	100.0 (99.7)
Rsym (%)	4.5 (25.1)	3.3 (9.4)	
Rmerge (%)			5.5 (17.2)
I/σ	59.25 (12.03)	73.05 (33.22)	137.2 (35.23)

Structure Statistics

R	20.2
R_{free}	25.2
R.M.S. Bond Length (Å)	0.016
R.M.S. Bond Angle (°)	1.42
Average B-factor ($Å^2$)	30.73

Table 3.1: Data processing and structure statistics from *Pfu*-542154 crystals used for phasing

Summary of reflections intensities and R-factors by shells
R linear = SUM (ABS(I - <I>)) / SUM (I)
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2)
Chi**2 = SUM ((I - <I>) ** 2) / (Error ** 2 * N / (N-1)))
In all sums single measurements are excluded

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
20.00	5.14		10285.9	205.3	30.5	1.072	0.035	0.073
5.14	4.09		10190.0	100.3	30.0	0.920	0.024	0.026
4.09	3.58		7472.6	73.9	27.3	0.917	0.026	0.028
3.58	3.25		4391.7	46.9	22.2	0.955	0.031	0.032
3.25	3.02		2800.4	33.0	18.7	0.960	0.036	0.037
3.02	2.84		1846.3	22.0	16.6	1.096	0.044	0.044
2.84	2.70		1387.6	19.5	15.8	1.003	0.050	0.050
2.70	2.58		1033.8	17.8	15.3	0.966	0.062	0.061
2.58	2.49		798.0	16.9	15.2	0.928	0.071	0.072
2.49	2.40		571.4	17.2	16.1	0.957	0.094	0.091
All reflections			4200.4	57.5	21.0	0.978	0.033	0.051

Table 3.2 Scaling statistics from *Pfu*-542514 crystal collected at UGA

Summary of reflections intensities and R-factors by shells
R linear = SUM (ABS(I - <I>)) / SUM (I)
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2)
Chi**2 = SUM ((I - <I>) ** 2) / (Error ** 2 * N / (N-1)))
In all sums single measurements are excluded

Shell limit	Lower Angstrom	Upper Angstrom	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
20.00	5.14		12359.1	188.9	38.2	0.793	0.030	0.034
5.14	4.09		11460.4	143.1	44.7	0.917	0.033	0.035
4.09	3.58		8108.2	122.4	43.4	0.924	0.039	0.040
3.58	3.25		4592.2	67.6	36.5	1.036	0.047	0.046
3.25	3.02		2781.1	48.6	32.4	1.086	0.060	0.057
3.02	2.84		1753.7	42.1	30.0	1.040	0.082	0.075
2.84	2.70		1295.1	38.0	29.9	1.135	0.104	0.096
2.70	2.58		887.6	34.5	29.4	1.213	0.142	0.126
2.58	2.49		641.8	33.0	28.1	1.326	0.187	0.175
2.49	2.40		442.6	36.8	30.9	1.435	0.251	0.217
All reflections			4591.5	77.5	34.5	1.065	0.045	0.038

Table 3.3 Scaling statistics from *Pfu*-542514 crystal collected at MSC

Summary of reflections intensities and R-factors by shells
R linear = SUM (ABS(I - <I>)) / SUM (I)
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2)
Chi**2 = SUM ((I - <I>) ** 2) / (Error ** 2 * N / (N-1)))
In all sums single measurements are excluded

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
20.00	5.14		10470.5	65.6	21.7	3.137	0.038	0.042
5.14	4.09		10617.4	59.7	24.3	3.192	0.044	0.047
4.09	3.58		7819.3	46.2	22.9	3.160	0.049	0.052
3.58	3.25		4595.1	31.4	19.2	3.026	0.060	0.063
3.25	3.02		2938.1	23.5	16.6	2.627	0.069	0.070
3.02	2.84		1937.8	19.0	14.9	2.096	0.081	0.081
2.84	2.70		1443.3	17.2	14.4	1.855	0.094	0.095
2.70	2.58		1074.3	16.1	14.1	1.643	0.114	0.113
2.58	2.49		822.0	15.3	14.0	1.675	0.138	0.144
2.49	2.40		581.3	16.5	15.7	1.587	0.172	0.171
All reflections			4349.2	31.7	17.9	2.432	0.055	0.050

Table 3.4 Scaling statistics from the first 360° of both crystals merged together

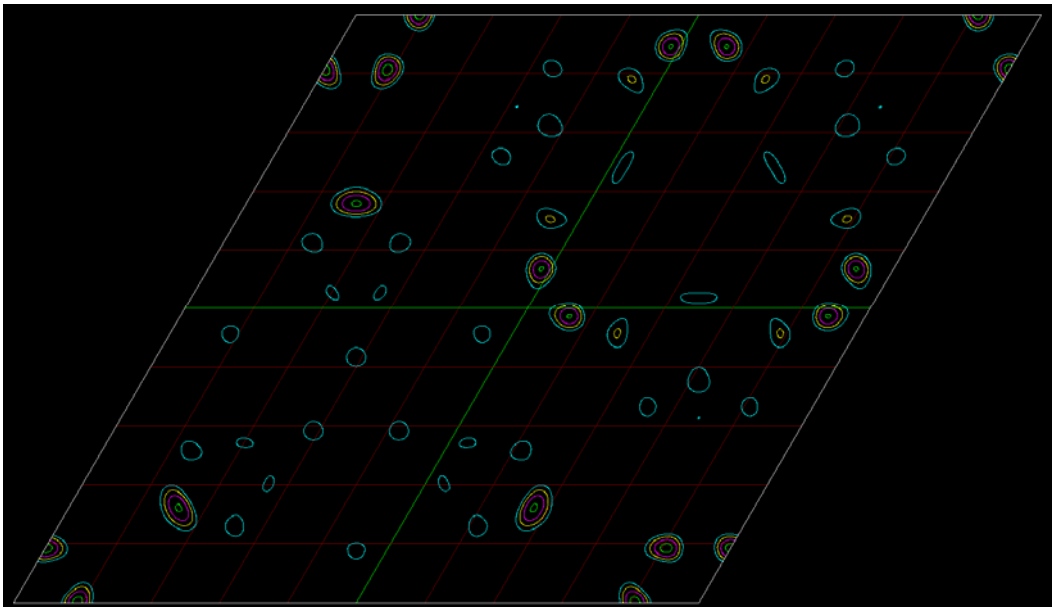
resolution, the edge of the detector's surface. First, each sphere of data was indexed, integrated, and scaled individually using HKL2000 (Otwinowski, 1997) and then merged together with the resulting statistics in Table 3.1. In addition to HKL2000, each sphere was indexed and integrated with D*TREK. The second integration was done in order to use 3DSCALE to calculate R_{as}, currently 3DSCALE can not read in the integration .x files of HKL2000. Based on the resulting R_{sym} from scaling, the spacegroup was determined to be either $P3_121$ or $P3_221$. The space group could not be resolved any further because the systematic absences for both spacegroups are identical. The resulting cell constants and spacegroup are consistent with the previous synchrotron high resolution dataset collected from the small crystal in the original additive screen crystallization.

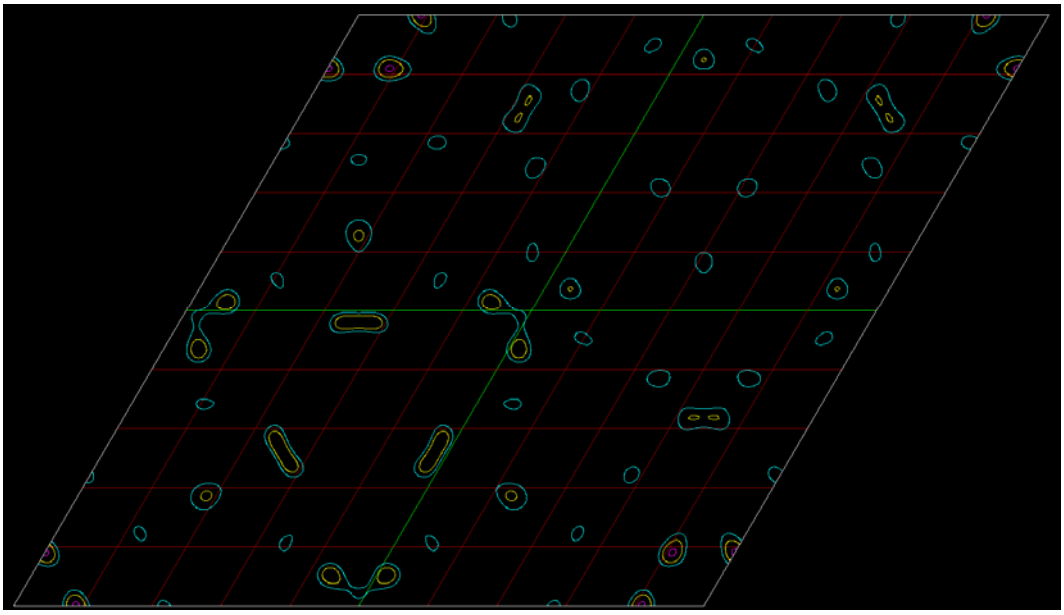
A "streak" crystal was mounted on the UGA Cr-rotating anode X-ray generator. Two 360° spheres of 100°K data were collected with 1° oscillation images at an exposure of 5 minutes per image with a crystal to detector distance of 98mm. The crystal diffracted to the edge of the Raxis-IV detector, 2.4Å resolution. As with the "bubble" crystal, the two spheres of data were processed individually and merged together with HKL2000 and D*TREK for R_{as} calculation. The cell constants, spacegroup and processing statistics are comparable to the two other crystals collected.

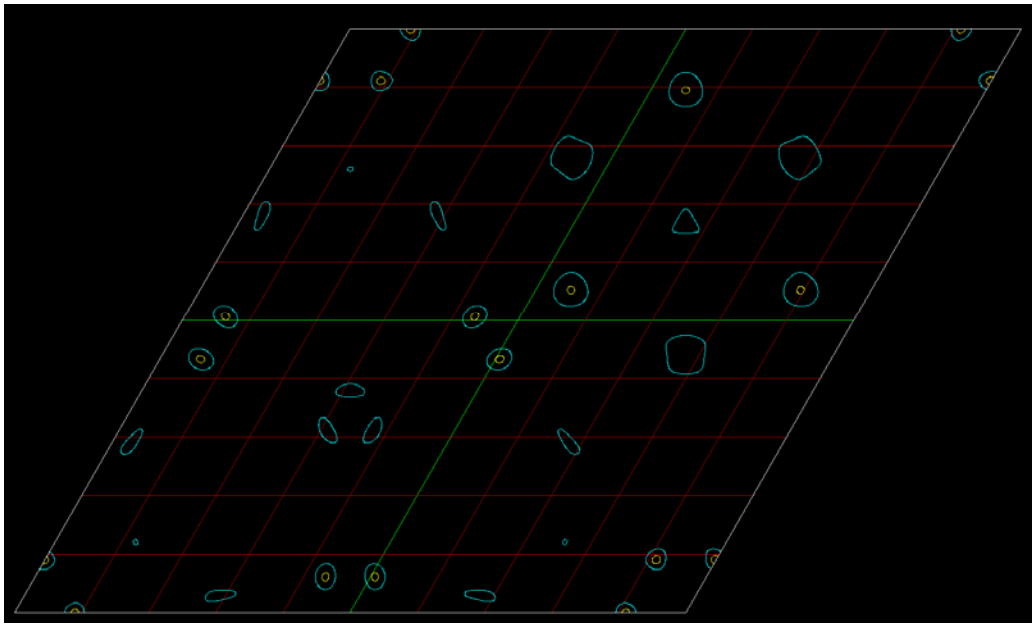
3.4: Phasing and structure solution

The merged .sca files from the two Cr-rotating anode dataset crystals were read into XPREP for creating an instruction file for heavy atom searching using XM. Also, XPREP allows for quick visual inspection of an anomalous Patterson map for possible heavy atom locations, Figure 3.3. XM heavy atom searching uses an automated Patterson peak search that produces a list file of the located positions with peak correlation statistics. Correlations of all the

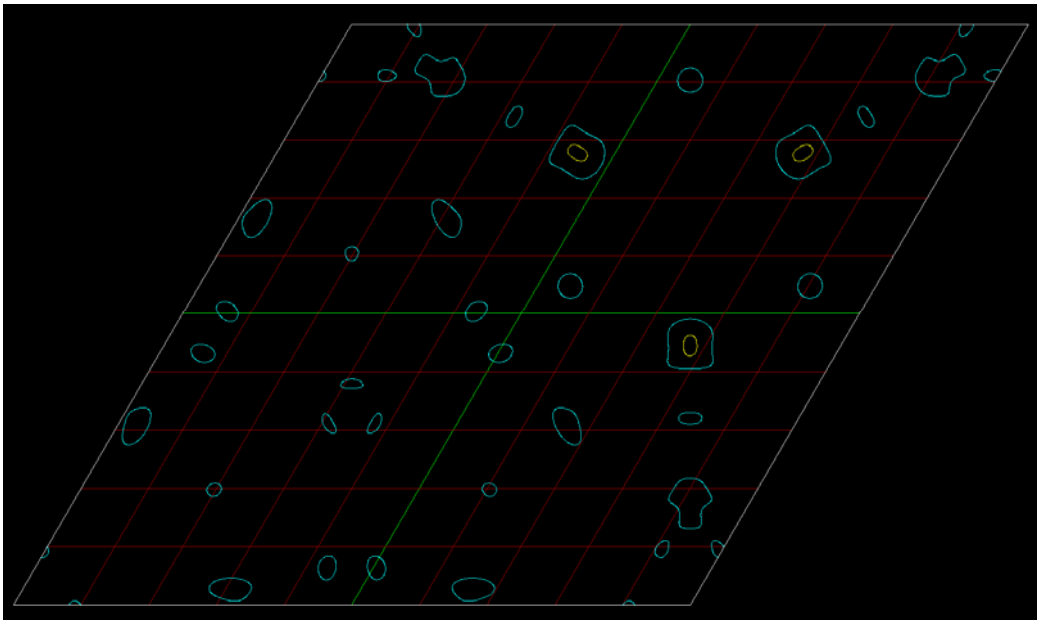
A







B



C

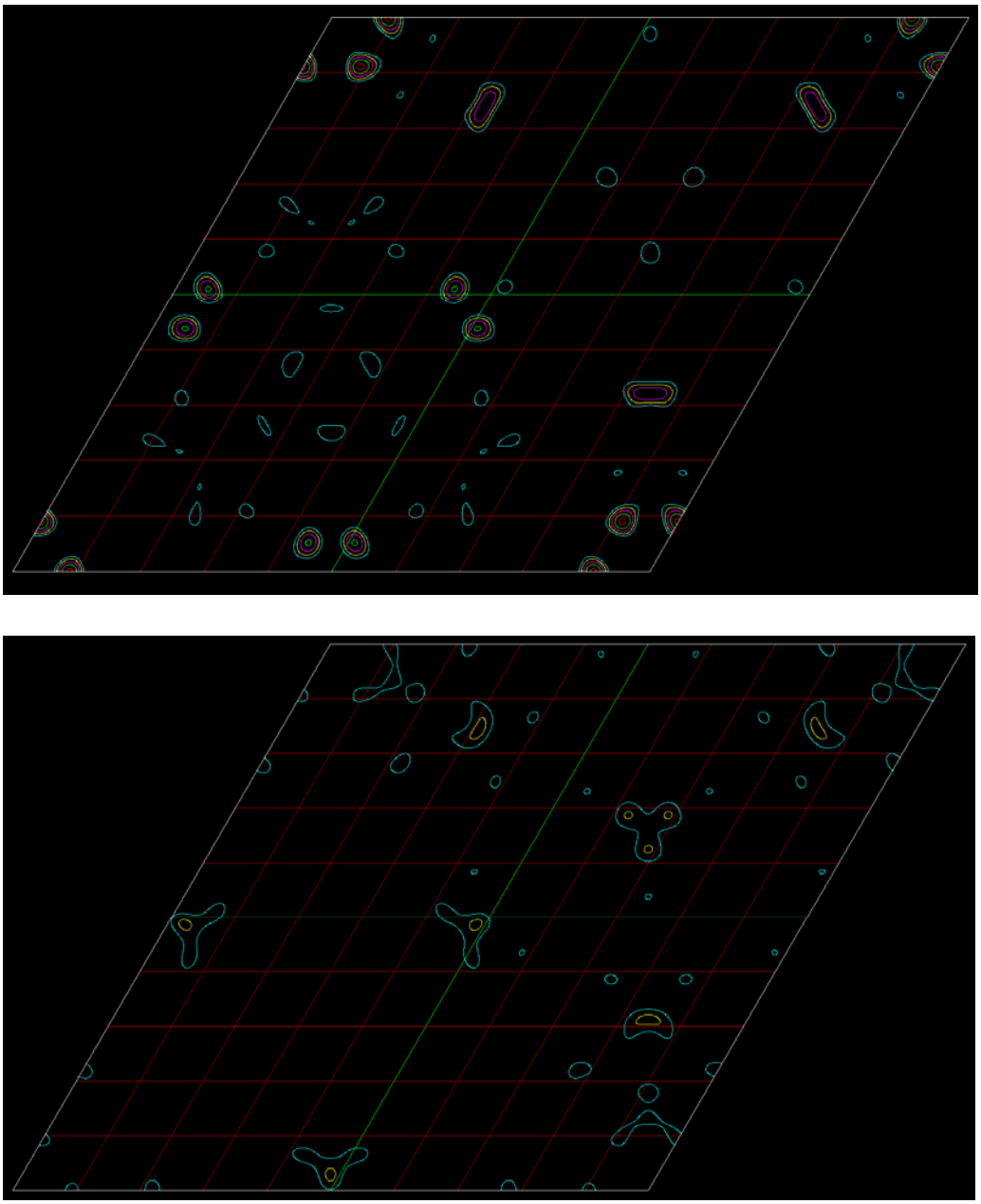


Figure 3.3: Anomalous Patterson maps of merged *Pfu*-542154
Anomalous Patterson maps generated with XPREP with a data resolution of 3.0Å. **(A)** Data from crystal collected on MSC Cr-rotating anode. The left panel is the first 360° and the right the merged 720° of data. **(B)** Data from crystal collected on UGA Cr-rotating anode. The left panel is the first 360° and the right the merged 720° of data. **(C)** Merged data from crystals collected on both the MSC and UGA Cr-rotating anode. The left panel is the first 360° of each dataset merged together (720° total) and the right, both complete datasets merged together (1440° total). Comparing the left panel to the right panel in each case, a decline in the peak height is seen when adding the second sphere of data indicating that more noise is being added than signal.

Data Set	Sulfur Sites Searched	All	Weak
UGA -720	3	30.56	13.88
MSC-720	3	30.57	16.24
UGA-MSC First Pass Merged	3	32.34	14.92
UGA-MSC First Pass Merged	2	33.45	18.38

Table 3.5: XM heavy atom peak searching statistics of *Pfu*-542154 chromium data

peaks and only the weak peaks are reported for every round of searching. For a positive solution, we would expect statistics of 30.0 for All and 20 for Weak. Table 3.2 shows the XM peak searching statistics for the two 720° datasets. The MSC data gives very good search statistics that look to be correct whereas the UGA data search statistics are significantly lower. Based on these results, it looks as if only the MSC data can be used for structure solution. However, by looking at the Patterson maps of the first 360° of the UGA data (Figure 3.3b left) compared to the 720° merged data, you see that the small peak heights that were there have all but disappeared; indicating that the second sphere of data is only adding noise to the data and not increasing the sulfur anomalous signal. To a lesser extent, the same result is seen in the anomalous Patterson maps of the MSC data (Figure 3.2a). Since the majority of the anomalous signal for both datasets is in the first sphere, if we merged together the first spheres of the MSC and UGA data we should see and increase in the peak heights, the XM heavy atom search statistics, and ultimately the R_{as} if anomalous signal truly can be added together. The first two 360° of each dataset were merged together with HKL2000 and the resulting .sca file fed into XPREP for the production of an XM instruction file. The peak searching statistics are slightly better than of the individual datasets alone (Table 3.2). Interestingly, the peak search statistics of the 720° UGA data are surprisingly higher than would be expected by comparing the peak heights in the anomalous Patterson maps, compare Figure3.2a to 3.2b. The .sca files from the 720° MSC data, 720° UGA data, and the merged first spheres of the MSC and UGA data were used for the next step of structure solution, phasing.

The 720° MSC dataset and the first 360° of data from both crystals were also indexed and integrated with D*TREK for R_{as} analysis. As seen in Table 3.1, the R_{as} for the MSC data looks very promising at 1.55. This value meets the threshold value proposed by Fu *et al.* and lies just

below the value proposed by the simulation study of *Cbot* neurotoxin type B. A problem arose when trying to integrate the UGA portion of the data. D*TREK discarded about 10-12% of the reflections and altering the parameters of D*TREK was not any more successful. These discarded reflections translate into less anomalous signal from the dataset. Ultimately, this reduces the R_{as} value to an artificially lower value. Even though the R_{as} value for the merged UGA/MSC data does not appear sufficient for phasing, the .sca file was still sent to the next step.

Phasing of *Pfu*-542154 was done through the SECSG web-based structure solution pipelines (Liu *et al.*, 2005). The .sca files from the merged MSC only data and the combined UGA-MSC data were loaded along with the sequence file of *Pfu*-542154 were loaded into the SCA2Structure pipeline (Liu *et al.*, 2005)(Figure 3.4). In both instances, sulfur was selected as the heavy atom, wavelength set to 2.29Å, resolution limit set to 2.4Å, sites searched set to 3 initially and then reduced to 2 for the best structure solution, 150 residues in the asymmetric unit, 45% solvent content, and both $P3_121$ and $P3_221$ spacegroups used since the data could have been either at that point. The pipeline employs SOLVE/RESOLVE (Terwilliger, 2000, 2003, Terwilliger & Berendzen, 1999) which finds and refines the heavy atom positions, calculates the initial electron density map, and then uses ARP/WARP (Perrakis *et al.*, 1999) to auto-trace that electron density. The best solution was obtained with 2 sites instead of 3; evidence supported by the .lst file from manual XM heavy atom searching. The .lst file indicates the peak heights of each heavy atom site found and the height of the third site was always significantly lower than the first two. The three cysteines could be distributed as a disulfide bond and free cysteine, even though that should be unlikely given the reducing environment *P. furiosus* lives in. Figure 3.5 is the output from the most successful pipeline run, most atoms automatically traced in sequence, with 2 sites located instead of three, and sorted by number of ARP/WARP atoms traced. For this

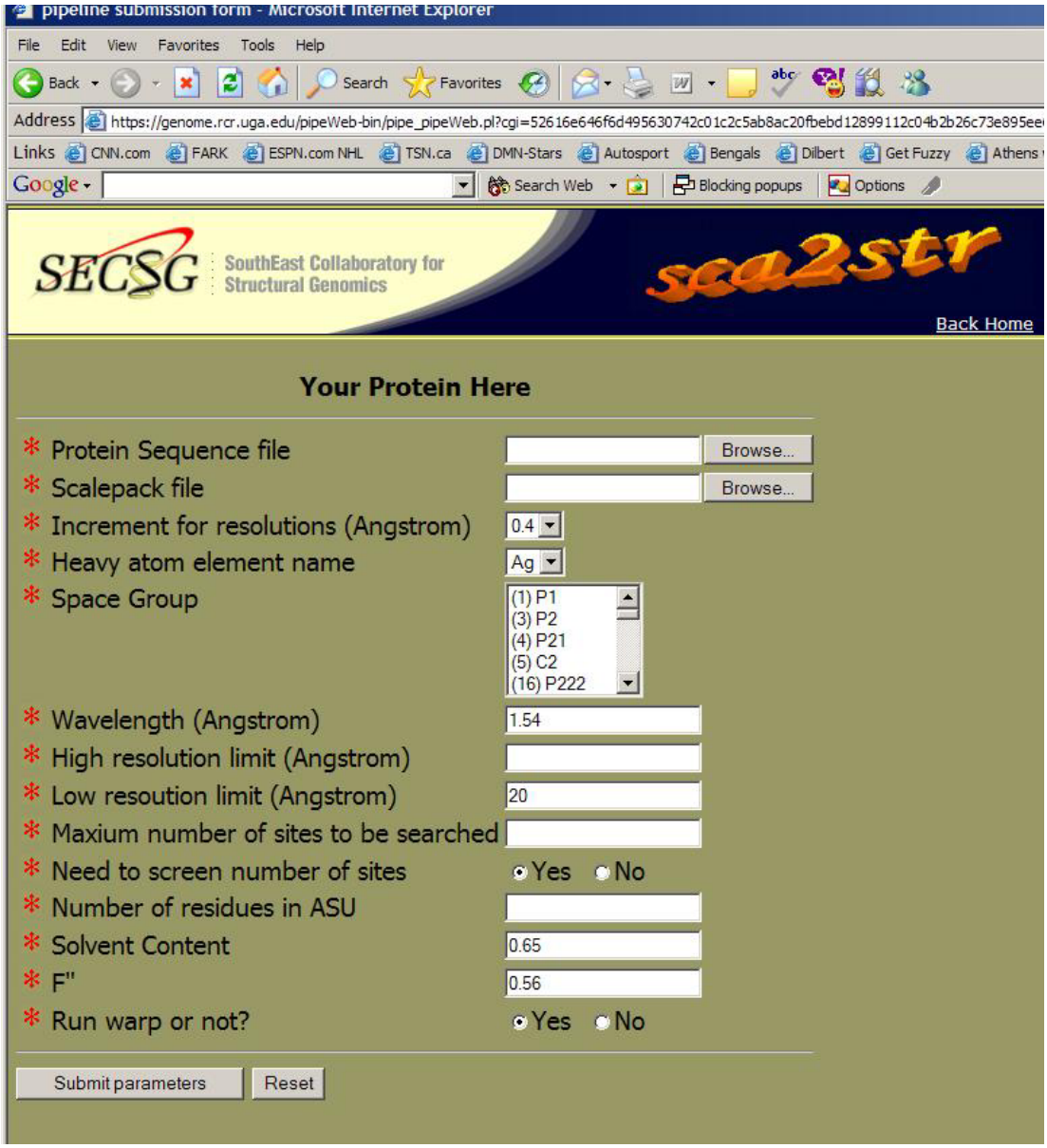


Figure 3.4: SCA2Structure pipeline input webpage
SCA2Structure input page once a job has been started. The .sca files of the merged MSC data and UGA-MSC data were loaded with the sequence file. The pipeline was run with both $P3_121$ and $P3_221$ spacegroups at the same time with a high resolution of 2.4Å. The best results were seen when searching for 2 sulfur sites, 150 residues in the asymmetric unit, and a solvent content of 0.45.

ResPhase	ResSolve	Site	Solvent	SpaceGroup	AtomNum	Zvalue	SolveFOM	ResovFOM	download
3	2.6	2	0.45	P3121	468	7.90	0.32	0.53	tarfile
3.2	2.7	2	0.45	P3221	463	3.94	0.20	0.49	tarfile
3	2.7	2	0.45	P3121	454	8.44	0.32	0.55	tarfile
2.5	2.4	2	0.45	P3121	448	9.16	0.30	0.54	tarfile
2.9	2.8	2	0.45	P3121	448	7.62	0.31	0.58	tarfile
2.8	2.5	2	0.45	P3121	432	9.24	0.32	0.53	tarfile
3.6	3.4	2	0.45	P3121	428	5.42	0.18	0.54	tarfile
2.7	2.6	2	0.45	P3121	426	8.49	0.31	0.57	tarfile
3.5	3.3	2	0.45	P3121	414	4.15	0.20	0.54	tarfile
2.7	2.5	2	0.45	P3121	406	8.03	0.31	0.54	tarfile
3.9	3.4	2	0.45	P3221	399	6.10	0.26	0.52	tarfile
3.6	3.5	2	0.45	P3121	398	2.99	0.25	0.56	tarfile
3	2.4	2	0.45	P3121	394	7.38	0.32	0.49	tarfile
3.5	2.8	2	0.45	P3121	391	3.99	0.22	0.49	tarfile
3.3	2.8	2	0.45	P3121	391	4.56	0.20	0.50	tarfile
2.6	2.5	2	0.45	P3121	386	9.16	0.30	0.55	tarfile
2.9	2.7	2	0.45	P3121	380	8.63	0.32	0.56	tarfile
3.4	3.1	2	0.45	P3221	379	3.12	0.20	0.52	tarfile

Figure 3.5: *Pfu*-542154 SCA2Structure pipeline results
A portion of the output from the most successful SCA2Structure pipeline run, most atoms automatically traced in sequence, with 2 sulfur sites instead of 3, and sorted by number of ARP/WARP atoms traced. For this run, the results from the first solution were downloaded, un-tarred, and the electron density map used for model building.

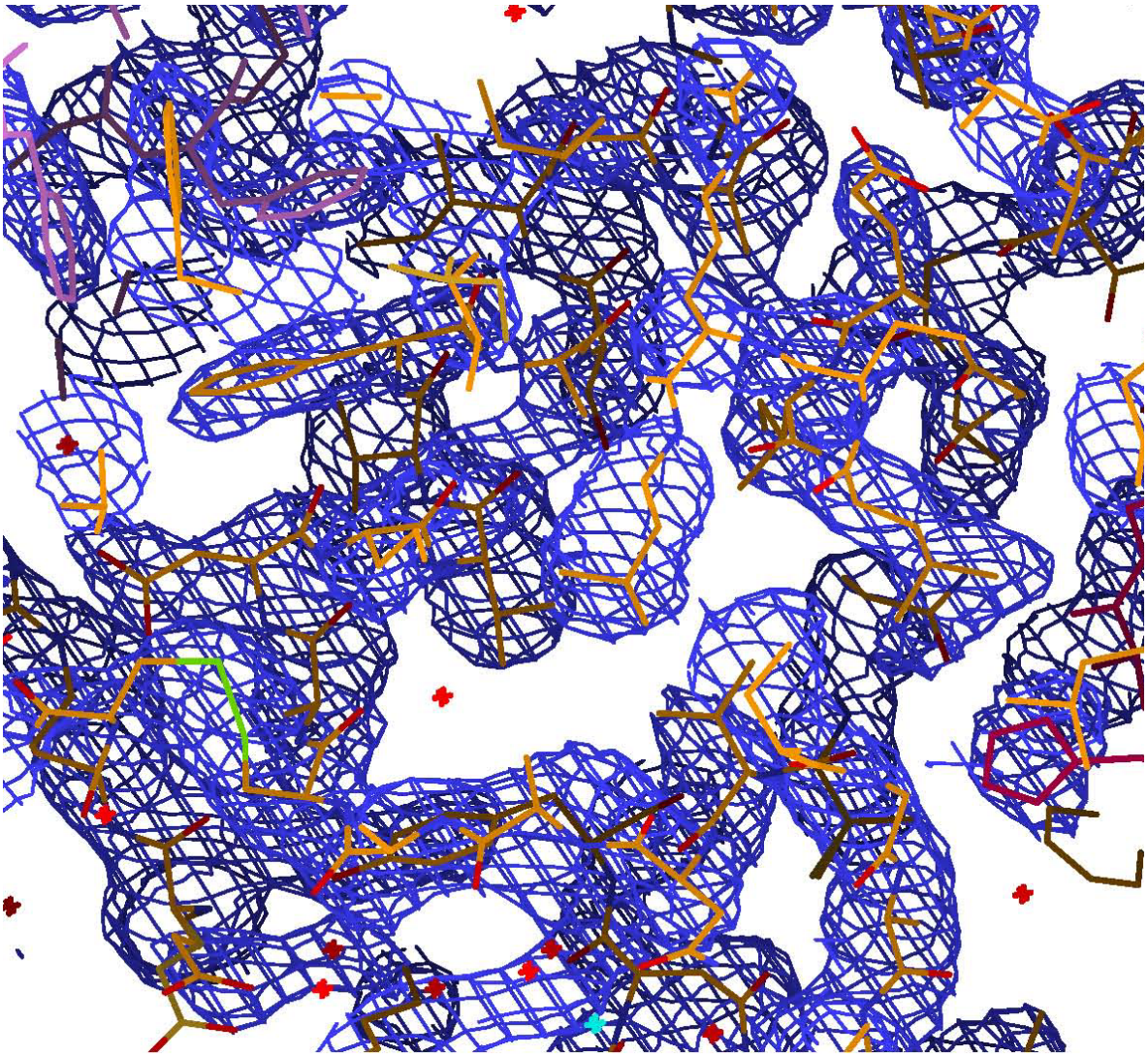


Figure 3.6: Initial F_o electron density from SCA2Structure pipeline
3.0Å F_o electron density map, at a level of 1σ, of the successful SCA2Structure pipeline result from RESOLVE used for initial model building via ARP/WARP automatic tracing visualized with XFIT. Carbon is yellow, oxygen red, nitrogen blue, and water molecules shown as cyan crosses. The model shown is a refined model, showing how well the model fits the observed data, and not the initial ARP/WARP model.

run, the results from the first solution were downloaded and un-tarred. The electron density from this solution, Figure 3.6, was visually inspected with the pipeline ARP/WARP model and determined to be a successful phasing result.

The pipeline traced about 48% of the model automatically. Autotracing was continued using the CCP4i interface of ARP/WARP and the 1.7Å resolution dataset collected with 12808eV (0.97Å wavelength) X-rays at SER-CAT of the Advanced Photon Source, Argonne, IL. The second round of autotracing was able to fill in about 80% of the protein model at an initial R/R_{free} of 31.6% and 38.8% respectively. REFMAC5 (Murshudov *et al.*, 1997) was used for all structural refinement and to generate new electron density maps, with the rest of the model added by hand using XFIT (McRee, 1999). After the protein molecule had been completely traced, the R/R_{free} was locked into a local minimum of 25/31%. The electron density of the majority of the structure was of a quality associated with 1.7Å resolution data, but the $2F_o$-F_c density associated with helix 1 and 5 was not near this quality and made model building within these regions difficult. Regardless of manually tweaking the model into the electron density, R/R_{free} would not decrease any further. A fortuitous posting on the CCP4 bulletin board made reference to using the TLS refinement (Winn *et al.*, 2001) option of REFMAC5 to dramatically reduce R/R_{free} when no other refinement would work. In tls refinement, the temperature factors, B-factors, of the protein structure are treated anisotropically, non-spherical, but not on an individual atomic basis. Instead, rigid bodies are defined (entire molecule, C_α's, single helix, etc.) and an anisotropic B-factor is created for each defined rigid body. The resulting B-factor is a combination of the isotropic atomic B-factor and the tls anisotropic B-factor. CCP4i was used to generate the entire protein molecule as a single tls rigid body, and a file of two defined tls groups, the two problems helices and the rest of the protein. A new round of refinement using REFMAC5 was setup using

the TLS refinement option on the protein as a single rigid body and the R/R_{free} dropped 4% to 21/26%. TLS refinement using more than one rigid body was not as successful in reducing R/R_{free}. In order to monitor the affects of TLS refinement on atomic position, the difference of x, y, and z-values individually before and after TLS refinement are squared and then summed together. The square root of that value is plotted versus atom number (Figure 3.7) and indicates not only did tls altered the B-factors of each atom, but, when coupled with positional least squares refinement, also moved the model with the majority of that movement in the regions of helix 1 and 5. The resulting electron density in the helix 1 and 5 regions improved, but not to the 1.7Å level of the rest of the protein. Water, hetero atoms (ethanol and Mg^{2+}), and unknown atoms (modeled as UNK) were assigned to continually reduce R/R_{free}. In order to make sure the hetero atoms were real, each was removed and then a difference map calculated to check for the appearance of positive electron density in the same location. Only those hetero atoms that reappeared were kept and those that did not discarded. The final steps in reducing R/R_{free} utilized structural clash and rotamer analysis using MOLPROBITY (Lovell *et al.*, 2003) carried out on the internet. Structural improvement via clash reduction (final clash score of 5.5), rotamer optimization, and Ramachandran plot outlier analysis (0 outliers in the end) resulted in a further decrease of R/R_{free} to the final values of 20.2/25.0. The R_{free} value is higher than would be expected from a 1.7Å dataset, but it is reasonable considering the problems with the F_o density associated with helix 1 and 5 (Figure 3.8). It is not uncommon for different portions of the model to have electron density that does not reflect the quality normally associated with the diffraction limit of the data (Krause *et al.*, 1987, Tanner *et al.*, 1993, Zhang *et al.*, 2004). While it adds a small acceptable level of uncertainty to the final model, it is better than trying to force the R/R_{free} lower by relaxing the restraints on bond lengths and angles.

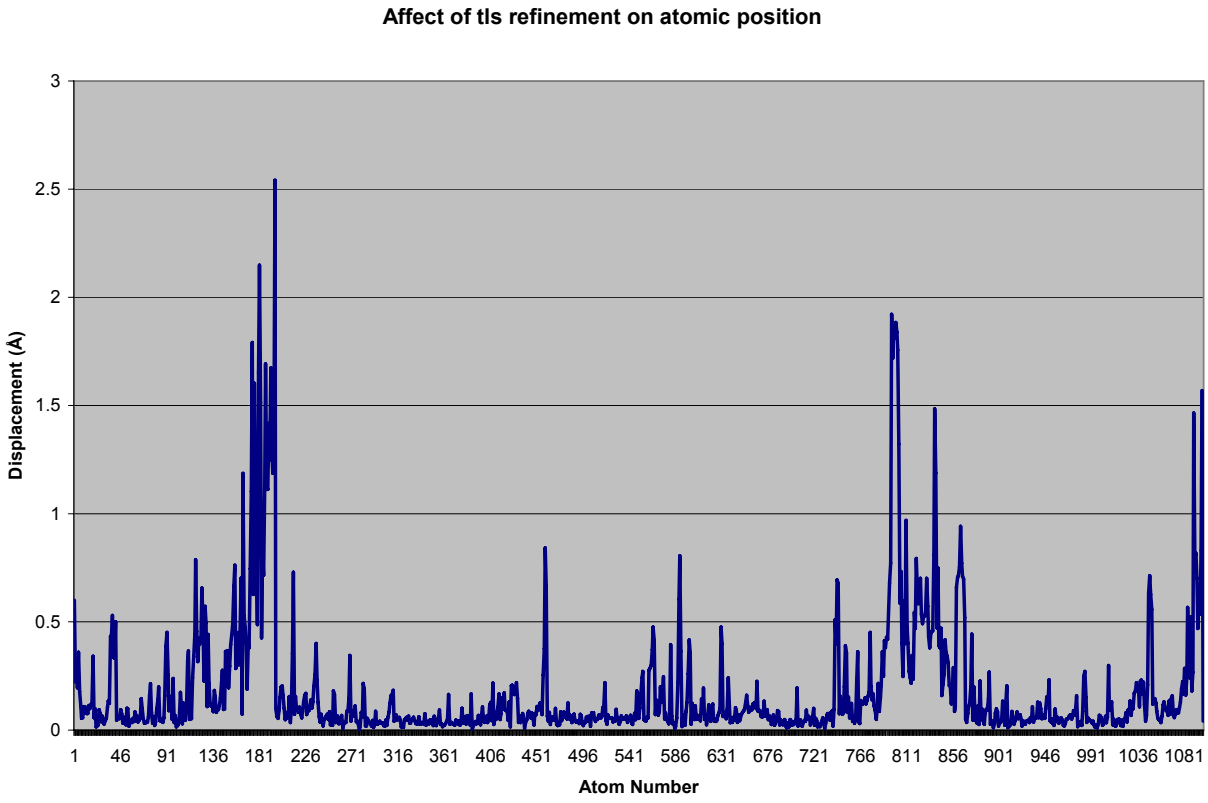
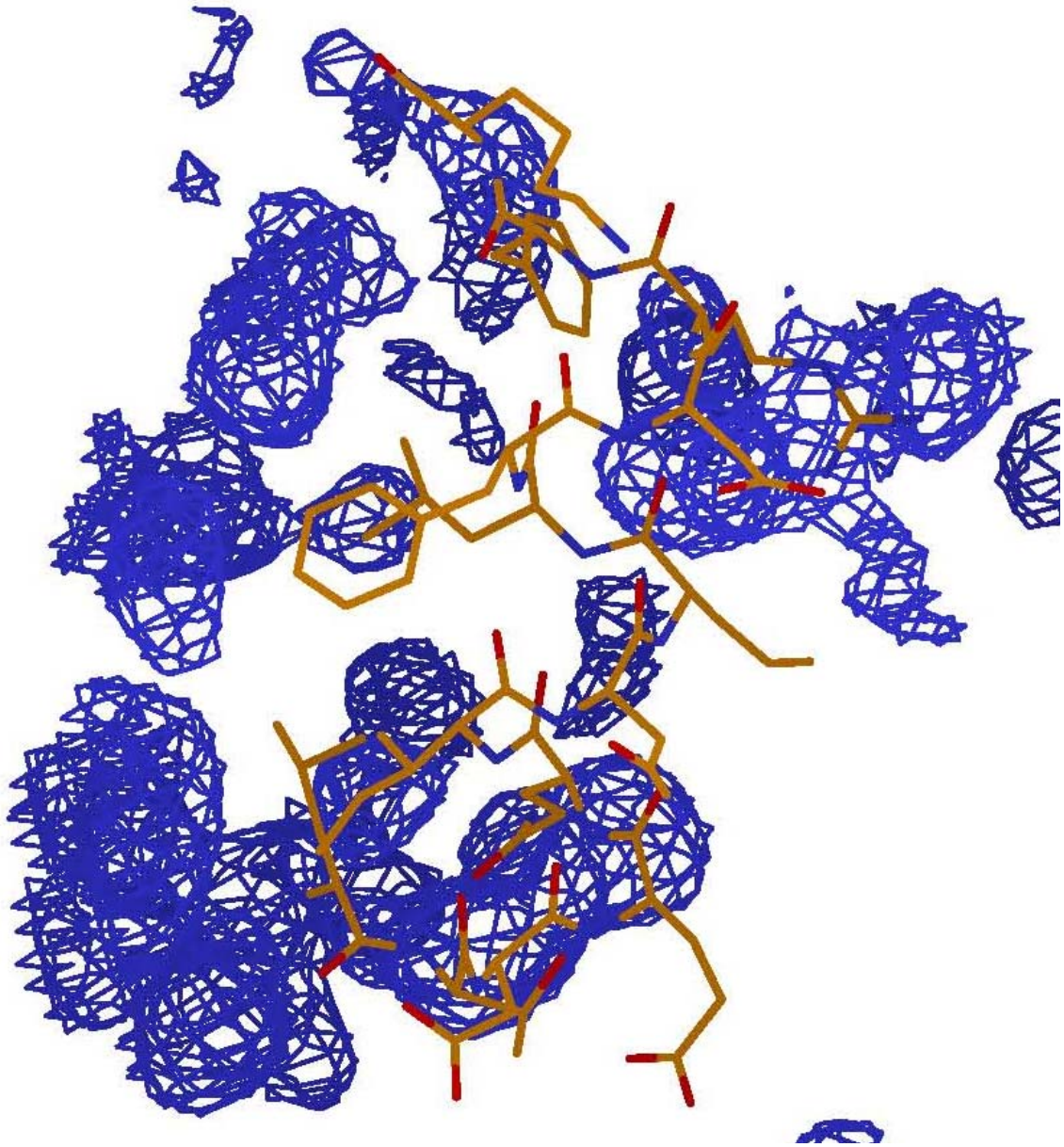
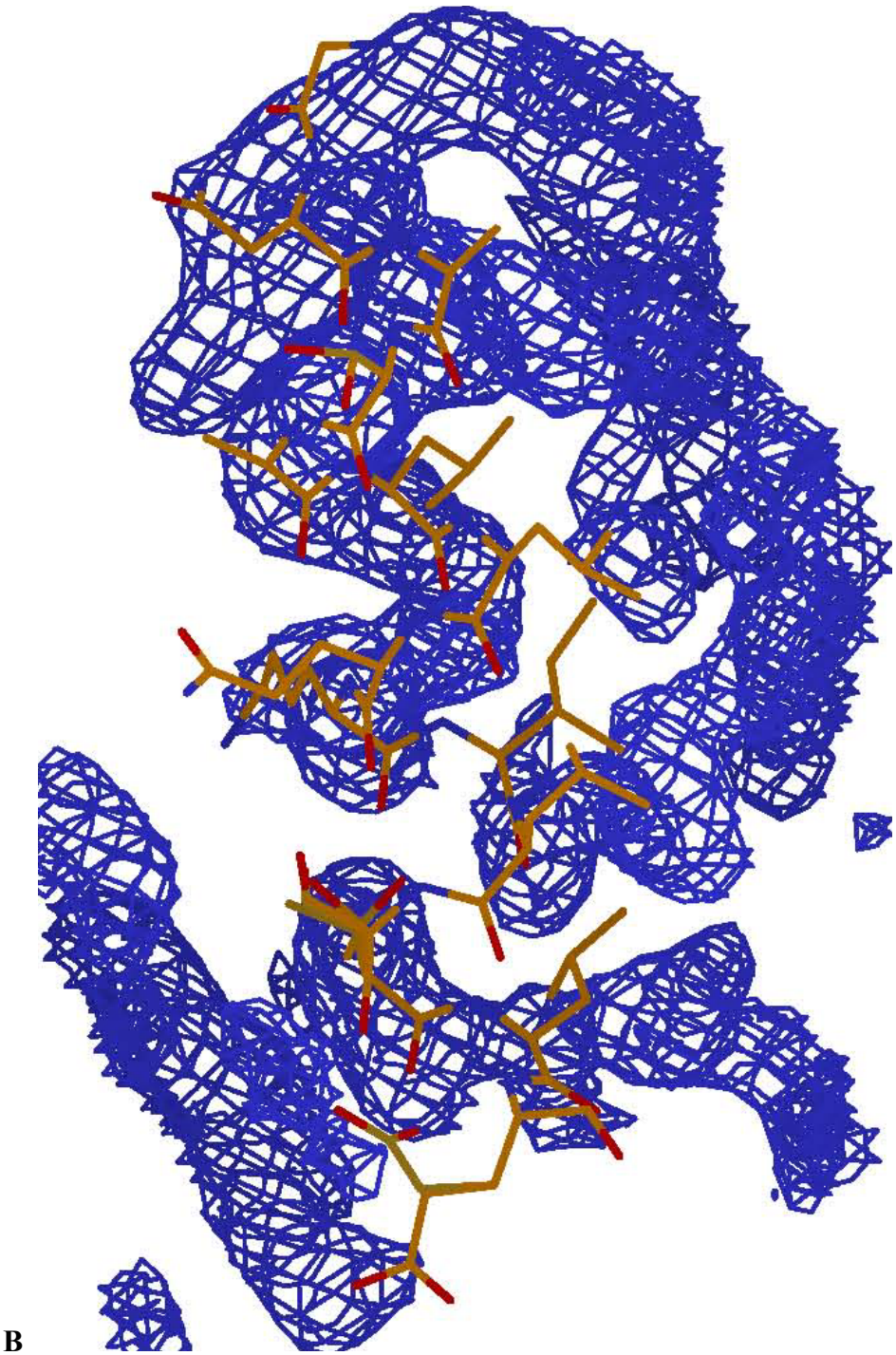


Figure 3.7: Graph of atomic displacement after tls refinement
Tls refinement using a single rigid body covering the entire protein coupled with least squares refinement reduced R/R_{free} by 4%. The reasoning behind the dramatic drop in R/R_{free} can be elucidated by examining the atomic displacement in the protein. Displacement is defined as the square root of the sum of the squared individual x, y, and z differences. Graph of the displacement versus atom number show that the majority of the movement of the model is within the difficult regions of helix 1, atom numbers 158-271, and 5, atom numbers 787-910.



A



B

Figure 3.8: F_o density of *Pfu*-542154 helix 1 and 5
(A) *Pfu*-532154 helix 1, a.a. 25-38, with the RESOLVE F_o electron density map showing the lack of density even with 1.7Å resolution data. **(B)** The F_o electron density map around helix 5, a.a. 108-122, showing the main chain breaks in electron density at the base of the helix. Both helices are on the same face of the protein towards the solvent channel. This helps explain the higher than normal R_{free} for a typical 1.7Å resolution dataset of the deposited final structure. Both images are generated using XFIT.

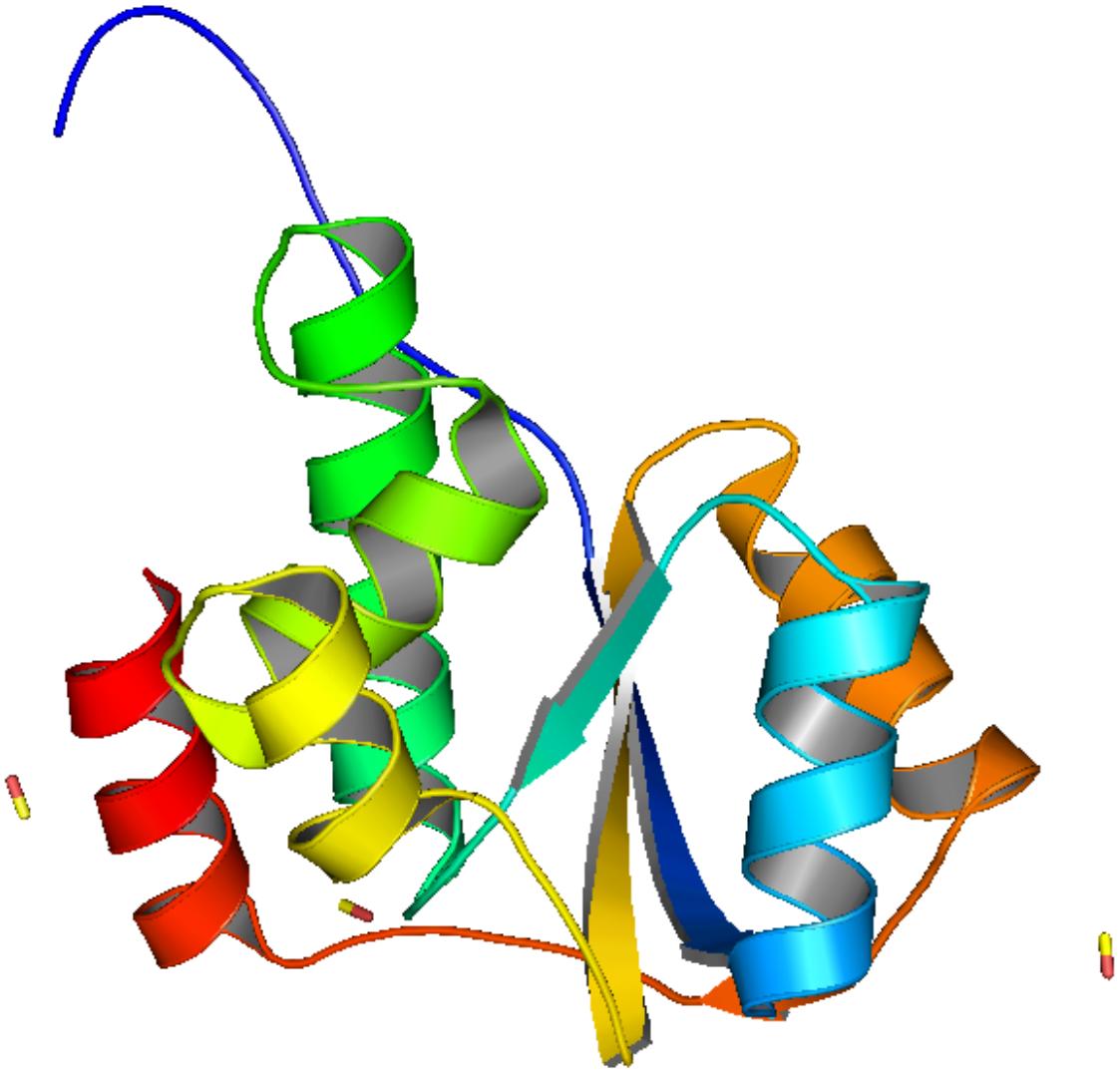


Figure 3.9: Structure of *Pfu*-542154
Ribbon diagram generated with PyMOL (DeLano, 2002) and colored N to C-terminus, blue to red respectively. The protein has two small domains, the α-helical domain on the left, helices 2, 3, 4, and 6; and the mixed α/β domain with a 3 stranded sheet and helices 1 and 5. Also seen in the image are the 3 molecules of methanol associated with the protein.

3.5: Structural characteristics of *Pfu*-542154

Pfu-542154 (PDBID 1ZD0), Figure 3.9, is a mixed α/β structure with 6 helices and a 3 stranded β-sheet. The structure appears to have two small domains; a small all α-helical domain composed of helices 2, 3, 4, and 6, and a mixed α/β domain composed of strands 1, 2, and 3, and helices 1 and 5. An interesting and somewhat unexpected structural feature is a disulfide bond bridging the long loop that connects helices 5 and 6 with the N-terminal portion of helix-2 (Figure 3.10). This helps stabilize the structure by locking the long flexible loop into place with the small α-helical domain. However, is this disulfide bond naturally occurring or an artifact of purification? *P. furiosus* is an obligate anaerobe living in the reducing environment of the hydrothermal vents off of the coast of Italy in the Mediterranean Sea. This organism should never be in contact with an oxidizing environment and therefore the thiols of the two cysteines should remain in the reduced form and considering that *P. furiosus* does not contain homologs of the proteins typically associated with disulfide bond formation. However, the bond obviously does exist and 5 other archeal structures in the Protein Data Bank (PDB) have a disulfide bond as well. Also, bioinformatics research from Mallick *et al.* (2002) carried out on archeal genomes characterized the probability that a cysteine could form a disulfide bond. While *P. furiosus* was not directly studied, two other species of *Pyrococcus*, *abyssi* and *horikoshii*, scored the highest probabilities within the study, 31 and 28% respectively. A general conclusion of the work was that the more hypothermophillic the organism, the higher the probability that its cysteines could be in disulfide bonds. This opens up an interesting avenue of further research on the debatable topic of disulfide bonds in archea.

An even more interesting structural discovery was the 6×His tag of one molecule sticking into a well defined cleft of a neighboring crystal packing related molecule (Figure 3.11a). It is

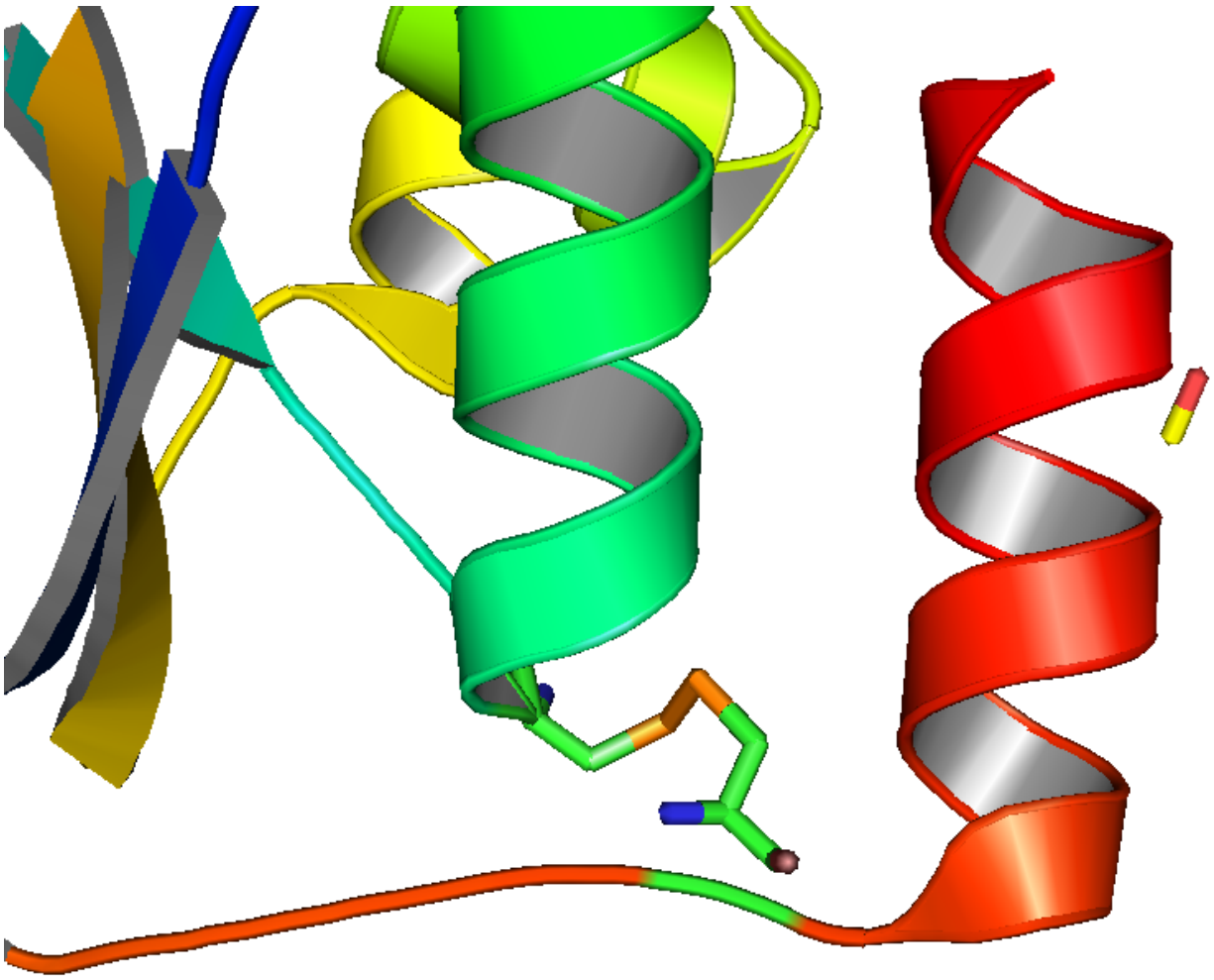
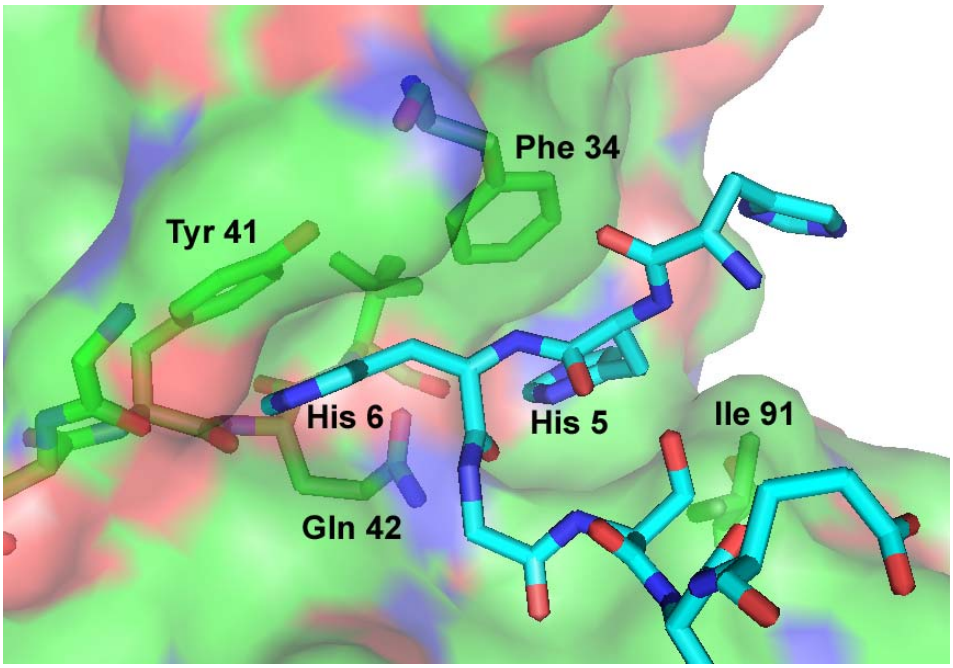


Figure 3.10 Disulfide bond in *Pfu*-542154
Disulfide bond between Cys-48 and 131 locking in helix 6 to the rest of the α-helical domain generated with PyMOL. Whether this disulfide bond is naturally occurring or an artifact of purification is a subject for debate and further research.

A



B

Figure 3.11: 6×His-tag coordination in *Pfu*-542154
(A) Surface representation of *Pfu*-542154 in PyMOL showing a cleft between the two domains binding the N-terminal of a crystal packing related molecule (cyan). **(B)** Close-up of the interactions holding the N-terminus in the cleft. His-5 is in a nice hydrophobic pocket formed between Ile-91 and Phe-36, also ring stacking with His-5. His-6 is involved in aromatic ring stacking with Tyr-41, and Gln-42 is hydrogen bonding with the mainchain atoms between His-5 and 6. The question remains as to what significance this plays in either enzymatic function or protein/protein interactions.

unusual for a 6xHis tag to be built into the model, but this situation is unique because of the coordination of His-5 and 6 (Figure 3.11b). Both histidines are stabilized in pockets where aromatic amino acids, Phe-34 for His-5 and Tyr-41 for His-6, are ringstacking with the histidine sidechains. Also, the sidechain of Gln-42 is hydrogen bonding directly with the mainchain of His6. The coordination helps explain why the crystals diffracted as well as they did, but it raises further questions about the function of *Pfu*-542154. This seems to suggest that *Pfu*-542154 coordinates some substrate in this cleft or has a protein partner that it complexes with for some purpose.

Pfu-542154 is a conserved hypothetical protein in *P. furiosus*, meaning that no known function exists for this protein. A sequence BLAST search against the PDB yielded no results and a PHI-,PSI-BLAST (Altschul *et al.*, 1997) searching the protein sequence yielded a homolog in two other species of *Pyrococcus* but no other positive hits based off of an E-value cutoff of 10^{-10}. Also, the search showed that *Pfu*-542154 belongs to the DUF509 pfam which is closely related to the COG1617 pfam. Unfortunately, DUF509 is a pfam of conserved hypothetical archaeal proteins with unknown function, and COG1617 is a family of conserved uncharacterized proteins. Neither one of these two families has a structural representative meaning that 1ZD0 is the first structure in either pfam. Since a sequence based alignment did not turn up a possible function, the next step was to look at a structure based alignment.

1ZD0 was uploaded to the European Bioinformatics Institute's, EBI, DALI structure based search program (Holm & Sander, 1994, 1994, 1996, 1998). The protein is taken and compared to every other sequence in the PDB to check for structural overlap. The results are emailed back as a rotation and translation matrix to align the possible solutions with the target. The best solution was PDB ID: 1V8C, MoaD related protein from *Thermus thermophillus*, with a

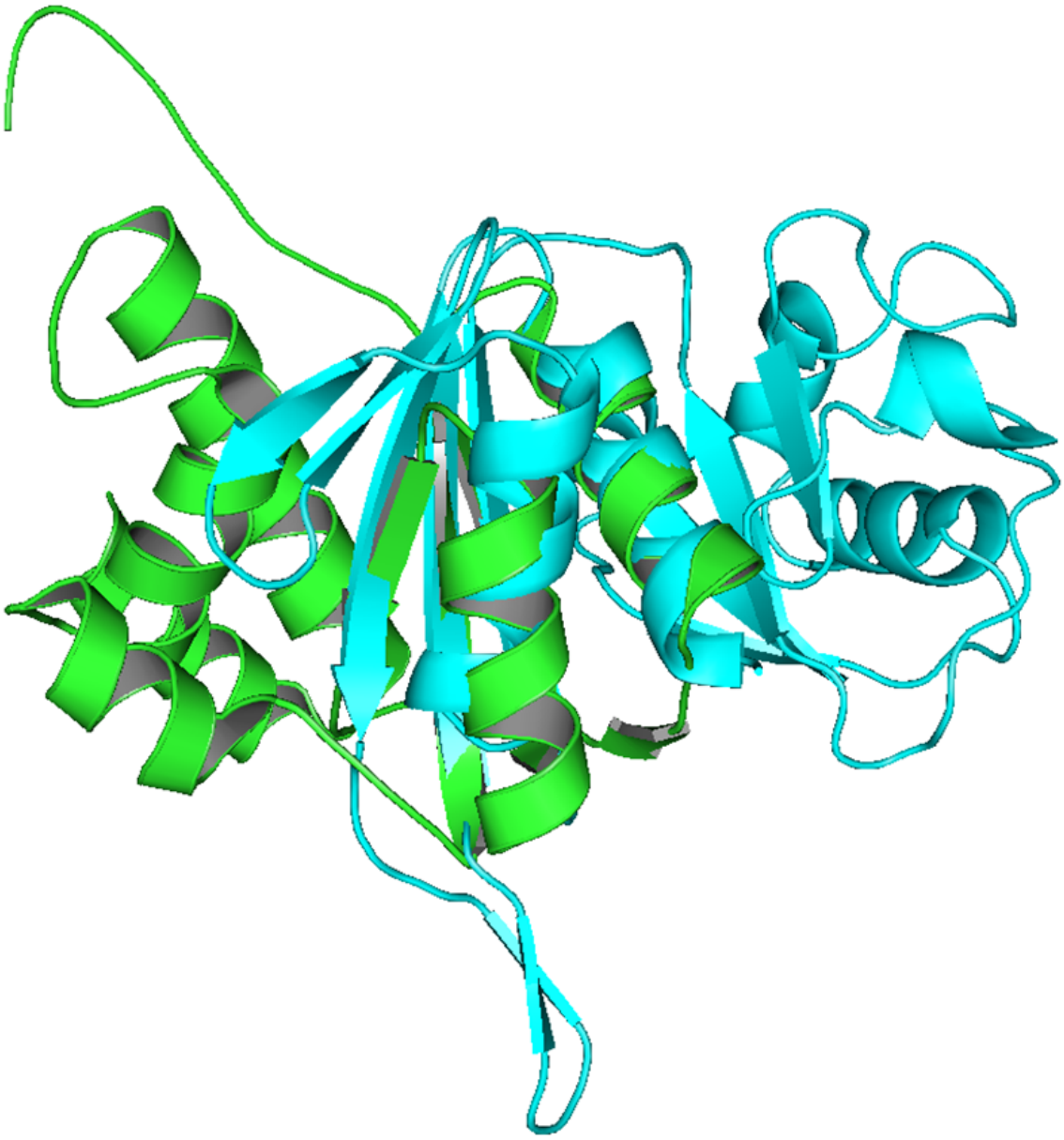


Figure 3.12: Structural alignment from the DALI server
The overlay of *Pfu*-542154, green, and the highest Z-score result from the DALI server 1V8C, cyan. The two structures have an RMSD of 2.6Å over the overlaid portion of the mixed α/β domain. While the two seem to possibly share this domain, the other domains of both proteins are left out. Also, the extra β-strands in 1V8C, to the left of the β-sheet of *Pfu*-542154, are in the area of the 6×His tag binding cleft. It is unlikely that function can be inferred from this structural comparison. Image generated with PyMOL.

mainchain root mean standard deviation (RMSD) of 2.1Å, Z-score of 3.7, and 13% sequence identity (Figure 3.12). As shown in Figure 3.12, the two proteins only slightly align in the α/β domain. While you could argue that this domain is conserved between the two proteins because of the RMSD value, the Z-score is lower than expected. Does this indicate that *Pfu*-542154 shares a similar function? It seems very doubtful. Looking at the cyan molecule of 1V8C, you can see that the β-sheet has two extra strands going to the left in the image. These two strands are filling the space where the 6XHis tag is binding. Also, the overlap between the two molecules doesn't take into account the other domains within either protein. Again, it seems very unlikely that these two proteins share a similar function.

Another way to determine possible function is to look and see if the gene is located in some operon at the gene localization level. In *P. furiosus*, *Pfu*-542154 is gene PF0523 with a conserved hypothetical annotation. Looking at The Insititute for Genomic Research's (TIGR) webpage there are three *Prococcus* genomes that have been sequenced, *furiosus, abyssi,* and *horikoshii* in the Comprehensive Microbial Resource database. In *furiosus* and *horikoshii*, the closest neighboring gene in the same direction is aspartate transaminase, but in *abyssi* the gene is isolated. Given the binding of the 6×His tag, it is possible that *Pfu*-542154 is forming a complex with the aspartate transaminase, but more research would be needed to verify.

In this case, unfortunately, none of the standard avenues of functional determination using the information from the structure or gene localization has provided any evidence for the possible function of *Pfu*-542154. Personal communication with UGA microarray scientists showed no increase or decrease of expression under the stress conditions they studied. Considering that this protein is a true unknown, any functional study would have to be long and exhaustive. Since this seems to be a *Pyrococcus* specific protein, it is unlikely that the function

of this protein will be ascertained within the near future. While *Pfu*-542154 lacks the aesthetics of a known function, the true beauty of this structure solution lies within the way it was phased.

Chapter 4

Conclusions

Sulfur-SAS represents an ideal situation for the protein crystallographer. In the most ideal situation, it allows them to merely mount the crystal and collect data from a single native crystal on a home rotating anode X-ray generator to solve the phase problem. There is no altering the contents of the unit cell or changing the protein to a non-wild type form; just data from native crystals. The easiest way to solve the phase problem is starting to mature into a viable option. While not commonplace yet, the future of protein crystallography will evolve towards only collecting data from unmodified protein crystals.

The idea of phasing an average protein *de novo* using only the anomalous signal of sulfur at a single wavelength has been theoretically possible since the simulation of B.C. Wang in the mid 1980's. The realization of this idea would come to fruition in 2000 with the structure solution of Obelin by Liu *et al*. Since that time, a number of small proteins with limited numbers of sulfur atoms have been solved at synchrotrons, Cu-rotating anode and, more recently, Cr-rotating anode X-ray generators. However, none have come close to the ratio of sulfur to amino acids, 57 amino acids per sulfur, originally established by the Wang simulation. The research presented here has a two fold affect on the area of sulfur-SAS. It expands the current ideas of the limitations associated with protein size and total number of sulfur positions able to be located in a sulfur-SAS experiment. Secondly, the *de novo* structure solution of *Pfu*-542154 is a realization of the original simulation study by phasing 50 amino acids per sulfur.

The simulation was successful not only in proving that a large protein could be solved with sulfur-SAS, but also validated the use of R_{as} as a monitoring tool for anomalous signal in a dataset. More importantly, it suggested a minimum R_{as} threshold value of 1.6 to strive for during data collection with the understanding that when the threshold value is achieved, phasing the protein should be successful. Also, the simulation demonstrated the additive affects of R_{as} in overcoming error associated with the data collection. Just as the original simulation established the first limits of sulfur-SAS, this simulation further expands those horizons into the realms of the larger single chained proteins and complexes. This simulation has the added affect of not only pointing to the future, but also showing you how to get there. With the advent of newer, more intense, and longer wavelength home X-ray generators, and increasingly more sensitive detectors, the future looks very bright for sulfur-SAS.

Even though the real world application of the simulation was not successful, the structure solution of *Pfu*-542154 represents the closest *de novo* structure solution to the original simulation study. *Pfu*-542154 was successfully phased using native data from 2 crystals collected on two separate Cr-rotating anode X-ray generators. This structure solution demonstrates that the anomalous signal of individual datasets is additive and systematic error from two different generator/detector setups can be overcome with more data in the form of redundancy. However, just adding any data will not be successful. A conscious effort must be made to incorporate data that contains more signal than noise. Careful examination of where to draw the lines in data incorporation during scaling can have a dramatic affect on obtaining a successful phasing outcome using sulfur-SAS.

Monitoring R_{as} in this case did prove somewhat successful. The 720° of MSC data had a R_{as} of 1.55 and by observing the electron density maps from the structure solving pipeline

results; the density was very close to being solvable. In this case, *Pfu*-542154 disagrees with the threshold value put forth by Fu *et al*. The numbers seem very close to each other, but the difference between a successful outcome and the need for more data is very slim in this case. While the merged UGA/MSC data's 1.44 R_{as} is significantly lower than 1.6, it seems to be an anomaly associated with reflection integration using D*TREK. While integrating with D*TREK, regardless of parameter modification about 10-12% of the reflections were being discarded. A direct correlation between amount of data and R_{as} would mean that the final value of R_{as} would be about 1.58-59 after a 10% increase and a maximum about 1.61-62 at 12%. If there was a direct correlation between R_{as} and data loss you could see that this data would be sufficient to solve the phase problem. While a direct correlation is unlikely, it is apparent that R_{as} increased to a level above 1.55 because the electron density maps became interpretable by auto-tracing protein sequence and visual inspection of the electron density for secondary structure elements.

The structure of *Pfu*-542154 has some interesting aspects both functionally and considering the environment the native organism strives in. The discovery of a disulfide bond in the structure was unexpected considering this is a *Pyrococcus* protein. Since *P. furiosus* is an obligate anaerobe, lives in a reducing environment at the opening of hydrothermal sulfur emitting underwater vents, and lacks the molecular machinery to establish and maintain disulfide bonds, it seems highly unlikely that this disulfide bond is naturally occurring instead of an artifact of expression and purification in an aerobic environment. However, protein structures from hyperthermophiles containing disulfide bonds do exist. All of these could in fact be artifacts or purification in an aerobic environment. Bioinformatic research on the genomes of hyperthermophiles did indicate that some hyperthermophiles, and specifically *Pyrococcus*, have

a 20-30% probability of having cysteines in disulfide bonds. Further research on *Pfu*-542154 could easily identify whether disulfides do exist in *Pyrococcus*. Two avenues could be explored; expressing, purifying, and crystallizing the protein in an anaerobic environment would provide direct evidence, or a simple mutagenesis study altering one of the cysteines to serine would disrupt the disulfide bond while maintaining the general hydrophilic properties of the sidechain. This way, the protein could be expressed, purified, and crystallized following the same protocols as before. Either way would help shed light onto the debatable topic of disulfide bonds in hyperthermophillic archea.

Another interesting structure aspect that has implications towards functions is the binding of a 6XHis-tag from a neighboring protein molecule in the crystal. The 6XHis-tag sits in a well defined cleft between the two domains of *Pfu*-542154. There is direct π-bond ring stacking between two of the histidine sidechains and Phe34 and Tyr41 of the protein. Electrostatic interactions also help stabilize the mainchain of the 6XHis-tag. Whether this has implications in substrate binding or protein-protein complex formation is yet to be seen. Sequence searches identify this protein as *Pyrococcus* specific and any structural comparisons show that this protein has no structural homolog in the PDB. Unfortunately, these do not help identify possible functions of the protein. Evidence for complex formation may be inferred from the gene localization of *Pfu*-542154. In *P. furiosus* and *horikoshii* this protein sits directly next to aspartate aminotransferase. It is possible that these two proteins could be forming a complex together. Along with the disulfide bond research, this is another avenue of work that can be further explored.

In the mid 1980's the initial bar was set for sulfur-SAS phasing. The structure solution of *Pfu*-542154 shows that we have met that bar and, using another simulation study, set an even loftier goal in the sulfur-SAS arena for crystallographers to strive for. Will it take another 20 years to reach this new goal? Only time will tell.

References:

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Nucleic Acids Res 25, 3389-3402.

Bijvoet, J. M. (1949). Kon.Ned.Akad.Wet. 52, 313.

Bijvoet, J. M. (1954). Nature 173, 888.

Box, G. E. P. a. M. E. M. (1958). Annals of Mathematical Statistics 29, 610-611.

Brown, J., Esnouf, R. M., Jones, M. A., Linnell, J., Harlos, K., Hassan, A. B. & Jones, E. Y. (2002). Embo J 21, 1054-1062.

Chen, L., Chen, L. R., Zhou, X. E., Wang, Y., Kahsai, M. A., Clark, A. T., Edmondson, S. P., Liu, Z. J., Rose, J. P., Wang, B. C., Meehan, E. J. & Shriver, J. W. (2004). J Mol Biol 341, 73-91.

Crick, F. H. C., and Magdoff, B.S. (1956). Acta Cryst. 9, 901-908.

Debreczeni, J. E., Bunkoczi, G., Girmann, B. & Sheldrick, G. M. (2003). Acta Crystallogr D Biol Crystallogr 59, 393-395.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.*

Fiala, G., and Stetter, K.O. (1986). Arch. Microbiol. 145, 56-61.

Friedel, G. (1913). Comptes Rendus 157, 1533.

Fu, Z. Q., Rose, J. P. & Wang, B. C. (2004). Acta Crystallogr D Biol Crystallogr 60, 499-506.

Gordon, E. J., Leonard, G. A., McSweeney, S. & Zagalsky, P. F. (2001). Acta Crystallogr D Biol Crystallogr 57, 1230-1237.

Harker, D. (1956). Acta Cryst. 9, 1-9.

Hendrickson, W. A. (1985). Trans. Am. Cryst. Assoc. 21, 11.

Hendrickson, W. A., and Teeter, M.M. (1981). Nature 290, 107-112.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). Embo J 9, 1665-1672.

Holm, L. & Sander, C. (1994). Nucleic Acids Res 22, 3600-3609.

Holm, L. & Sander, C. (1994). Proteins 19, 165-173.

Holm, L. & Sander, C. (1996). Nucleic Acids Res 24, 206-209.

Holm, L. & Sander, C. (1998). Nucleic Acids Res 26, 316-319.

Jenney, F. E., Jr., Brereton, P. S., Izumi, M., Poole, F. L., 2nd, Shah, C., Sugar, F. J., Lee, H. S. & Adams, M. W. (2005). J Synchrotron Radiat 12, 8-12.

Kahn, R., Fourme, R., Bosshard, R., Chiadmi, M., Risler, J. L., Dideberg, O. & Wery, J. P. (1985). FEBS Lett 179, 133-137.

Kendrew, J. C., Bodo, G., Dintzis, H.M., Kraut, J., and Wyckhoff, H.W. (1956). unpublished.

Kleywegt, G. J. (1992-2001). *MOLEMAN.*

Kleywegt, G. J., and Jones, T.A. (1996). Acta Crystallographica D 52, 826-828.

Krause, K. L., Volz, K. W. & Lipscomb, W. N. (1987). J Mol Biol 193, 527-553.

Lemke, C. T., Smith, G. D. & Howell, P. L. (2002). Acta Crystallogr D Biol Crystallogr 58, 2096-2101.

Li, S., Finley, J., Liu, Z. J., Qiu, S. H., Chen, H., Luan, C. H., Carson, M., Tsao, J., Johnson, D., Lin, G., Zhao, J., Thomas, W., Nagy, L. A., Sha, B., DeLucas, L. J., Wang, B. C. & Luo, M. (2002). J Biol Chem 277, 48596-48601.

Liu, Z. J., and Wang, B.C. (2002). *TAB.*

Liu, Z. J., Chen, L.R., Rosenbaum, G., Chrzas, J., Fu, Z.Q., Rose, J.P., Wang, B.C. (2004). *American Crystallographic Association Annual Meeting*. Chicago, IL

Liu, Z. J., Fu, Z.Q. and Wang B.C. (2002). *RNDME.*

Liu, Z. J., Lin, D., Tempel, W., Praissman, J. L., Rose, J. P. & Wang, B. C. (2005). Acta Crystallogr D Biol Crystallogr 61, 520-527.

Liu, Z. J., Vysotski, E. S., Chen, C. J., Rose, J. P., Lee, J. & Wang, B. C. (2000). Protein Sci 9, 2085-2093.

Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Proteins 50, 437-450.

Mallick, P., Boutz, D. R., Eisenberg, D. & Yeates, T. O. (2002). Proc Natl Acad Sci U S A 99, 9679-9684.

McRee, D. E. (1999). J Struct Biol 125, 156-165.

McRee, D. E. (1999). Journal of Structural Biology 125, 156-165.

Micossi, E., Hunter, W. N. & Leonard, G. A. (2002). Acta Crystallogr D Biol Crystallogr 58, 21-28.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Acta Crystallogr D Biol Crystallogr 53, 240-255.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. Acedemic Press.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Nat Struct Biol 6, 458-463.

Phillips, J. D., Whitby, F. G., Warby, C. A., Labbe, P., Yang, C., Pflugrath, J. W., Ferrara, J. D., Robinson, H., Kushner, J. P. & Hill, C. P. (2004). J Biol Chem 279, 38960-38968.

Ramachandran, G. N., and Srinivasan, R. (1970). Fourier Methods in Crystallography. New York: Wiley (Interscience).

Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B. & Dunn, D. M. (2001). Methods Enzymol 330, 134-157.

Schneider, T. R. & Sheldrick, G. M. (2002). Acta Crystallogr D Biol Crystallogr 58, 1772-1779.

Stura, E. A., and Wilson, I.A. (1991). J Cryst Growth 110, 270-282.

Swaminathan, S. & Eswaramoorthy, S. (2000). Nat Struct Biol 7, 693-699.

Tanner, J. J., Smith, P. E. & Krause, K. L. (1993). Protein Sci 2, 927-935.

Terwilliger, T. C. (2000). Acta Crystallogr D Biol Crystallogr 56 (Pt 8), 965-972.

Terwilliger, T. C. (2003). Acta Crystallogr D Biol Crystallogr 59, 38-44.

Terwilliger, T. C. & Berendzen, J. (1999). Acta Crystallogr D Biol Crystallogr 55 (Pt 4), 849-861.

Wang, B. C. (1983). *FCAL.*

Wang, B. C. (1985). Methods Enzymol 115, 90-112.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). Acta Crystallogr D Biol Crystallogr 57, 122-133.

Yang, W., Hendrickson, W. A., Crouch, R. J. & Satow, Y. (1990). Science 249, 1398-1405.

Yang, W., Hendrickson, W. A., Kalman, E. T. & Crouch, R. J. (1990). J Biol Chem 265, 13553-13559.

Zhang, M., White, T. A., Schuermann, J. P., Baban, B. A., Becker, D. F. & Tanner, J. J. (2004). Biochemistry 43, 12539-12548.

APPENDICES

Appendix A

De-twinning and Structure Solution of a Putative Acetyltransferase from *Pyrococcus furiosus*, *Pfu*-35386

Abstract

The production of diffraction quality crystals is rarely straight forward. According to the latest xml file released, the structural genomics centers of the Protein Structure Initiative are reporting a purified protein to crystal success rate of almost 40% but, the purified protein to diffraction quality crystal rate is, on average, only around 13%. The Southeast Collaboratory for Structural Genomics (SECSG) target protein *Pfu*-35386 was unable to achieve diffraction quality crystals through screening and optimization because of twinning. Streak seeding was employed and single diffraction quality crystals obtained. The structure was solved using platinum single wavelength anomalous scattering (SAS). This structure represents a putative acetyltransferase from *Pyrococcus furiosus* that is structurally similar to the GCN5-related N-acetyltransferase superfamily found in all kingdoms of life.

Keywords: acetyltransferase, single wavelength anomalous scattering, structural genomics, streak seeding, GNAT

Introduction

After generating a purified soluble protein, the production of a diffraction quality crystal sufficient for structure solution represents the last bottle neck of structural genomics. Structural genomics centers across the country are only having, on average, a 13% success rate from purified protein to diffraction quality crystal. These proteins represent the "easiest" targets available and yet diffraction quality crystals are still

difficult to obtain. Failing to produce diffraction quality crystals via high throughput methods will often end the structure solution of that protein. A putative acetyltransferase from *Pyrococcus furiousus*, gene PF0028 or open reading frame *Pfu*-35386, failed the crystallization standards set forth by the Southeast Collaboratory for Structural Genomics (SECSG) and was removed from high throughput structure determination.

Pfu-35386 encodes a 150 amino acid protein that contains a GCN5-related N-acetyltransferase (GNAT) domain in the latter two thirds of its sequence. The GNAT superfamily spans all kingdoms of life with over 10,000 representatives and utilizes acetyl-CoA as the acyl donor for transfer to their substrates. The superfamily encompasses a diverse set of substrates and its members are involved in areas such as antibiotic resistance, histone acetylation, biosynthesis of melatonin in humans, and generation of the branched cell wall in gram positive bacteria. However, this diverse group shares a relatively low sequence homology (Shaw *et al.*, 1993) but with conserved structural homology (Dyda *et al.*, 2000). *Pfu*-35386 represents a *Pyrococcus furiosus* specific member of the GNAT superfamily.

2. Materials and methods

2.1 Crystallization, data collection and structure determination of *Pfu*-35386

Pfu-35386 was expressed and purified using the standard SECSG Protein Production Core protocol (Jenney *et al.*, 2005). Following the SECSG Crystallization Core protocol (Shah *et al.*, 2005), modified microbatch crystallization experiments, using the Douglas Instruments Oryx-6 robot, generated plate clusters in Hampton Research's MembFac condition #19, 100mM magnesium chloride hexahydrate, 100mM tri-sodium

citrate pH 5.6, and 4% (v/v) 2-methyl-2,4-pentanediol (MPD). Grid screen optimization around MembFac #19 (50-250mM magnesium chloride, 100mM tri-sodium citrate pH 5.2-6.6, and 2-10% (v/v) MPD) thickened the plates sufficiently to test for diffraction (Figure 1a), but always resulted in severe split spot diffraction pattern. *Pfu*-35386 was detwinned by lowering the protein concentration from 40 mg/mL to 20 mg/mL, switching the crystallization experiment from modified microbatch with an 80:20 paraffin to silicon oil mixture to a batch experiment with a pure paraffin oil overlay where water does not evaporate out through the oil layer, and ultimately by streak seeding (Stura, 1991). The streak seeding protocol used involved setting up the MembFac #19 optimization screen with a pure paraffin oil overlay and allowing it to equilibrate over night. A cat whisker was used to touch plate clusters from conditions that reached nucleation and then streak through one well of the equilibrated batch optimization tray with no signs of a crystal. The process was repeated until all wells without crystals were streaked. Crystals typically appeared within 1 day and reached maximum crystal dimensions of 250 × 150 ×100μm 10 days after streaking (Figure 1b). 0.2μL of a 20% glycerol cryoprotectant solution, 8μL of precipitating solution and 2μL of 100% glycerol, was injected directly through the oil layer into the crystallization drop using a 0.5μL Hamilton 7000 syringe (Fisher #14-813-100) and allowed to equilibrate for 5 minutes. For derivatives, a 50μm or smaller crystal of potassium tetrachloroplatinate (II) (Hampton Research Heavy Atom Screen Pt HR2-442 #1) or potassium iodide (Sigma P4286), was placed directly into the crystallization drop after cryoprotection equilibration and allowed to sit for another 5 minutes. 100K data on the Pt derivative crystals were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline, and 100K data on the KI

derivative crystals were collected at Structural Biology Center (SBC) 19-BM beamline both at the Advanced Photon Source, Argonne National Laboratory. All data were processed with HKL2000 (Otwinowski, 1997), and data from two of the Pt derivatives were merged together with the resulting .sca file fed into the SECSG's SCA2STR pipeline (Liu *et al.*, 2005). SOLVE/RESOLVE (Terwilliger, 2000, 2003, Terwilliger & Berendzen, 1999) were used to phase the structure to 3.0Å and ARP/WARP (Perrakis *et al.*, 1999) produced an initial partial structure with an R of 30.8%. A potassium iodide derivative diffracted to the highest resolution, 1.9Å, and was used for refinement. Diffraction and structural statistics are presented in Table 1. Rounds of positional, B-factor and simulated annealing refinement were carried out using CNS (Brunger *et al.*, 1998) and a random selection of 10% of the reflection data were excluded and used to calculate the free R (R_{free}) as a monitor of model bias (Brunger, 1993). Model building and corrections were carried out using XFIT (McRee, 1999) and the final R/R_{free} of the deposited structure (PDB ID: 1VKC) after structural validation using MOLPROBITY (Lovell *et al.*, 2003) was 21.2/24.4%.

3. Results and discussion

3.1 De-twinning crystals of *Pfu*-35386

As a structural genomics target, *Pfu*-35386 was sent through the Crystallization Core of SECSG and initial crystallization screening using a modified microbatch under 80:20 paraffin to silicon oil showed Hampton Research's MembFac condition #19 to produce thin plate clusters. Grid screening around MembFac #19 produced crystals of sufficient thickness in all three dimensions to test for diffraction (Figure 1a). The crystals diffracted to 2.6Å but the diffraction spots were always severely twinned. Continued

optimization and additive screening never produced a crystal whose diffraction spots weren't split. At this point the target was taken out of the high throughput crystallization pipeline. De-twinning *Pfu*-35386 was directed by the observations made from the grid screen optimization experiment. Crystals formed within 24 hours of mixing and that is very fast for modified microbatch. It is possible that water is not evaporating out of the drop and the crystals are forming under batch conditions. To test this, the 80:20 paraffin to silicon oil layer mixture was replaced with pure paraffin oil, which does not allow water to evaporate out of the drop, changing the crystallization to a pure batch experiment. Plate clusters still formed within a day verifying the nature of the crystallization. Next, crystals formed even though protein precipitation always appeared when the protein and precipitating solutions were initially mixed; a clear indication that the protein concentration is too high. Reducing the protein concentration by half and setting the optimization plate back up resulted in noticeably less protein precipitating during initial mixing but slowed the formation of crystals from within 1 day to 3 days. However, even with reduced protein concentration and under batch conditions, the super nucleus of the plate cluster still persisted. In order to deposit single nuclei, streak seeding was used after mapping out the boundary between nucleation and the metastable region. The optimization grid screen was set up and allowed to sit for 3 days and afterwards a cat whisker was used to touch the plate clusters that appeared in the tray and then streak the wells without crystals present. Crystals along the streak line appeared within 1 day and reached their maximum dimensions within 10 days after streaking (Figure 1b). Crystals were mounted with a 20% glycerol cryoprotectant and tested for twinned split spot diffraction. Apparent in the diffraction pattern of Figure 2, the combination of slowing

down crystal formation and streak seeding was sufficient to de-twin the crystals of *Pfu*-35386 for structure solution.

3.2 Structure of *Pfu*-35386 and similarity to the GNAT superfamily

Once de-twinned, *Pfu*-35386 was phased using platinum single wavelength anomalous scattering (SAS) and initially traced utilizing the SECSG SCA2STR pipeline. CNS refinement and MOLPROBITY structure validation resulted in the crystallographic dimer structure deposited in Figure 3. *Pfu*-35386 is a mixed α/β structure with a 4 stranded β-sheet wrapping around α-helix 4 characteristic of the GNAT fold superfamily. Submission to the DALI server (Holm & Sander, 1994, 1994, 1996, 1998) confirmed the structural similarity to the GNAT fold super family (Table 2). The top two results have known functions, animoglycoside 6'-N-acetyltransferase (PDB ID: 1S3Z) (Vetting *et al.*, 2004) and Hpa2 histone acetyltransferase (PDB ID: 1QSM) (Angus-Hill *et al.*, 1999), and when overlaid with *Pfu*-35386 structural similarity (Figure 4a) in the absence of sequence similarity (Figure 4b), 16 and 19% identity with 1S3Z and 1QSM respectively, becomes readily apparent. Sequence based BLAST search (Altschul *et al.*, 1997) also verifies the possible functional annotation with the highest hits , $<10^{-6}$, all from COG 0454, histone Hpa2 acetyltransferases, corresponding to amino acids 57-150 of *Pfu*-35386 (data not shown). Comparing the binding site of Acetyl-CoA in 1S3Z or 1QSM to the same location on 1VKC, you see that this binding site is located at the crystallographic dimer interface. If modeled into 1VKC, the acetyl-CoA would sterically disrupt the dimer interface helping to explain the current MembFac #19 co-crystallization difficulty being experienced and suggests that re-screening the protein-ligand complex is necessary, already underway. While substrate identification may not be easily resolved

due to the numerous different targets available within the cell, it appears that the annotation of *Pfu*-35386 as an acetyltransferase is correct based on structural and sequence information.

Acknowledgements

Work was supported in part by funds from the National Institutes of Health (GM62407), The Georgia Research Alliance and The University of Georgia Research Foundation. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. W-31-109-ENG-38. SER-CAT supporting institutions may be found at http://www.ser-cat.org/members.html. Use of the Argonne National Laboratory Structural Biology Center beamlines at the Advanced Photon Source, was supported by the U. S. Department of Energy, Office of Energy Research, under Contract No. W-31-109-ENG-38. The author would also like to acknowledge Michi Izumi, Dr. Frank Jenney, and Dr. Michael Adams for the production of purified protein, Dr. Randy Alkire for data collection assistance at SBC, and Dr. Nadia Leyarovska for data collection assistance at SER-CAT.

A

B

Figure 1: Twinned and de-twinned crystallization of *Pfu*-35386
(A) Crystal image of *Pfu*-35386 from MembFac #19 grid screen optimization, showing sufficient thickening of the plate cluster for diffraction analysis. The thickest plate in the image has the dimensions 200 × 200 × 75μm **(B)** Single crystals of *Pfu*-35386 were grown by reducing protein concentration, altering the crystallization experiment, and ultimately through streak seeding. The image shows the streak line through the protein precipitate and final crystal dimensions of 250 × 150 × 100μm.

Diffraction Statistics

	KI Derivative	Merged Pt Derivatives
Spacegroup	$P2_1$	$P2_1$
Unit Cell		
a (Å)	46.92	46.91
b (Å)	67.39	67.45
c (Å)	49.41	49.35
β (Å)	91.55	92.203
Wavelength (Å)	0.97	0.97
Resolution (Å)	1.9	2
Redundancy	7.8 (5.3)	13.9 (10.2)
Completeness	99 (95.6)	98.4 (91.4)
R_{sym} (%)	4.9 (31.7)	
R_{merge} (%)		8.2 (28.8)
I/σ	31.78 (2.64)	41.95 (12.34)

Structure Statistics

R	21.2
R_{free}	24.4
R.M.S. Bond Length (Å)	0.013
R.M.S. Bond Angle (°)	1.078
Average B-factor (Å^2)	28.811

Table 1: Diffraction and structural statistics of *Pfu*-35386
$R = \Sigma ||F_o| - |F_c|| / \Sigma |F_o|$
$R_{free} = \Sigma_{test} ||F_o| - |F_c|| / \Sigma_{test} |F_o|$ where "test" refers to a randomly selected 10% of the reflection set aside prior to refinement.

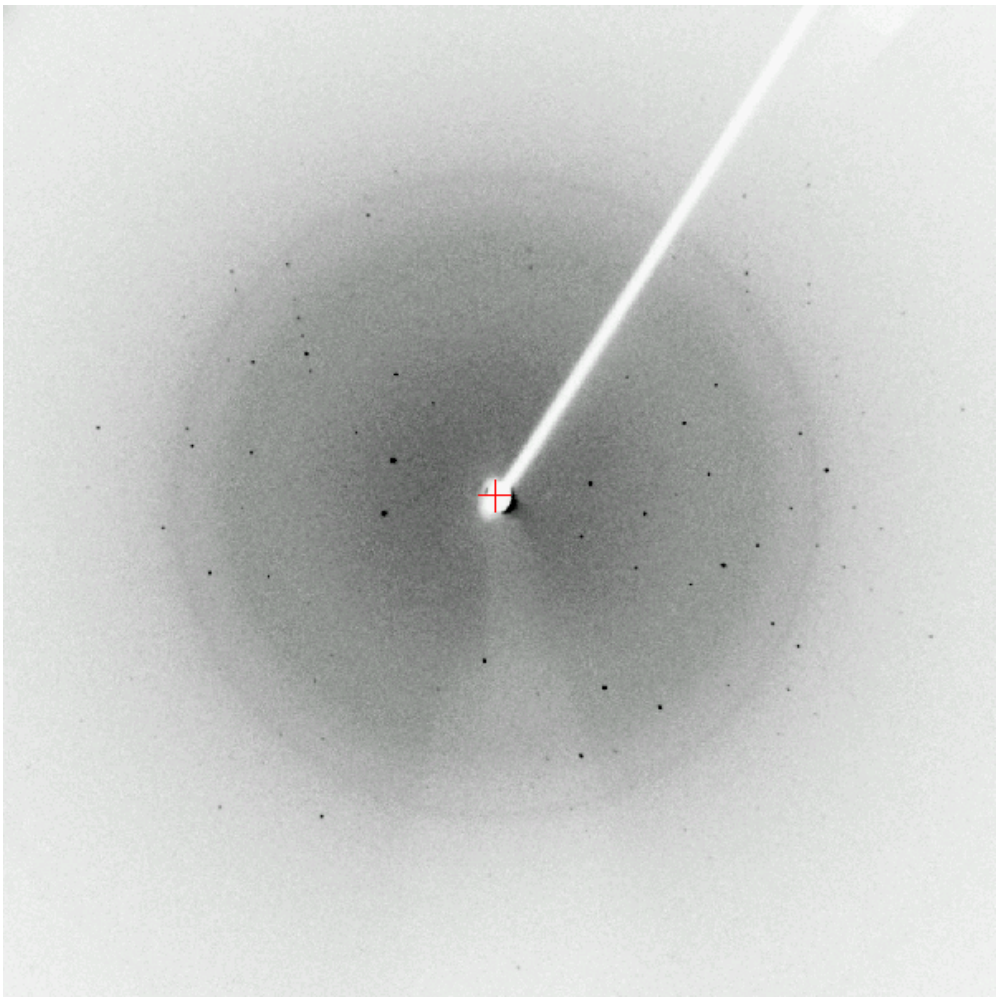


Figure 2: Diffraction of streak seeded *Pfu*-35386 crystals
Diffraction experiment image from a streak seeded *Pfu*-35386 crystal, soaked with potassium tetrachloroplatinate, using a Rigaku FRD Cu-rotating anode X-ray generator with Saturn92 CCD detector showing that diffraction spots are no longer split.

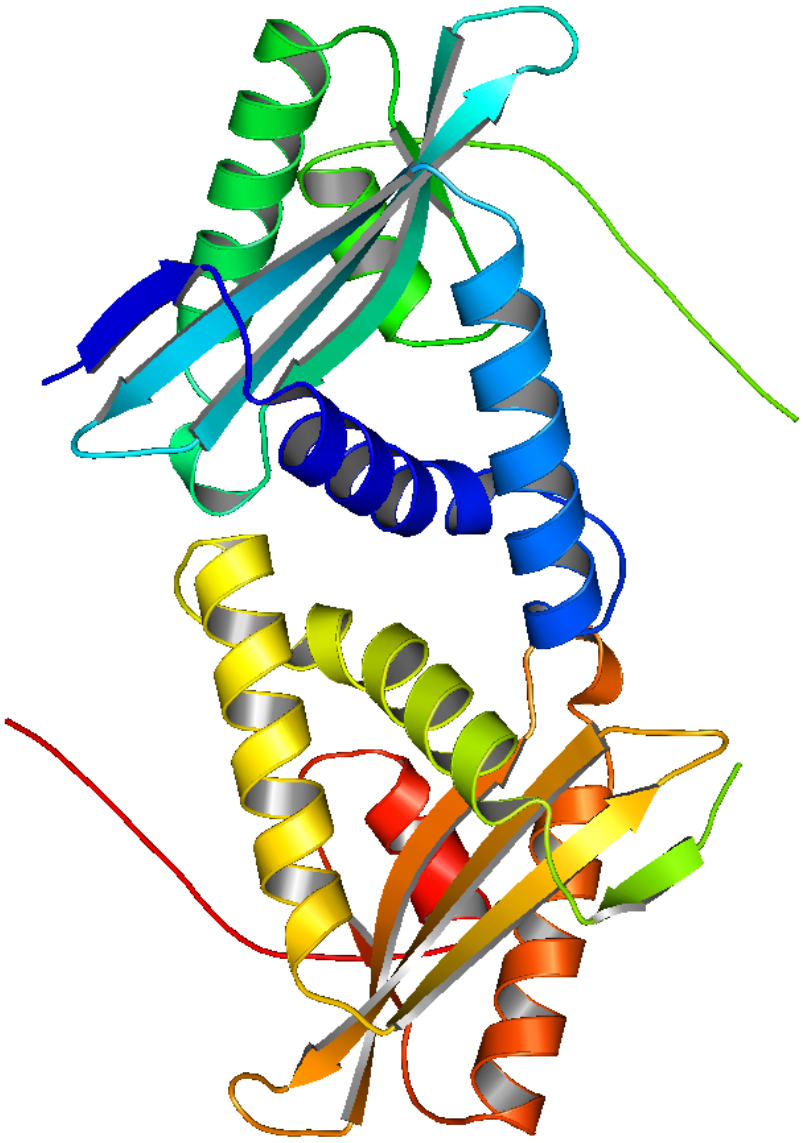
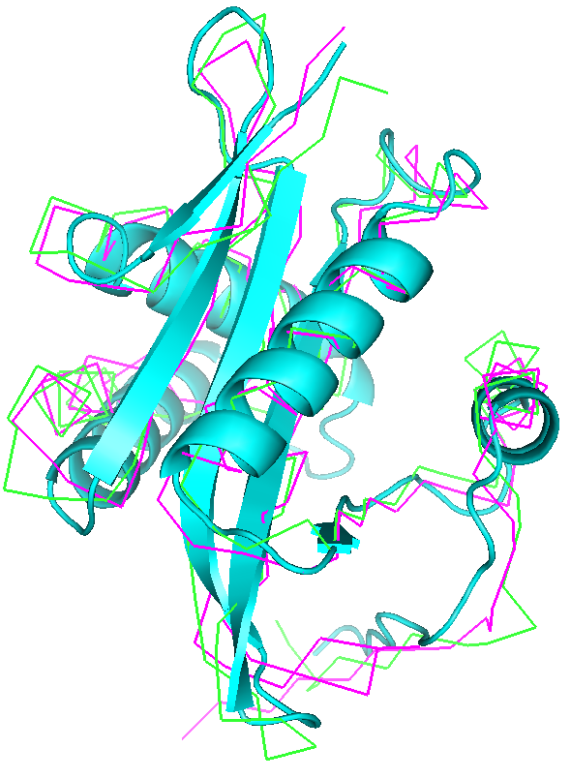


Figure 3: Structure of *Pfu*-35386
Ribbon diagram of the crystallographic dimer of *Pfu*-35386 (PDB ID: 1VKC) colored blue to red from N to C-terminus across the two monomers. Image generated using PYMOL (DeLano, 2002).

NR.	STRID1	STRID2	Z	RMSD	LALI	LSEQ2	%IDE
1	1vkc	1s3z	16.5	2.9	137	147	16
2	1vkc	1qsm	15.4	2.7	140	150	19
3	1vkc	1tiq	14.4	2.8	140	166	24
4	1vkc	1on0	14.2	3.4	141	152	21
5	1vkc	1wk4	14	3	142	173	18
6	1vkc	1q2y	13.6	2.1	122	140	19
7	1vkc	1b87	13.6	2.7	130	181	19
8	1vkc	1ozp	13.3	2.5	132	290	18
9	1vkc	1pu9	13.2	2.6	125	163	13
10	1vkc	1cjw	13.2	2.6	129	166	16
11	1vkc	1i21	13.1	4.1	135	155	18
12	1vkc	1xeb	12.8	2.4	123	146	16
13	1vkc	1u6m	12.8	3.4	136	188	26
14	1vkc	1mk4	12.7	2.6	126	157	16
15	1vkc	1vhs	12.5	3	133	161	21
16	1vkc	1m44	11.7	2.4	125	177	16
17	1vkc	1ygh	11.6	2.9	123	164	11
18	1vkc	1ne9	11.2	3.3	130	335	15
19	1vkc	1sqh	11	2.7	119	293	13
20	1vkc	1bo4	11	3.6	114	136	17
21	1vkc	1lrz	10.9	3	122	400	9
22	1vkc	1yx0	10.7	3.6	129	151	20
23	1vkc	1ro5	10.7	2.7	128	197	8
24	1vkc	1yre	10.3	3.1	130	183	12
25	1vkc	1yk3	8.6	3.2	132	198	10

Table 2: DALI Server results using IVKC as the search model



A

```
1VKC_A          GSEYTIVDGEEYIEEIKKLDREISYSFVRFPISYEEYEERHEELFESLLSQGEHKFFVAL 60
1QSM_A          -SEDNIT--VRFVTENDKEGWQRLWKSYQDFYEVSFPDDLDDFNFGRFLDPNIKMWAAVA 57
1S3Z_A          ---GLVPRGSHMDIRQMNKTHLEHWRGLRKQLWPGHPDDAHLADGEEILQADHLASFIAM 57
                     :    .   .  :      :   :        :: .      :*. .      .

1VKC_A          NERS--ELLGHVWICITLDTVDYVKIAYIYDIEVVKWARGLGIGSALLRKAEEWAKERGA 118
1QSM_A          VESSSEKIIGMINFFNHMTTWDFKDKIYINDLYVDENSRVKGAGGKLIQFVYDEADKLGT 117
1S3Z_A          ADGVAIGFADASIRHDYVNGCDSSPVVFLEGIFVLPSFRQRGVAKQLIAAVQRWGTNKGC 117
                 :     : .       :   *     :: .: *    *  * .  *:  .   . : *

1VKC_A          KKIVLRVEIDNP-AVKWYEERGYKARA-LIMEKPI- 151
1QSM_A          PSVYWCTDESNHRAQLLYVKVGYKAPK-ILYKRKGY 152
1S3Z_A          REMASDTSPENTISQKVHQALGFEETERVIFYRKRC 153
                 .:   .. .*  :   :   *::    ::  :
```

B

Figure 4: Structural and sequence similarity *Pfu*-35386 to known GNATs
(A) Overlay of *Pfu*-35386 (PDB ID: 1VKC, the cyan ribbon) and the two highest results from the DALI Server, animoglycoside 6'-N-acetyltransferase (PDB ID: 1S3Z, magenta C_α-trace, RMSD 2.9Å) and Hpa2 histone acetyltransferase (PDB ID: 1QSM, green C_α-trace, RMSD 2.7Å) using PYMOL. **(B)** CLUSTALW (Thompson *et al.*, 1994) sequence alignment of the three structures in (A) showing the lack of sequence identity, 16% identity between 1VKC and 1S3Z and 19% between 1VKC and 1QSM, where "*" means that the residues or nucleotides in that column are identical in all sequences in the alignment, ":" means that conserved substitutions have been observed, according to the CLUSTALW color table on the webpage http://www.ebi.ac.uk/embnet.news/vol4_3/clustalw1.html, "." means that semi-conserved substitutions are observed.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Nucleic Acids Res 25, 3389-3402.

Angus-Hill, M. L., Dutnall, R. N., Tafrov, S. T., Sternglanz, R. & Ramakrishnan, V. (1999). J Mol Biol 294, 1311-1325.

Brunger, A. T. (1993). Acta Crystallogr D Biol Crystallogr 49, 24-36.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Acta Crystallogr D Biol Crystallogr 54 (Pt 5), 905-921.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.*

Dyda, F., Klein, D. C. & Hickman, A. B. (2000). Annu Rev Biophys Biomol Struct 29, 81-103.

Holm, L. & Sander, C. (1994). Nucleic Acids Res 22, 3600-3609.

Holm, L. & Sander, C. (1994). Proteins 19, 165-173.

Holm, L. & Sander, C. (1996). Nucleic Acids Res 24, 206-209.

Holm, L. & Sander, C. (1998). Nucleic Acids Res 26, 316-319.

Jenney, F. E., Jr., Brereton, P. S., Izumi, M., Poole, F. L., 2nd, Shah, C., Sugar, F. J., Lee, H. S. & Adams, M. W. (2005). J Synchrotron Radiat 12, 8-12.

Liu, Z. J., Lin, D., Tempel, W., Praissman, J. L., Rose, J. P. & Wang, B. C. (2005). Acta Crystallogr D Biol Crystallogr 61, 520-527.

Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Proteins 50, 437-450.

McRee, D. E. (1999). J Struct Biol 125, 156-165.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. Acedemic Press.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Nat Struct Biol 6, 458-463.

Shah, A. K., Liu, Z. J., Stewart, P. D., Schubot, F. D., Rose, J. P., Newton, M. G. & Wang, B. C. (2005). Acta Crystallogr D Biol Crystallogr 61, 123-129.

Shaw, K. J., Rather, P. N., Hare, R. S. & Miller, G. H. (1993). Microbiol Rev 57, 138-163.

Stura, E. A., and Wilson, I.A. (1991). J Cryst Growth 110, 270-282.

Terwilliger, T. C. (2000). Acta Crystallogr D Biol Crystallogr 56 (Pt 8), 965-972.

Terwilliger, T. C. (2003). Acta Crystallogr D Biol Crystallogr 59, 38-44.

Terwilliger, T. C. & Berendzen, J. (1999). Acta Crystallogr D Biol Crystallogr 55 (Pt 4), 849-861.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Nucleic Acids Res 22, 4673-4680.

Vetting, M. W., Magnet, S., Nieves, E., Roderick, S. L. & Blanchard, J. S. (2004). Chem Biol 11, 565-573.