GIVEN A MODERATELY DIFFRACTING CRYSTAL, DOES THE CHOICE OF DATA REDUCTION APPROACH EFFECT THE S-SAD PHASING RESULT?

by

JAMES TUCKER SWINDELL II

(Under the Direction of Bi-Cheng Wang)

ABSTRACT

The inherent difficulty of solving the phase problem in macromolecular crystallography has been somewhat alleviated due to the advances in all areas of the diffraction experiment including protein purification, crystallization, X-ray sources, cryo-crystallography, detection technologies, and data reduction.

A phasing approach, Sulfur-ISAS (**I**terative **S**ingle Wavelength **A**nomalous **S**cattering), aimed at removing the generic necessity of either including selenium via protein engineering or derivatization using heavy atoms was hypothesized in 1985 by B.C. Wang and is, albeit gradually, increasing in popularity.

We have recently determined the structure of AF1382, a small 95-residue protein encoded by *Archaeoglobus fulgidus* by S-SAD (**S**ingle Wavelength **A**nomalous **D**iffraction) phasing using two 360° data sets collected on a moderately (2.65Å) diffracting crystal. Producing an interpretable electron density map at the time of data collection required additional assistance from the SER-CAT support staff. The eventual phase solution was achieved by merging the two data sets in conjunction with expert processing, involving both the HKL2000-GUI and command line scaling via SCALEPACK.

The downsides associated with removing the need for experienced crystallographers when dealing with data reduction becomes most evident when a data set does not yield an

immediate structure solution, as in the AF1382 case. Due to the reliance on a single data reduction program, characteristic of many of this generations Structural Biologist, using point-an-click processing without a fundamental understanding programs operation causes data to often be discarded in lieu of mounting a second or third crystal in hopes of a better processing result. Discarding data for this reason illustrates pitfalls from an experimental point of view. First "difficult" proteins may produce only a few or even a single crystal and second, employing more than one data reduction program during phasing efforts could be advantageous to phasing results.

Considering the difficulties involved with generating the AF1832 phases an obvious question presented itself. "Given a moderately diffracting crystal such as AF1382, does the choice of data reduction approach affect the S-SAD phasing results?" Herein we report a comparative analysis of data sets produced by five data reduction programs (HKL2000, d*TREK, XDS, MOSFLM, PROTEUM2) to the success rate of S-SAD phasing for data collected on a moderately diffracting crystal using 1.9Å SER-CAT (22ID) X-rays.

INDEX WORDS: Phasing, **S**ingle Wavelength **A**nomalous **D**iffraction, S-SAD, phase problem, data reduction approach, macromolecular crystallography

GIVEN A MODERATELY DIFFRACTING CRYSTAL, DOES THE CHOICE OF DATA REDUCTION APPROACH EFFECT THE S-SAD PHASING RESULT?

by

JAMES TUCKER SWINDELL II

B.A., North Carolina State University, 2001

M.S., North Carolina A&T State University, 2003

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

GIVEN A MODERATELY DIFFRACTING CRYSTAL, DOES THE CHOICE OF DATA APPROACH REDUCTION EFFECT THE S-SAD PHASING RESULT?

by

JAMES TUCKER SWINDELL II

Major Professor: Bi-Cheng Wang

Committee Members: John P. Rose
William M. Dennis

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

DEDICATION

For My Family

My efforts throughout my academic career are not purely the result of intellect or ardor, the foundation of my abilities are rooted in the members of my family. Form those in my history whose sacrifice can never be properly told, yet persevered through hard work, love, and faith I thank you. For my grandparents James Tucker and Larcenia Swindell, Thomas Nelson and Manetta Walker this is for you. For my parents James Wilson and Annie Walker Swindell your children are a reflection of the values and work ethic you instilled in each of us, we thank you. For my brothers and sister thank you for being supportive through everything.

To my wife who has been with me through most of my academic career I cannot say thank you enough, those words do not reflect your support and perseverance. A special thanks to my “proofreader” Elaine and James we will see you soon.

ACKNOWLEDGEMENTS

My time at the University of Georgia, involving both Physics and Biochemistry, has had the greatest impact on my scientific and personal development. It not only takes years to train a scientist but also talented individuals to mentor and advise along the way. Dr. Bi-Cheng Wang has been instrumental in my scientific growth. His foresight and constant vigilance served as a guide for my training and learning from Dr. Wang will be a memorable experience throughout my life.

I would also like to extend my thanks to Dr. John P. Rose who always made time for my questions never let a single query go unanswered. Thank you Dr. Rose. I am extremely grateful to the Department of Physics for allowing me to be the first student to peruse a concentration in Biochemistry to truly become a multi-disciplinary scientist.

To the members of the Wang lab I appreciate the comradery and support as project focus changed and results rose and fell to exterior scrutiny. I consider the effort and perseverance required to acquiring a PhD was best phrased, …if you want this degree you should wake every day feeling slightly empty because you do not have it, only this constant reminder will shape your mind for the sacrifice and exertion ahead of you.

What I have learned and those whom I have met made this experience something to be truly thankful for.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

To determine a protein's structure using X-ray diffraction is a tiered process. Several factors must be considered to accomplish successful data collection and processing. Let us begin by discussing the basic experiment involved. Little has changed in fundamental experimental procedure since the first structure of Myoglobin (1). The experimenter will in some cases purify then crystallize the target protein. The crystal is then mounted in the path of an X-ray beam and rotated by a predetermined small angle Φ, about a chosen experimental axis. A detection device is placed in the path and the resultant diffracted X-rays are recorded. With each predetermined crystal rotation a single corresponding image is captured. These images record the positions of the deflected X-rays versus the rotation angle. Diffraction patterns contain intensities characteristic with atomic scattering due to X-ray interaction with each component of the protein(s), which comprise the crystal. Additional scattering due to solvent, mounting devices and air does occur, but these factors contribute primarily to background noise.

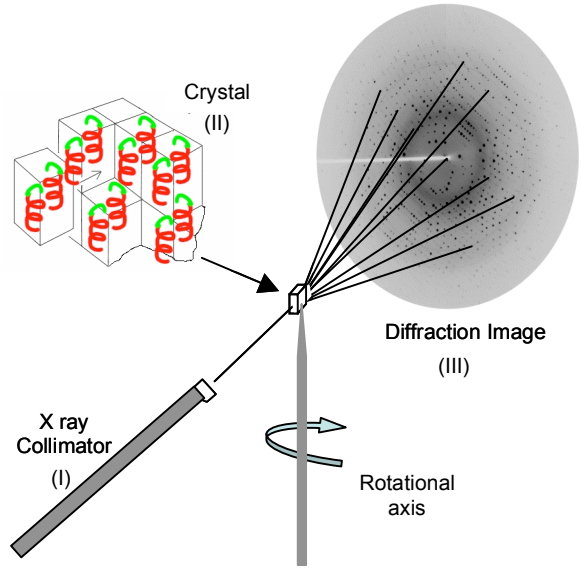

Figure 1.1 Diffraction Experiment: (I) X rays of a chosen wavelength are directed at the target crystal (II) composed of identical and repeating arrangements of the protein, (III) Diffraction image-visual representation of the scattering contributions from each atom within the protein.

The scientist then analyzes the diffraction patterns, which are a result of constructive and destructive interference depending on differences in the travel path of incident X-rays being equal to integer multiples of the wavelength. This is best understood by using Bragg's Law; $n\lambda = 2d(\sin\theta)$ from which n represents the integer "order" of each reflection, λ is the wavelength of the incident X-rays, d or "d-spacing" measure the interplanar spacing of the crystal and the incident angle of reflection is θ. The diffraction peaks, which result from a crystallographic experiment, characterize various planes through a crystal lattice. These peaks are assigned Miller (hkl) indices based on the different spacing and location of the diffraction peaks (also called reflections or spots). After several rounds of experimental refinements and corrections the individual reflection intensities are determined. Intensity values are calculated for each reflection by subtracting the background or noise inherent to the experiment. These measurements are used comparatively or in combination with information from related experiments to discern the phase information. Crystallized proteins diffract incident X-rays altering their direction producing

unique diffraction patterns, Figure 1.1. The structure of the protein is indicated from the diffraction pattern based on the relative spot angles, distances, and intensities generated from a collection of diffraction images. Precise knowledge of the diffraction intensities allow for the formulation structure factor, F_{hkl}.

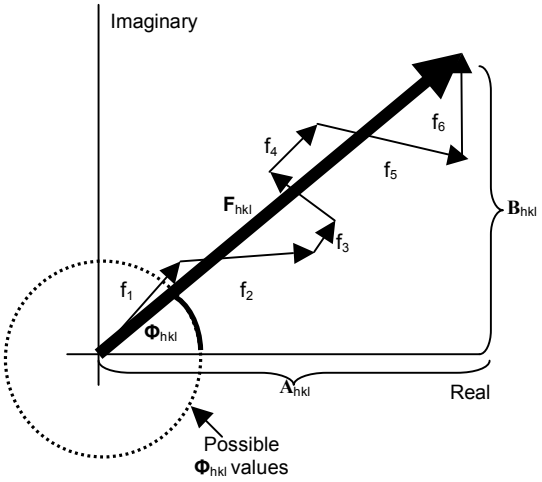

Figure 1.2: Structure Factor, F_{hkl}: F_{hkl} in phaser notation represents the sum of all the atomic scattering vectors in the unit cell. The magnitude of F_{hkl} can be calculated as $I_{hkl}=F_{hkl}^2$ but the phase information cannot be so simply devised.

F_{hkl} is the summation of all contributing scattering factors f_j containing their own phase and amplitude. The intensity of each diffraction peak (I) is approximately equal to square magnitude of its corresponding structure factor ($|F_{hkl}|^2$).

$$F_{hkl} = \Sigma_j f_j e^{-i\Phi_j} \quad \text{eq. 1.0}$$

$$= \Sigma_j f_j e^{[2\pi i(hx_j + ky_j + lz_j)]} \quad \text{eq. 1.1}$$

$$= \Sigma_j f_j \cos[(2\pi(hx_j + ky_j + lz_j) + i\Sigma_j f_j \sin[(2\pi(hx_j + ky_j + lz_j)] \quad \text{eq. 1.2}$$

$$= A_{hkl} + iB_{hkl} \quad \text{eq. 1.3}$$

Next, the phase angle of each reflection can be expressed in terms of A and B.

$$\Phi_{hkl} = \tan^{-1}\left(\frac{B_{hkl}}{A_{hkl}}\right) \quad \text{eq. 1.4}$$

Depending on the wavelength used, the magnitude and eventual phase of F_{hkl} will differ if the wavelength is approaches the excitation edge of atom(s) contained within the protein. As seen in equations 1.0-1.3, F_{hkl} includes the whole set of scattering factors f_j of all atoms (j) within the

unit cell. The term $2\pi(hx_j+ky_j+lz_j)$ describing the fractional coordinates of the j^{th} atom in x,y,z and the position of the hkl-reflections. A Fourier transform can be applied to F_{hkl} to provide a formula for an electron density function, ρ.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h,k,l} F_{hkl} e^{(-2\pi i(hx+ky+lz))} e^{i\phi_{hkl}} \qquad \text{eq. 1.5}$$

Figure 1.3 The Electron Density Equation: The Fourier transform of F_{hkl} used for calculating electron density.

This equation allows for the calculation of the electron density belonging to the various atoms, a graphical representation of the probability of the presence of an electron, at any point (x,y,z) within the unit cell by summing the atomic scattering factors F_{hkl}(2). The solution to the electron density equation allows for the calculation of electron density map, which represents the probable positions of the electrons of the amino acids within the protein. The electron density Fourier transform equation requires five parameters. First, the position of each spot within the diffraction pattern, which current data processing programs determined by use of a grid pattern assigned to the detectors surface (in x,y,z̸,ϕ), which can be described by a unique hkl index). Second, the intensity measurements (I_{hkl}), recorded as spots in the diffraction pattern. The intensity of these X-rays are represented by the following:

$$I(2\theta) = I_o\left[\frac{(ne^4)}{2r^2m^2c^4}\right]\left[\frac{1+\cos^2(2\theta)}{2}\right] \qquad \text{eq. 1.6}$$

Figure 1.4: Intensity Function: Where *n* is the number of electrons, *e* is the charge of an electron, *r* is the distance from scattered wave to detector, *m* is the mass of the electron, *c* is the speed of light, and the $[(1+\cos^2(2\theta))/2]$ term is the scattered photons partial polarization. –excerpt from Cryst. Notes and Man, Dept of Chem UWIC 1999

The third parameter needed is the volume of the unit cell (V), which is determined during the hkl assignment or indexing process. Fourth, the magnitude of the structure factor F_{hkl} (a mathematical description of how a material scatters incident radiation) and finally the phase information (Φ_{hkl}) associated with the structure factor F_{hkl}.

Of the components needed to satisfy the Fourier transform equation, all variables can be precisely determined from a single diffraction experiment save one, the phase (Φ_{hkl}). Herein lies the problem. Without the phase component, the Fourier transform equation cannot be solved. This issue is known as the phase problem.

1.1 Types of scattering

Typical X-ray diffraction experiments utilize X-rays of a precise wavelength. Interaction of the incident X-rays with electrons belonging to the atoms within the protein produce either coherent Thomson or incoherent Compton scattering. The scattered X-rays are recorded using a detector and are collectively referred to as a diffraction pattern. In 1913 a French mineralogist Georges Friedel proposed a theory which states the intensities of h,k,l and –h.-k.-l reflections are equal. Friedel was correct except in the case for resonance scattering. When the energy of the X-rays scatter from a quantum mechanical system, such as and atom, the energy is absorbed and almost simultaneously emitted. An excitation event occurs if the energy of incoming X-rays is tuned to the transition energy of the electrons within the atom. This excitation event involves the transition of an electron in the atom from its current energy state to a higher one. If the incoming X-rays posses the energy required for this transition they are said to be at resonance with electrons of the atom in question. As the excited electron ascends from a higher energy state back to its previous level a photon of nearly the same energy as the incident X-ray is emitted in nearly random directions. This excitation results in resonant or “anomalous” scattering in

addition to expected coherent scattering. The anomalous scattering signal is a by-product of X-ray absorption and re-emission, which disrupts the magnitude and phase of the normally coherently scattered X-rays. Johannes M. Bijvoet outlined this observation, which proved an exception to Friedel's rule, in 1949 during his study of cholesteryl iodide (3). The application of Bijvoet's findings would not be applied to crystallography until 1956 with the determination of sperm whale myoglobin (1). This was archived by exploiting this deviation from Friedel's rule, coining the term "anomalous" scattering, which is actually resonant scattering.

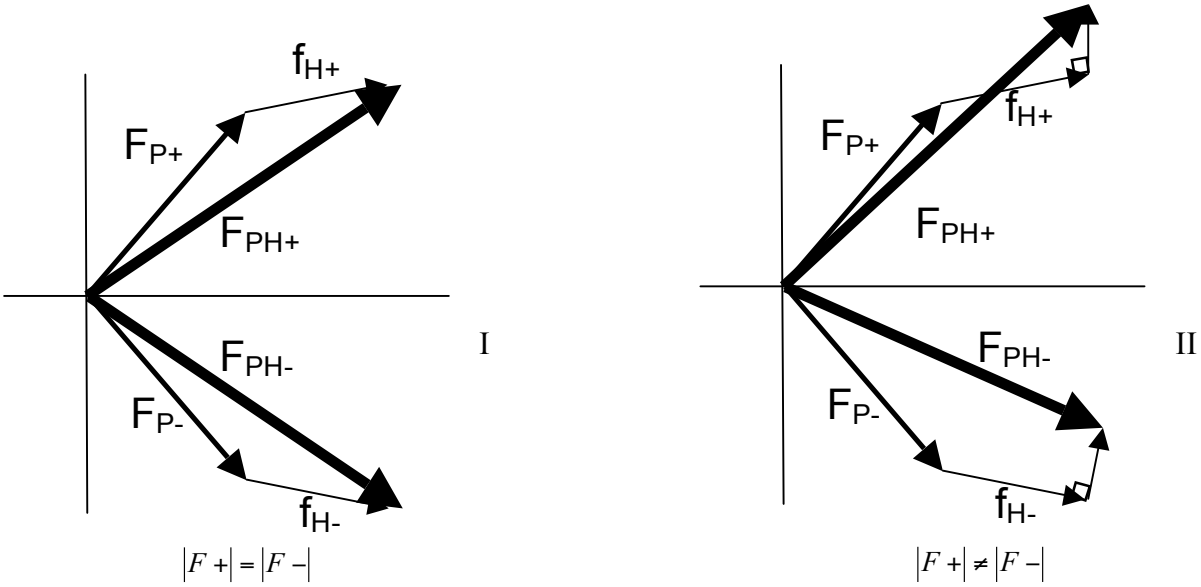

Figure 1.5: Effects of Resonant Scattering: (I) Satisfies Friedel's Rule for various structure factors, +F and -F while the inclusion of resonant scattering (II) adds an additional resonant component which is phase shifted 90 degrees, this causes an inequality amongst the magnitude and phases of ±F.

The total scattering contribution of any atom within the protein can be written as;

$$f = f_o + f' + if'' \qquad \text{eq. 1.7}$$

in which f is the total scattering factor, f_o is the coherent or Thomson scattering factor. This value is related to the energy of the incident and eventually scattered X-rays. The remaining $f' + if''$ terms represent the real and imaginary portions of any anomalous or resonance contribution.

The previously mentioned Compton scattering occurs simultaneously with Thomson scattering and is incoherent with the incident radiation. This type of scattering occurs when incident X-rays scatter inelastically, losing energy, and scattering at an angle which increases depending on the loss of energy. These lower energy X-rays have no specific directionality and are of low intensity. Their contribution is general background intensity or noise.

1.2 The Problem with the Phase Problem

Traditionally an Argand plane is used to illustrate the structure factor F_{hkl}, using both "real" and "complex" axis. The structure factor is illustrated as scalar of magnitude F_{hkl} and the associated angle/direction from the real axis is the phase angle (Φ_{hkl}).

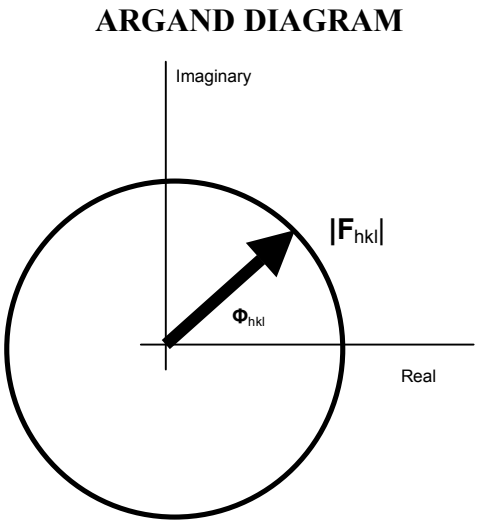

Figure 1.6: Argand Diagram A single data set, with no anomalous signal, is incapable of identifying what the proper phase (location) of the atoms contained in the protein. This loss of phase information is represented as a circle with no fixed direction for F_{hkl}.

F_{hkl} is proportional to the square root of the intensity (I_{hkl}) measured on the detector. With the magnitude of F_{hkl} being a known value, locating the correct phase from a seemingly

immeasurable number of possible phase solutions, which could accompany $|F_{hkl}|$ is the "Problem with the Phase Problem".

Without the critical phase angle information, no solution to the electron density equation can be discerned (eq 1.5). To overcome this problem various techniques can be implemented. The most commonly used methods to determine the proper phase angle corresponds to the use of differing F_{hkl} values. This is usually accomplished by (1) using naturally occurring anomalous scatters, (2) the introduction of heavy atoms (or anomalous scatterers) into the crystal lattice via soaking, or (4) the engineering of selenium during protein expression. Tuning the wavelength of the X-rays at or around the specific resonant excitation energies of the anomalous scatterers incorporated into the crystal/proteins will result in large $f' + if''$ values. Therefore, different F_{hkl} values can be generated using the correct heavy atom and its corresponding wavelength. The key to heavy atom incorporation is isomorphous distribution. Crystals grown in the absence of heavy atoms are referred to as native crystals. Those crystals exposed to heavy atoms (via engineering or soaking) are referred to as derivatives crystals. If the incorporation is truly isomorphous the differences between the native structure factor (F_P) and derivative crystal structure factor (F_{PH}) should only be the contribution of the heavy atom structure factor (F_H) (Figure 1.5). In addition, no disruption of the native protein's structure or changes in the crystal lattice should result from the isomorphous incorporation of heavy atom by soaking or engineering. If heavy atom soaking is the chosen method of crystal derivation the X-ray wavelength used for data collection should be as close to the heavy atom absorption edge as experimentally possible, in order to produce the largest anomalous signal. This phenomenon is referred to as resonant scattering, which will result in a detectible shift in the magnitude and phase of F_{hkl} when comparing native and heavy atom derivative crystals.

Methods used in determining the phase from X-ray diffraction data fall into three categories. The first and oldest method is Isomorphous Replacement. Two techniques, which fall explicitly within this category, are Multiple Isomorphous Replacement (MIR) and Single Isomorphous Replacement (SIR). The second category is Anomalous Dispersion, which contains both the **M**ultiple-wavelength **A**nomalous **D**ispersion (MAD) and **S**ingle-wavelength **A**nomalous **D**ispersion (SAD) methods. These methods can utilize heavy atom derivatives (having a large anomalous scattering component) but are more commonly used with either engineered or naturally occurring "heavy" atoms. The third category is Direct Methods. This method is generally used for small molecules or peptides containing ~1000 atoms and usually not applicable to large proteins. The fourth method, Molecular replacement is a commonly known method for generating protein structures, requires a model (structure) having a similar arrangement of main chain atoms. The sequence identity can be as low as 10% providing that the overall structure of the molecular replacement model is similar to the unknown structure. The phases of the model structure are used as a starting point for target phase determination.

The earliest method of phase determination employed Multiple Isomorphous Replacement (MIR) by Kendrew in 1956 with the structure of Myoglobin (1). This was the dominant means to solving the phase problem until the creation of synchrotron sources in the late 1980's. In 1983 Single Isomorphous Replacement (SIR) presented and additional method to address the phase problem with the use of Dr. B.C. Wang's Noise Filtering technique. This method used heavy atom derivatives to determine the correct phase, just as MIR but saw limited use. During the early 1980's SIR phasing was soon followed by Single-wavelength Anomalous Scattering (SAS), which is also referred to as Single-wavelength Anomalous Dispersion (SAD) phasing. Unlike its predecessors this method requires a single crystal and data set. Although this

method preceded MAD phasing it was its usefulness was not realized by the scientific community until the early 2000's. In 1981 a single structure was solved using the special case of Resonance Anomalous Diffraction via Sulfur by Hendrickson *et al*. (4), but this idea was not specifically visited again until 2006 (5).

During the time between the initial SIR solution and the emergence of routine SAD phasing in the early part of this decade, Multi Wavelength Anomalous Dispersion (MAD) became the dominant structural determination method, eventually surpassing MIR (6], 7). The largest factors which propelled the popularity of MAD phasing past MIR were crystal cryo-cooling (8), the creation of tunable synchrotron X-ray sources and proteins engineered using Selenium (i.e. Seleno-methionine) as an anomalous scatterer. Although SIR and SAD predated MAD experiments, issues such as cryo-cooling, intense tunable X-ray sources and area detectors necessary for consistent collection of data were nearly a decade away. As these advancements in technologies became available in conjunction with the commonplace engineering of proteins with Seleno-methionine, the use of native anomalous scatterers within proteins allowed SAD (9) to eventually surpass MAD phasing as the preferred methods of structure solution.

1.3 Phasing techniques:

The aforementioned procedures (MIR, SIR, MAD and SAD) uniformly employ phase triangles to identify the true or correct phase(s), (Φ_{hkl}). The three sides of the triangle represent the magnitudes of F_P, F_H, and F_{PH} scattering scalars from the native protein, heavy atom, and derivatized protein respectively, just as in Figure 1.5. The differing values of the structure factor scalars are a direct result of data collection at varying wavelengths (or the use of different heavy atoms). If the phase of any one scattering factor, such as Φ_P, Φ_H, or Φ_{PH}, were known, the orientation of the triangle will be geometrically fixed. Therefore, knowing the phase of any one

of the three sides F_P, F_H, and F_{PH} is to also know (via calculation) the phase of the other two sides through geometry. Of the scalars, which comprise the phase triangle, both F_P and F_{PH} are measured values while F_H can be calculated. In theory, if the location of heavy atoms or anomalous scatterer is known it can be used as a reference point for calculating the phase of F_P and F_{PH}.

$$F_{PH} = F_P + F_H \qquad \text{eq. 1.9}$$

$$|F_{PH}|^2 = |F_P|^2 + |F_H|^2 + 2|F_P||F_H|\cos(\Phi_P - \Phi_H) \qquad \text{eq. 2.0}$$

$$\Phi_P = \Phi_H + \cos^{-1}\left[\left(|F_{PH}|^2 - |F_P|^2 - |F_H|^2\right) \Big/ 2|F_P||F_H|\right] \qquad \text{eq. 2.1}$$

$$\Phi_P = \Phi_H \pm \beta \qquad \text{eq. 2.2}$$

Using only two (F_P and F_{PH}) structure factors will create a phase triangle possessing two possible orientations for the true or correct phase. Using a single heavy atom derivative will generate a bimodal phase ambiguity illustrated as β in eq 2.2.

Although the phase ambiguity can be expressed mathematically, the solution to this problem can be discerned through experiments. Each procedure discussed herein solves this ambiguity by a combination of mathematical and experimental means. When using MIR or MAD techniques for phase determination, multiple data sets are required which contain either different heavy atoms or wavelengths about a specific anomalous scatters absorption edge. The use of multiple data sets can resolve the best phase (either $\pm\beta$) by utilizing the heavy atom or anomalous contribution within a MIR or MAD experiment. Using SIR and SAD techniques, a mathematical approach can be enlisted termed iterative noise filtering to remove the necessity for multiple heavy atoms or data sets to solve the bimodal phase ambiguity (10).

It is important to understand what data are collected for each procedure and why. Each of the structural determination processes utilizes a least a single native data set. MIR and MAD

phasing techniques rely on three (or more) data sets, one devoid of anomalous or heavy atom contribution (native set) and the other two containing heavy atoms (MIR) or anomalous scatterers (MAD) soaked or engineered into the lattice of the protein crystals. SAD and SIR use either one or two data sets, for SIR a native and a heavy atom derivative data sets are used while for SAD a single data set processed keeping the anomalous scattering data separate set will suffice. In both cases an anomalous scatter must be present for proper phase identification.

The value of these techniques is amplified by understanding the infinite number of phase solutions, which are possible for a native protein data set. The scalar F_P, Figure 1.6, could be directed at any point at a radial distance from the origin. The use of phase triangles, generated for example from the anomalous data, will solve this phase ambiguity from an experimental standpoint. Each diffraction experiment produces an Argand diagram depicting the native or anomalous contributions for F_P (native) and F_{PH} (native + anomalous) scalars. Subsequently, using vector relationships F_H can be determined. The radii illustrating the possible phases corresponding to each of the varying structure factors will prove pivotal in resolving the phase ambiguity.

It is important to mention that perfect experiments are rare and there will be some error involved with locating the heavy atoms and properly measuring F_{PH}. This error is called LOC (Lack Of Closure), ε. This is defined as the difference between $F_{PH,observed}$ and $F_{PH,calculated}$. In truth, it is best to use the equation $F_{PH} \approx F_P + F_H$ in lieu of eq 1.9. Visually one can imagine LOC in terms of identifying the proper phase. For example the phase circles seen in Figure 1.7 will not experimentally intersect at a single point, the error exists within the data such there are Gaussian distributions describing the magnitude of LOC error. Combating elevated LOC errors requires additional data sets to be collected.

Before each phasing method can be discussed it the type of experiment used for each should be reviewed. The following table can be used as a reference for each section highlighting the experiments conducted for MIR, SIR, MAD, and SAD.

Method	Crystals Needed	Data sets Needed	Wavelengths Needed	Phasing probe
MIR	3 +	Native Derivative 1 Derivative 2	1	Atoms containing many electrons Hg, Pr, Au, Os
SIR	2	Native Derivative 1	1	Atoms containing many electrons Hg, Pr, Au, Os
MAD	1	Peak λ Inflection point λ Remote λ	3	Atoms with a measurable anomalous scattering signal at the peak λ and a tunable low band pass X-ray source
SAD	1	Native or Derivative	1	Atoms with a measurable anomalous scattering signal at the wavelength used to collect the data

Table 1.0 Experimental Phasing Requirements: Highlights the experimental requirements of each phasing method.

1.3.1 MIR

MIR phasing requires the collection of three data sets from at least three different crystals, one native and two heavy atom isomorphs. The two isomorphous data sets need not use different heavy atoms. Only the locations of the heavy atoms must differ within the protein. Successfully soaking identical protein crystals using the same heavy atom derivative but finding more than one orientation of the heavy atom within the protein is unlikely and at best a random occurrence. Thus it is more common for experimenters to use multiple heavy atom types when conducting such experiments. The experimenter would determine the structure factors for each data set ($F_{P(protein)}$, $F_{PH1(first\ derivative)}$ and $F_{PH2(second\ derivative)}$) as well as calculating the heavy atom contribution from F_{H1} and $F_{H2,}$ respectively. These scalars contain both the native (real) scattering component and anomalous (real and imaginary) contribution depending on the heavy atom chosen, as seen in an Argand diagram Figure 1.2. Using vector notation illustrates how the phase problem is solved using a common origin for each $F_{P,}$ $F_{H1,}$ and F_{H2} structure factors.

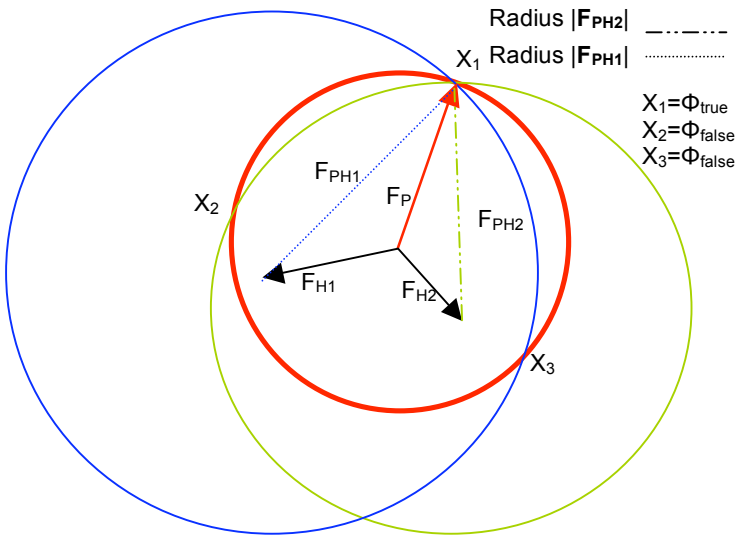

Figure 1.7: Phase Solution by MIR: The relationship between the vectors describing the heavy atoms ($F_{H1\ and\ 2}$) and native data (F_P) are illustrated the correct phase solution. Recall if a lack of closure error exist the circles would not intersect at precisely the same point.

Each intersection of the circles generated from F_P, F_{PH1}, and F_{PH2} scalars represent possible phase solutions. F_P defines the radius of the initial phasing circle. The terminal ends of the calculated F_{H1} and F_{PH1} scalars are used as origins for circles described in radius by the F_{PH1} and F_{PH2} scalars. Using only F_P and F_{PH1} an experimenter will be left with two possible solutions for the correct phase X_1 and X_3. Likewise, if F_P and F_{PH2} are used the two possible phases would be X_1 and X_2. Only by using all three data sets do the intersecting circles highlight the true phase solution X_1.

Assuming Equation 1.9 is correct the different lengths of the structure factors ensure that radial circles drawn from each F_H terminus will intersect at an approximately common point. This intersection marks the best phase needed to complete the electron density function, ρ (eq 1.5).

This method of phase determination was the earliest used and most taxing of the techniques reviewed. In a best case scenario the experimenter would collect multiple crystals grown in an

identical solution. These crystals would be screened to ensure a high quality resolution, and a portion of the total number harvested crystals were set aside for heavy atom soaking. Two types of heavy atoms are used in most cases. It is most difficult to incorporate heavy atoms into a crystal lattice without damaging the crystal/diffraction quality or significantly changing the dimensions of the lattice to ensure 100% isomorphism. If successful, the results of these efforts would be at least three crystals, a native crystal with no heavy atoms, and two additional crystals isomorphously accepting heavy atom derivatives. At this point data collection would begin.

1.3.2 MAD

MAD phasing can be conducted using a single crystal but this is dependent on the resiliency of the crystal. Both MIR and MAD phasing are similar in concept but differ in experimental complexity. The largest difference between the two procedures is the necessity of synchrotron X-ray sources because the data must be collected at three defined wavelengths the absorption maxima (peak) having the largest anomalous signal, the inflection point (the first derivative of the absorption curve) having the highest dispersive signal and a remote higher energy remote wavelength (having little or no anomalous signal). MAD experiments do not require the use of multiple heavy atoms derivatives. Only one anomalous scatterer is necessary which could be an anomalous scatterer already present in the protein, Seleno-methionine incorporation during expression is most commonplace method of introducing an of anomalous scatterer into the crystal (11). Selenium modification of methionine residues was ideal in creating reproducible and 100% isomorphous heavy atom derivative crystals. The correct phase can be calculated using three data sets collected at several different wavelengths from the same crystal. The three wavelengths used are directly related to the energy corresponding to the absorption edge of the anomalous scatterer used. The wavelengths at which the data are collected are the

absorption edge - f_2 corresponding to the largest f ” values; f_1 slightly off the peak considered the rising edge or inflection point with the largest f ’ value; and f_3 collected at a point at least 1000eV away from the peak or absorption edge to ensure the data would contain as little anomalous signal as possible.

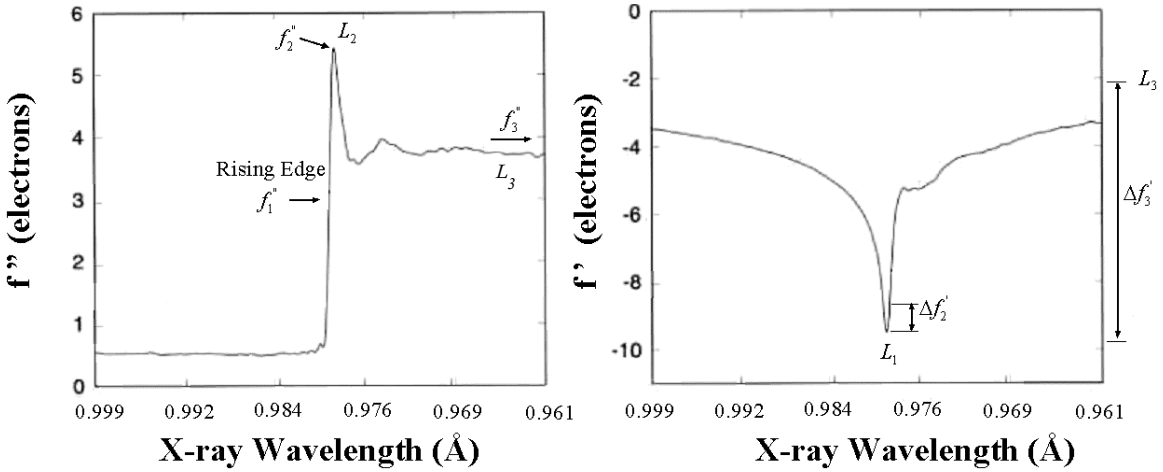

Figure 1.8: X-ray Florescence Scan, for Selenium: Three data sets are collected, firstly the data collected at the point producing the highest anomalous scattering value of *f* ” (L2), secondly the Rising Edge/Inflection Point which yields the largest f ’ value (L1), and finally a remote point data collection such that no significant anomalous signal is included (L3). - Adapted from: Ramakrishnan, V and Biou, V. Methods in Enzymology Vol. 276 New York, Academic Press 1997

Typically a Florescence scan (electrons) vs. Energy plot is used to identify the anomalous scatter and corresponding absorption edge within a protein (Figure 1.8), unless the experimenter has prior knowledge of the anomalous scatterer, to be used during data collection.

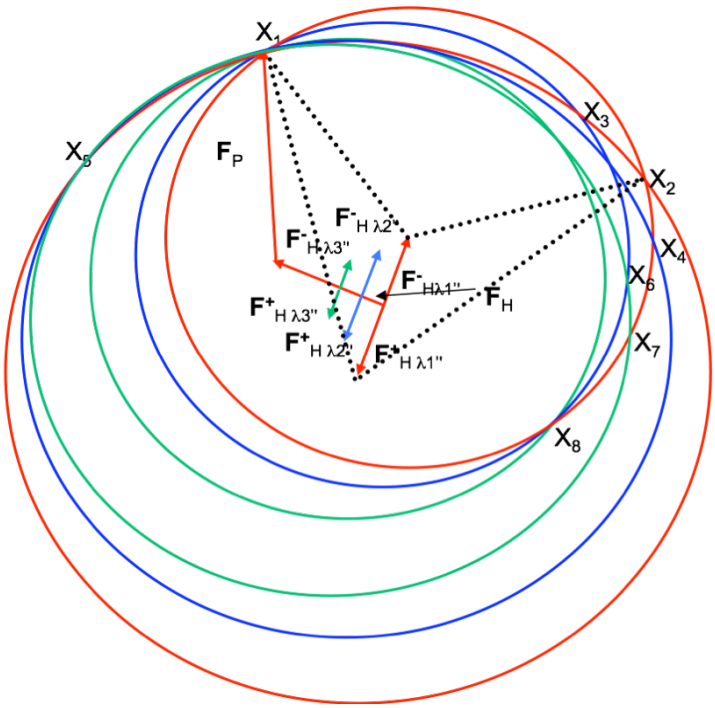

Figure 1.9: Phase solution by MAD: Within a MAD experiment the result of using three wavelengths can be represented by three separate phase diagrams. First the largest contribution of anomalous signal (collected at the absorption edge) is represented in red (L2). The blue circles correspond to data collected away from the absorption edge (L3) and the green circles represent the minimum anomalous signal collected (L1).

The use of three different wavelengths and the resulting anomalous contribution provides the same result (in terms of calculating the phase solution) as using three different heavy atoms in the MIR case. Calculating the structure factors from each anomalous experiment and combing them with the native can use vector diagrams can be used in both cases to clearly illustrate the correct phase solution.

1.3.3 SIR

The natural evolution of most experimental techniques is often accompanied by the progression of technology and accompanying theory. The data collection methods used in a typical MIR experiment are labor and time intensive. Reducing the number of data sets required needed to solve the phase problem from three to perhaps two or one would be a vast

improvement. Single Isomorphous Replacement and Single Wavelength Anomalous Dispersion (SAD), which will be described later, satisfy these requirements. SIR experiments require data collected on a native and one isomorphous derivative crystal. These data collection parameters can be described as 2/3 of a MIR experiment. SAD experiments further simplify requirements of SIR needing only a single crystal containing an heavy atom, naturally occurring or introduced.

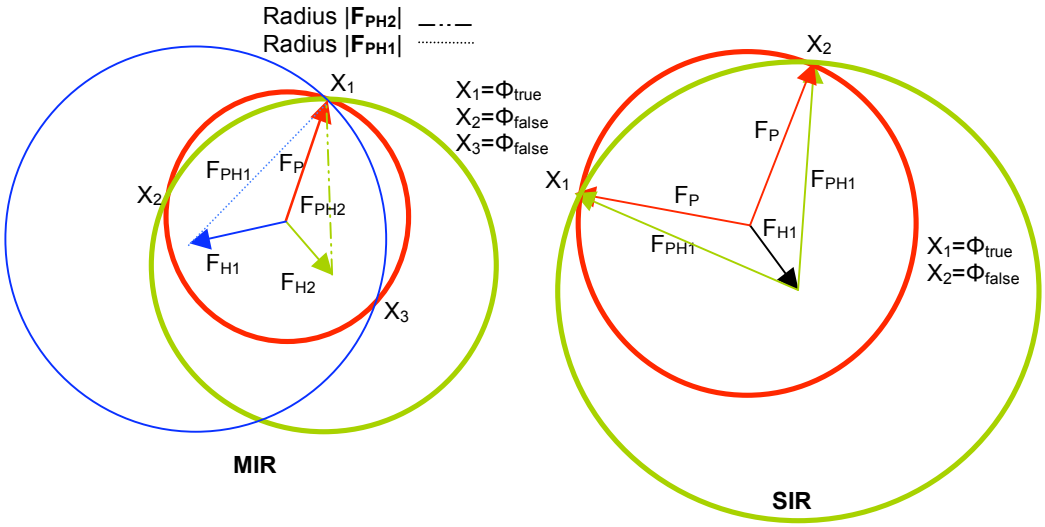

Figure 1.10: Phasing Diagrams Generated by MIR and SIR Methods: The SIR method represents the next generation of Isomorphous replacement phasing techniques. MIR experiments rely on at least three data sets to identify the appropriate phase while SIR experiments require two data sets and a computational means of solving the phase ambiguity.

Neither SIR nor SAD experiments alone could be effectively used to identify the true phase necessary for protein structure determination until the early 1980's. Improvements in discerning the correct phase originally required averaging both predicted phases and the values of the F_{hkl} vectors for each solution (12). This method was able to produced acceptable electron density maps up to 2Å resolution for myoglobin. Several equivalent techniques such as Double Phase and β-isomorphous synthesis (13-15) have also been used. The inclusion of the false phases with the correct phases creates electron density maps with contain considerable noise thereby

increasing the time and effort to discern the actual structure, if it is even possible to do so. As technology improves the current need for such heroic efforts involved with interpreting electron density maps has diminished. This is largely due to the 1983 release of B. C. Wang's method for phase determination commonly referred to as Solvent Flattening. The appropriate title is Iterative Single Isomorphous Replacement when working with SIR data and Iterative Single Wavelength Anomalous Scattering if using SAS data (Wang, 1985). This was the first approach to filter out the noisy background inherent to earlier techniques of correct phase identification. Both SIR and SAD techniques rely on a mathematical approach for solving the phase ambiguity. Several methods addressing phase ambiguity of SIR and SAD experiments had been proposed from the early 1950's to 1960's with varying degrees of difficulty (16). To resolve the phase ambiguity Dr. B. C. Wang in 1982 developed a multi-step computing process using "filters" and iteration by Fourier Transform between real and reciprocal spaces for the purposes of removing the false phase solution and enhancing the electron density maps. This process has been named by the crystallographic community as solvent flattening or density modification, which has become the backbone of most current software for phase improvement in SAD, MAD, SIR and MIR phasing methods. (Fig 1.11; Fig 1.12). Within this study a common program was used for structure determination that employs the Dr. Wang's technique to remove the phase ambiguity found in SAD experiments.

1.3.4 ISIR/ISAS

The foundation of the ISIR/ISAS method involves utilizing the electron density map to establish a molecular envelope from low-resolution diffraction data. An initial electron density map is generated by choosing the "best phase" as a starting point for inclusion in the electron density equation.

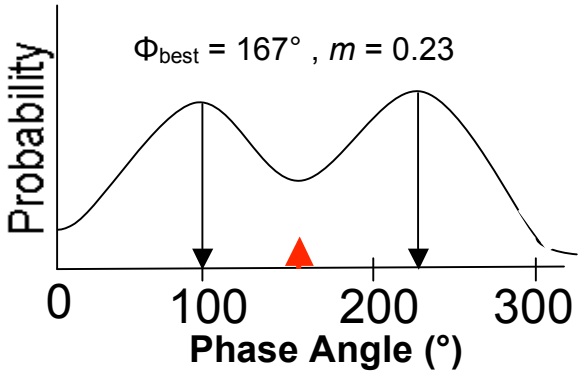

Figure 1.11: Bimodal Phase Distribution from SIR/SAD Phasing: The approximation of the "best phase" is considered the bisection (red triangle) the two possible phases identified. The *m* value represents the figure of merit; *m*=1 represents no phase error, *m*=0.5 reflects approximately 60° phase error, and *m*=0 describes all phases having equal probability.

Using this "best phase" will of course produce an electron density map, which appears to be more noise than secondary structure. From this initial map the molecular boundary is used to produce a mask, which identifies the protein and solvent regions. The density identified as solvent is assigned a value of zero and negative density discarded. The new masked density is held at a constant value while a positive constraint is applied to all the electron density encompassing the protein portion of the density. This process is called density filtering. The newly modified electron density map is reverse Fourier transformed producing a phase (Φ_{NEW}) value. This newly derived phase information is combined with the original "best" phase via a phase filter producing an improved "combined" phase treated as the new "best" value. Four iterative cycles of this process are conducted and the finial output combined with the initial "best phase", to prevent creating bias during phase calculation. Twenty rounds of this iterative cycle from real (electron density) to reciprocal (calculated and combined phases) are conducted. The finial phase calculated from this process is used to calculate the electron density map corresponding to the correct phase solution, this is illustrated in Figure 1.12.

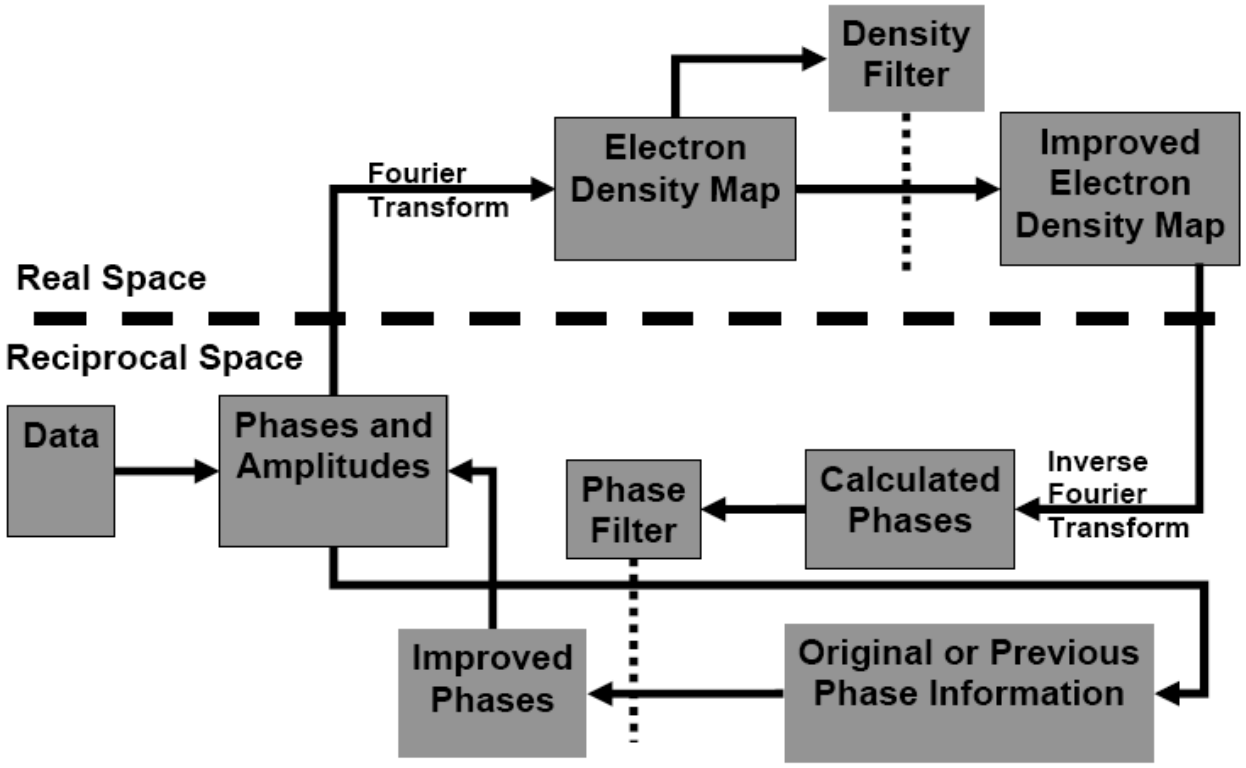

Figure 1.12: Simplified flowchart of the Wang method: In 1982 Dr. B. C. Wang devised a method in which solving the phase problem via anomalous scattering became independent of the percent contribution of the scatterers used. –adapted from Habel (2005).

This technique was the first to offer a reliable means to solving the phase ambiguity from any anomalous scatter used in SAD experiments.

1.3.5 SAD

Just as SIR is an abbreviated version of MIR experiments, SAD uses only one of the three experiments required for MAD experiments. Although SAD experiments did predate MAD these were few in number due to the necessity of having (I) a crystal that would not decay in the X-ray beam, (II) an accurate and sensitive detection device and full incorporation of the anomalous scatterer in the protein/crystal. The same scatterer used for a MAD experiment can be used in SAD data collection as well. Of course the necessity of identifying the scattering atom(s)

remains, such that data collection at the optimal wavelength, peak or L_2 (Figure 1.8), could be assured.

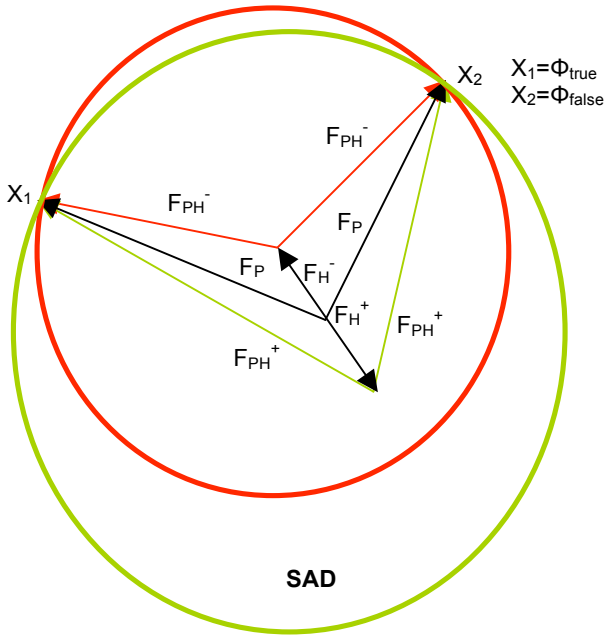

Figure 1.13: The Phase Ambiguity of SAD: Two possible phases arise from the intersection of both native (green) and anomalous (red) phase probabilities.

SAD has surpassed MAD phasing due impart to the community's final acceptance that SAD data alone could produce an interpretable electron density map. This realization coupled with crystal cryocooling, tunable X-ray sources and next generation detectors has made Se-SAD the dominant method for de novo structure determination as illustrated in Figure 1.14.

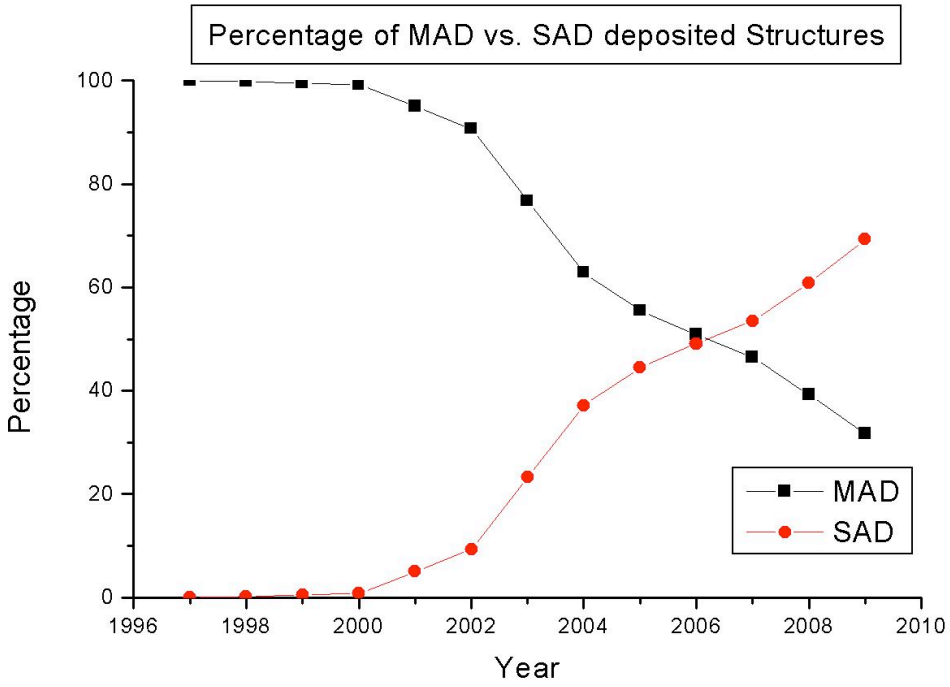

Figure 1.14: Graphical representation of MAD vs. SAD popularity: In 2006 SAD phasing surpassed MAD as the preferred method of determining protein structures. This graph was compiled using the submitted information contained within the PDB.

SAD experiments are best conducted at a synchrotron source where the X-ray can be tuned to maximize the anomalous scattering signal. Importantly, unlike MAD, SAD experiments can also be carried out in the home lab using either Copper or Chromium X-rays, provided that the anomalous scatterer being exploited for phasing has a measurable anomalous scattering signal.

Element	Absorption Edge Wavelength (Å)	Δf"(e^-) @ λ=1.0Å	Δf"(e^-) @ λ=1.54Å	Δf"(e^-) @ λ=2.29Å
S	5.02	0.24	0.54	1.12
Se	0.97	3.69	1.14	2.52
Fe	1.74	1.12	3.33	0.75
Mn	1.89	2.7	1.4	0.66
Ca	3.07	1.2	1.2	2.95
Zn	1.28	0.7	0.74	1.44
Hg	1.01	10	7.7	14.2
Pt	1.07	9.1	6.9	12.9
Au	1.04	9.1	6.9	13.46

Figure 1.15: Various Anomalous Scattering Values; Selected wavelengths including the typical wavelength used at synchrotron radiation sources (1.0Å), Copper (1.54Å), and Chromium (2.29Å) measured in electrons (e^-).

The advantage of using engineered Seleno-methionine labeled protein lies in the fact that the absorption or transition edge for Selenium is ~0.980Å within the experimental envelope of most synchrotron beamlines. However some Seleno-methionine labeled proteins do not readily crystallize and not all proteins contain methionine. In these cases, experimenters must consider the atoms contained within the amino acid sequence, which comprise the protein. Of these, the only atom which could serve as an anomalous scatter, is Sulfur.

1.4 The Special case of Sulfur-SAD

The concept of using Sulfur, a naturally present anomalous scatters to phase a crystal structure was initially realized with the phasing of Crambin by Hendrickson and Teeter in 1981 (17). This was accomplished by utilizing the anomalous contribution of 6 Sulfur atoms inherent to the protein. Although the ratio of Sulfur atoms to amino acids for Crambin were quite high and largely uncharacteristic for macromolecules (6 Sulfur atoms per 45 amino acids) the concept of S-SAD was affirmed. Hendrickson also formulated an equation to determine the percent of scattering contribution each atom contained in a protein that adds to the net intensity of diffraction measurements, depending on the wavelength used during data collection. Research

conducted by Dr. B.C. Wang in the early 1980's showed that using his ISAS method a ratio of 57 residues per Sulfur could be used for phasing if the data were accurate enough (10). From Dr. Wang's experiment the Bijvoet ratio exemplified by the Magdoff equation (4) yielded, $<\Delta F>/F \approx 0.6\%$. To this day the Wang "limit" still holds as an excellent measure of the probability of achieving a structural solution. The AF1382 protein targeted in this study possesses a Bijvoet ratio of 1.05%, which is well within the Wang "limit" for structure determination. Dr.Wang showed by simulation that initial phases calculated by anomalous scatter substructures could be improved upon independently of the level of contribution of the anomalous scatters, thus allowing for the extension of SAD experiments to larger proteins. Within five years of this prediction the first SAD structures were phased via Seleno-methionine and a selenobiotinyl derivative (18, 19). Wang's 1985 results were further validated by work conducted by Dauter (20), proving if an anomalous signal could be measured accurately then ISAS is an efficient means of utilizing the substructure Sulfur scatterers within the protein despite a relatively low anomalous continent. This discovery heralded the possibility of using the weak scattering potential of naturally occurring Sulfur atoms as a means to phase protein structures. Although hypothesized by Wang in 1985, the first *de novo* S-SAD structure using ISAS was not realized until 15 years later when the 22KDa protein Obelin was solved (21). Although still not as common place as Selenium SAD, Sulfur-SAD phasing is steadily gaining a foothold in structural biology. The absorption edge of Sulfur, 5.015Å or 2475eV, remains far beyond practical exploitation at synchrotron sources due to extreme loss of beam intensity and air absorption during data collection. There are, however, several accessible wavelengths that have been utilized in solving S-SAD structures both at both home and synchrotron sources.

Popular wavelengths for Sulfur SAD Studies

Genertors	Wavelength	Δf" (e-)	Author/Year
Cu RA	1.5418	0.56	Dauter/1999
Synchrotron	1.74	0.67	Lui/2000
Synchrotron	1.9	0.83	Weiss/2004
Cr RA	2.2909	1.14	Habel/2005

Figure 1.16: Several of the Wavelengths used for S-SAD Experiments: The fact that home source generators can be used to solve the S-SAD phasing remove the absolute need for synchrotron sources.

Sulfur is an excellent candidate for an anomalous scatter due to its availability throughout many genomes. To illustrate the utility of Sulfur as an anomalous scatterer, several reprehensive genomes were chosen to discern possible targets of S-SAD studies.

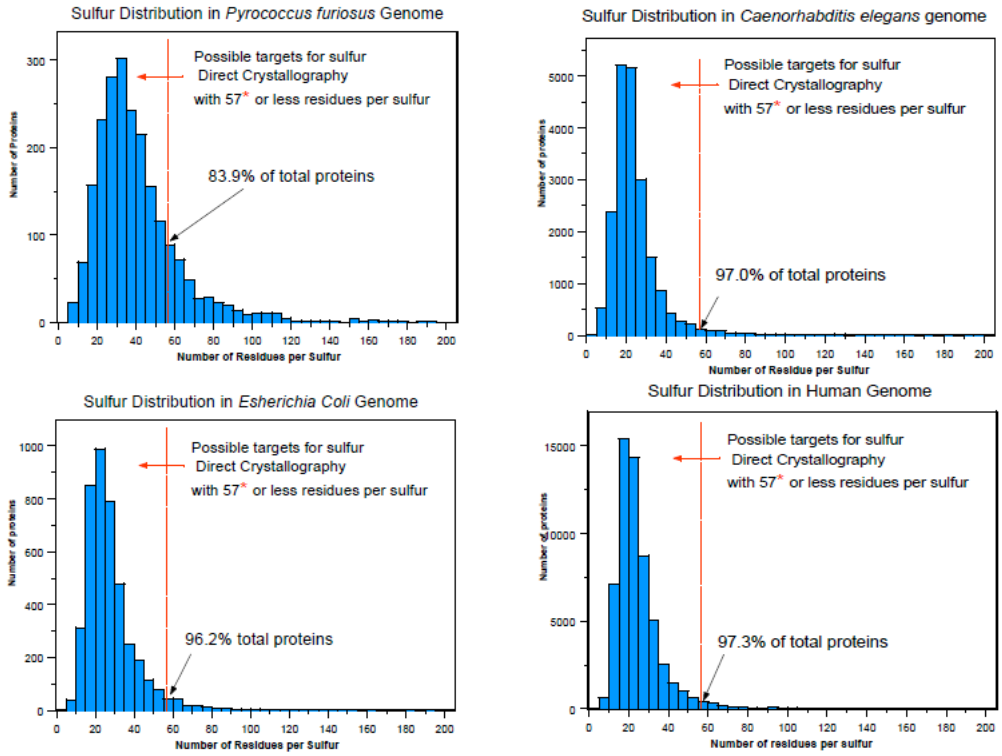

Figure 1.17: Graphical analysis of the Sulfur atoms within the "Wang limit": This diagram illustrates ~83 to 97% of proteins comprising several genomes provide excellent targets Sulfur SAD phasing. -adapted from New Frontiers in Neutron Macromolecular Cryst. Wang 2005

This information, coupled with the inherent roadblocks of traditional SAD phasing, offers a clear reason for increased attention to S-SAD structure determination. The weak scattering potential of Sulfur results in an anomalous signal can be used for successful phasing, but the data must be recorded very accurately. Standard data collection methods used for SAD can be extended to the special case of S-SAD. However, due to the sensitivity of S-SAD to radiation decay, there are several studies which illustrate the use of multiple data sets collected at low power being processed in combination providing a better phasing solution than a single higher power data collection (21, 22).

Additional factors should be considered when attempting to optimize the signal to noise ratio such as sacrificing high-resolution data via moving the detector away from the crystal. The resolution range required for successful location of the anomalous substructure is typically within the lower resolution range, 3-4Å. This must be done carefully to avoid loss of Sulfur scattering signal due to air-absorption. Collecting smaller oscillation steps will, in theory, decrease the noise generated from background scattering (23, 24). Decreasing the background contribution will increase the signal to noise ratio of the data, thus highlighting the anomalous contribution of Sulfur. Another method of increasing the overall Sulfur contribution is achieved by collected redundant data.

During a Sulfur-SAD experiment, it is common to collect data at the longest stable wavelength available. Since the resonant wavelength of Sulfur, 5.02Å, is prohibitively long for collecting data at either a home or synchrotron sources the contribution of Sulfur anomalous scattering will be diminished. However several wavelengths below the absorption edge of Sulfur have been used for successful S-SAD studies (Figure 1.16). The biggest disadvantage associated with the use of long X-ray wavelengths is radiation induced crystal decay. Prolonged exposure to

low energy (long wavelength) radiation has been found to damage Sulfur containing residues preferentially due to Sulfur's significantly higher absorption cross section compared to other atoms commonly found in a proteins such as carbon, oxygen or nitrogen. This is most easily seen in the destruction of disulfide bonds in which two Sulfur atoms coordinated by a covalently bound cystine residues will produce a favorably large anomalous signal commonly referred to as "Super Sulfurs" at low resolution (25, 26).

This manuscript considers the role of data processing when attempting to successfully phase data from moderately diffracting crystals containing weak anomalous scatterers such as Sulfur by addressing the question, " For moderately diffracting crystal does the choice of data reduction approach effect the resulting S-SAD phasing?"

To answer this I have analyzed the top five programs currently used to process X-ray diffraction data. From the inception of X-ray diffraction being used to solve protein structures, technological emphasis has been placed on the hardware applications associated with improving X-ray generation, stability, cryogenic techniques, accurate detection devices, improved optics, wavelength accessibility and data collection methods (27). Various software packages have been devised over the years for data reduction and phase determination. Some of these packages have produced great success and others have not. However, very little emphasis has been placed on the best route for processing S-SAD data. Typically labs are relegated to using a single data reduction program due to the general success and or knowledge base within the group concerning its operation. This can be disadvantageous to the laboratory due to different programs interpreting identical data in a different manner. Special care should be taken not to disregard a data set because of a single program's failure. The effort and expense necessary to prepare and crystallize a protein sample, then properly mount, and collect data on it should not be written off

because a single data reduction program was unsuccessful in generating a viable solution. As previously mentioned, the weak anomalous signal from Sulfur in conjunction with the necessary consideration and accuracy involving Sulfur-SAD phasing is an excellent candidate to determine the effect of data reduction choice versus phasing prowess. There exists very little tolerance for error when considering Sulfur-SAD data before the signal is lost in translation. Therefore, with the hardware technology gap bridged for routine S-SAD, crystallographers must recall a time-tested concept: how data is interpreted is just as important as how the data was collected.

1.5 AF1382 Data Processing at SERCAT, 22ID line:

The test data set used in this work was collected with the original intent of describing the crystal structure of a non-Pfam protein from the organism *Archaeoglobus fulgidus*, titled AF1382. This data set was chosen for the following reasons: (1) the viable resolution of the data extended to only 2.65Å – providing for a moderate resolution test set; (2) this level of resolution will test the previously established upper limits currently used for S-SAD phasing; and (3) the most commonly used data reduction program, HKL2000(28), required 720° of data and expert assistance to phase the structure. Thus, this data offers a excellent test set for determining the effect of various data reduction programs not only because of the weak anomalous signal associated with Sulfur-SAD but also the difficulties associated with processing will truly test the methodologies of each data reduction program studied in this work.

Initially the AF1382 protein was recombinantly expressed using Seleno-methionine with the intent of conducting a Se-Met SAD phasing experiment. The crystals were shipped to the 22-ID beamline at SER-CAT located at the Advanced Photon Source (APS), Argonne National Laboratory, and data collected using 0.979Å X-rays. This resulted in a poorly phased solution that did not produce a structure (29). Using Selenium as an anomalous scatterer is often

attempted before Sulfur phasing because the scattering potential of Selenium far exceeds that of Sulfur at wavelengths commonly used at both synchrotron and home sources (Figure 1.15).

The difficulties accompanying the production of Seleno-methionine protein with respect to time and effort have been discussed previously. Incorporation of Seleno-methionine is also likely to affect steps involved in protein crystallization such as; decreased in protein solubility, variation in known crystallization conditions, and non-isomorphous incorporation leading to decreased resolution. The possibilities of mixed oxidation states of Seleno-methionine can also cause non-isomorphic distribution of Selenium which will attribute to a reduction of anomalous signal (30).

During a second attempt at structure solution, the protein was again expressed recombinantly but without Selenium incorporation. The AF1382 protein was purified via nickel affinity and gel filtration chromatography and crystallized. The crystals were again shipped to the 22-ID line at SERCAT for the purpose of S-SAD data collection. When the initial data set failed to produce a structure, the crystal was remounted and a second 360º data collected which eventually was used to discern anomalous scattering sub-structure and the phases. Both data sets were collected using 1.9Å X-rays, the maximum stable wavelength currently achievable on the SER-CAT 22ID beam line. Hereinafter, these data sets will be referred to as the as the R1 and R2 data sets. Both R1 and R2 data sets were collected from the same crystal using exposure times of 3 and 2 seconds respectively. It is important to note the R2 data set was collected after the crystal was removed and remounted, this will be discussed later. AF1382 consisted of 95 amino acids with 4 Sulfur atoms (3Methionines and 1 Cystine), well within the historical ratio (10) for structures solved using S-SAD. However, the resolution range of 2.65Å is at the approximate upper edge of known S-SAD phasing solutions. Knowledgeable observers accustom to working

with diffraction images would conclude that the diffraction images from this crystal seem (at least qualitatively) to be of poor quality. This is conceivably due to high background and highly mosaic diffraction pattern. Dr. Zheng-Qing Fu, an extremely talented staff member at SERCAT and author of the structure determination program SGXPRO, initially processed and merged the two data sets using the HKL2000 gui interface. The HKL2000 processed data either the R1 or R2 data sets alone or merged R1-R2 data set did not yield a solution. The R1, R2 and merged R1-R2 electron density maps showed inconsistently traced alpha carbon chains of varying length and no apparent secondary structure. After several days of arduously re-processing both the singular and merged data sets, Dr. Fu was able to determine the appropriate HKL2000 integration parameters and data merging strategy that finally yielded the AF1382 structure. Key to this success is the SCALEPACK scaling script shown in Figure 1.18.

Scaling script developed for HKL2000/SCALEPACK scaling

```
print user interface
scalepack log file '/Users/bcllab1/Desktop/with-lorentz-ugadefSite-R1_R2/NoREINDEX-
Project2a/tucker1.log'
resolution 50.00 2.30
number of zones 10
estimated error
0.05 0.05 0.05 0.05 0.05
0.05 0.05 0.05 0.05 0.05
error scale factor 1.3
default scale 10
rejection probability 0.0001
reference film 7
scale restrain 0.01
Absorption zo3
Lorentz
space group P42
output file '/Users/bcllab1/Desktop/
with-lorentz-ugadefSite-R1_R2/
REINDEX-Project2a/tucker1.sca'
Anomalous
ignore overloads
intensity bins
add partials 1 to 360 361 to 720
fit crystal cell 1 to 360 361 to 720
fit crystal mosaicity 1 to 5 6 to 10 11 to 15
16 to 20 21 to 25 26 to 30
31 to 35 36 to 40 41 to 45
46 to 50 51 to 55 56 to 60
61 to 65 66 to 70 71 to 75
76 to 80 81 to 85 86 to 90
91 to 95 96 to 100 101 to 105
```

```
106 to 110 111 to 115 116 to 120
121 to 125 126 to 130 131 to 135
136 to 140 141 to 145 146 to 150
151 to 155 156 to 160 161 to 165
166 to 170 171 to 175 176 to 180
181 to 185 186 to 190 191 to 195
196 to 200 201 to 205 206 to 210
211 to 215 216 to 220 221 to 225
226 to 230 231 to 235 236 to 240
241 to 245 246 to 250 251 to 255
256 to 260 261 to 265 266 to 270
271 to 275 276 to 280 281 to 285
286 to 290 291 to 295 296 to 300
301 to 305 306 to 310 311 to 315
316 to 320 321 to 325 326 to 330
331 to 335 336 to 340 341 to 345
346 to 350 351 to 355 356 to 360
361 to 365 366 to 370 371 to 375
376 to 380 381 to 385 386 to 390
391 to 395 396 to 400 401 to 405
406 to 410 411 to 415 416 to 420
421 to 425 426 to 430 431 to 435
436 to 440 441 to 445 446 to 450
451 to 455 456 to 460 461 to 465
466 to 470 471 to 475 476 to 480
481 to 485 486 to 490 491 to 495
496 to 500 501 to 505 506 to 510
511 to 515 516 to 520 521 to 525
526 to 530 531 to 535 536 to 540
541 to 545 546 to 550 551 to 555
556 to 560 561 to 565 566 to 570
571 to 575 576 to 580 581 to 585
586 to 590 591 to 595 596 to 600
601 to 605 606 to 610 611 to 615
616 to 620 621 to 625 626 to 630
631 to 635 636 to 640 641 to 645
646 to 650 651 to 655 656 to 660
661 to 665 666 to 670 671 to 675
676 to 680 681 to 685 686 to 690
691 to 695 696 to 700 701 to 705
706 to 710 711 to 715 716 to 720
fit batch rotx 1 to 360 361 to 720
fit batch roty 1 to 360 361 to 720
postrefine 10
write anomalous rejection file
format denzo_ip
sector 1 to 360
FILE 1 '/Users/bcllab1/Desktop/ok/compare/Project_3a/R1_try2-SameAsAlbert/040707-
8_2_1_1_0###.x'

HKL MATRIX 0 1 0
           1 0 0
           0 0 -1
sector 1 to 360
FILE 361 '/Users/bcllab1/Desktop/ok/compare/Project_2a/R2-try3-SameasAlbert/040707-
8_2_2_1_0###.x'
;
```

Figure 1.18: HKL2000 Scaling script: Script-1 developed by Dr. Zing-Quin Fu creates a reflection rejection file and reindexes the h,k,l indices to ensure the unit cell axis between data sets R1 and R2 are equivalent.

HKL2000 was not able to generate the phases for AF1382 using either single data set. Instead, the phases were generated by relying on increasing anomalous signal from the additional redundancy contained in 720º of diffraction data instead of a standard 360º data set (29). An increase in redundancy provides multiple measurements of the same diffraction peaks which can be averaged along with there associated error thus increasing the signal to noise ratio and the anomalous signal within the data.

An interesting fact surfaced after processing the data concerning are the average intensities divided by the average standard deviations for the entire data set, abbreviated $\langle I/\sigma_I \rangle$. The $\langle I/\sigma_I \rangle$ value of the merged R1-R2 data sets was 12.36% higher than the R1 data set alone. However, the R1-R2 merged data set was 11.9% less than the R2 data set alone. Yet the R_{sym} and R_{merge} values for the merged data did not differ greatly from the individual R1 and R2 processing runs. Classically R_{sym} is difference between multiple symmetry related reflections throughout the data while R_{merge} represents differences symmetry related differences when combining datasets, these terms are however often used interchangeably.

Set	R_{sym}(%)	$\langle I/\sigma_I \rangle$
HKL-R1	4.2	72.3(3.34)
HKL-R2	4.1	93.7(17.58)
HKL-R1/R2	5.6	82.5(10.7)

Table 1.1: Initial Data Processing results conducted by Dr. Zheng-Qing Fu using HKL2000: Dr Fu used hand written scaling scripts. R_{sym} is a data quality indicator measuring the agreement of symmetry related observations of a individual (hkl) reflections. $\langle I/\sigma_I \rangle$ is the average signal to noise ratio for the entire data set, and highest resolution shell. – adopted from Jinyi Zhu, 2007

Annealing and/or more accurate crystal alignment are the likely reasons for the R2 data set having a higher $\langle I/\sigma_I \rangle$ than the R1 data set. Annealing is the process by which cryo-cooled crystal is warmed (not necessarily to room temperature) in an effort to allow repacking of the protein lattice and provide better lattice packing (31, 32). The act of removing this crystal

possibly resulted in an annealing event. The second data set was collected after the crystal was remounted and subsequently removed from the goniometer and replaced in its storage container at the conclusion of the R1 data set collection. Annealing is usually considered an all-or-nothing option. It can be useful if the mosaicity or diffraction is of such poor quality that processing data from such a crystal would prove difficult or impossible. By warming the crystal and allowing for protein repacking both the mosaicity and resolution can be improved upon (33). The risk involved with annealing is irreversible loss of diffraction. However, in most cases the possible advantages to annealing outweigh the risk. Annealing is the only remaining option for improving the diffraction quality of a crystal which has already been mounted and subjected to cryogenic temperatures (31). As for crystal alignment, the more care taken during the experimental setup will have a direct relationship to the quality of the data collected.

With a ratio of Sulfur atoms to amino acid of ~1:23, using S-SAD to solve the structure of AF1382 should be an easy choice for structural studies. Despite this theoretical fact, crystals are seldom ideal. In practice the viable resolution of a crystal and anomalous scattering potential are issues, which are related to the quality of the data collected. The viable resolution range of the R1 and R2 data are currently at the approximate upper threshold of S-SAD phased structures. The level of effort and expertise required by Dr. Fu in processing the data sets dictates that this data presents a substantially challenging task for processing. Although the initial S-SAD phases produced the structure, the final model (PDB entry 2QVO) (32) was built and refined against a 1.85Å data set collected later on a better diffracting crystal.

1.6 Significance of this Work:

The effects of this research are far reaching and will bolster the field of SAD phasing and especially S-SAD structure determination. In recent years a large portion of NIH funding

involved with X-ray crystallography has shifted from large scale structural genomics towards more concise fields of study such as protein-protein complexes, membrane bound and human proteins. These targets may prove to be less resilient in respect to derivatization or expression as Seleno-methionine proteins for structure determination**.** This research highlights the necessity of considering multiple options involving data processing, especially when considering Sulfur-SAD or other SAD experiments using weak anomalous scatterers.

An increasing number of non-formally trained scientists have shown interest in crystallography as a means to better understanding protein mechanisms. This is an excellent development for the science of crystallography. The prevailing issues that hinder non-traditional crystallographers is the lack of crystallographic knowledge including data collection and processing knowledge. Approaching any task as a novice can be quite daunting (ask any first year graduate student), especially for a topic such as crystallographic data reduction without having a fundamental understanding of the statistical output from each portion of data reduction process. These limits restrict a novice experimenter from accurately processing data sets that present problems. Without knowledge of what the statistical values represent, no intelligent changes can be made to the processing procedure and thus the slightest challenge will shackle a novice.

This work offers a "real world" test, from a novice perspective, of each of the five major software packages (HKL2000(28), PROTEUM2(34), d*TREK(35), XDS(36), MOSFLM(37)), which will be especially beneficial to two groups: (1) the non-traditional crystallographers who are interested in target proteins which contain weak anomalous scatterers (e.g.-Sulfur-SAD) but lack the experience to adequately manipulate data reduction programs when dealing with such data. Second, beamlines located at various synchrotron sources that are in constant competition

to provide the highest quality of service to their users. Thus, identification of which processing program is most likely to produce straightforward and high quality results will save time and money during the structure solution process.

In addition, although multiple data processing programs are generally available at most beamlines most novice users consult the opinion of the beamline staff member on call who chooses data reduction programs based on his/her past experiences and proficiencies. Furthermore, nearly all data reduction programs in use today would not qualify as "novice friendly", unless of course, the data being analyzed requires nothing more than straight forward or "black box" processing. This black box approach is reinforced by the fact that many of the programs studied in this work have an automatic or nearly autonomous function, works well for high quality data but produces less than optimal results for challenging problems.

This work exhibits the need to provide multiple data reduction options for synchrotron users as well as individuals experienced enough to offer assistance with a variety of platforms when needed. A comprehensive walk-through of the most popular data reduction programs is presented with the goal of determining the best route for processing moderate resolution S-SAD data, which has application to any SAD data set regardless of the anomalous scatterer in question. Working with well studied proteins such as glucose isomerase, bovine insulin, or lysozyme would not provide results as convincing as those acquired from a novel unstudied protein such as AF1382. This protein offers a real life example of what users face during *de novo* structure solution using mediocre diffraction data exhibiting increased background scattering. Finally, a consistent method for both quantitatively and qualitatively testing the results of each data reduction program based on the phased solution is presented

As the science and funding of X ray crystallography expands towards more complex, sensitive, and costly targets, attention must be given to the methods used to process data. Yet, little effort has been applied to identifying which data reduction program(s) use the best approach in dealing with less than ideal data. The amount of care and effort used in protein preparation, crystal growth, mounting and data collection is irrelevant if attention is not paid to data processing. Sub par data can be “rescued”, as illustrated by Dr. Fu’s HKL2000 contribution to the AF1382 structure, with proper data processing. Conversely processing highly accurate data incorrectly significantly lowers the chances of producing a structure. This fact is of the utmost importance for those structural biologists who wish to use crystallography but lack the knowledge to fully exploit optimization options within data reduction programs.

Chapter 2

Data Reduction Overview

The data reduction programs utilized in this work are HKL2000, PROTEUM2, D*trek, XDS, MOSFLM. Each program performs essentially the same functions: (1) indexing, (2) refinement, (3) integration, and (4) with the exception of MOSFLM, scaling.

Indexing uses the positions of the reflections in the diffraction pattern to determine~~s~~ the orientation of the crystal axes with respect to the X-ray beam and to assign Miller indices to each reflection. Indexing also produces the unit cell parameters a, b, c, α, β, γ and crystal orientation in reference to the detectors surface, the orientation matrix. Refinement, often considered part of indexing, is a means of adjusting experimental parameters (e.g. crystal to detector distance) such that the predicted spot locations better match the experimental positions. Indexing is often not given the measure of care it should. When dealing with weak anomalous scatterers it is important to accurately predict centroid positions to capture the complete intensity profile of each ~~spot~~ reflection.

The most critical part of the data reduction process is integration in which the intensity of each predicted reflection in the diffraction pattern is calculated. The orientation matrix determined during indexing generates the predicted positions of the diffraction spots on the face of the detector. In addition to calculating the intensity values, machine and counting error terms proportional to the square of the intensity are generated as well (22, 36).

The final step of the of data reduction process is scaling. During scaling the information generated during integration from each diffraction image is collected, and if necessary, a scaling factor is applied to adjust and discrepancies among identical or symmetry related measurements

placing the intensities on the same numerical scale. Reflections are merged and averaged (user preference) producing a single list of unique intensity measurements and related errors.

During the course of this study the reflection intensities produced by the aforementioned ~~various~~ programs were scaled using their respective scaling algorithms with the exception of MOSFLM. MOSFLM's integration results are commonly scaled by a stand alone program titled SCALA (38, 39). An additional scaling program titled 3DSCALE, a portion of the SGXPRO structure determination suite (40), was also used to scale integration results from these programs where possible. The 3DSCALE approach is unique when compared to the scaling programs included in this work and will be discussed later.

To independently evaluate the various data reduction programs used in this study, a real world S-SAD test case was chosen. The data sets used for this work represents the current cutting edge for Sulfur phasing due to the viable diffraction resolution of 2.65Å, a lack of "super Sulfurs" or disulfide bonds and the difficulties involved with previous processing attempts.

Before I begin to review the individual data reduction programs, first I offer a concise overview of indexing, refinement, integration, and scaling. A general explanation of topics will also be offered pertaining the specific inner workings of each program in later sections.

2.1 Indexing

The initial step in data processing X-ray crystallography data is indexing. The goal of indexing is the creation of an orientation matrix. This calculation lies at the center of the entire data collection process. Once this has been determined, the correct position of any reflection in terms of in X, Y (on the detector face) and phi (rotation angle) can be calculated for any reflection. The initial orientation matrix and unit cell dimensions are refined to produce the best

fit of all available reflections with the unit cell and predicted spot locations. This is the manner into which spot predictions are created for use during integration.

To initiate indexing diffraction, images are loaded into each data reduction program along with the four necessary values for processing crystallographic data; (1) the crystal to detector distance, (2) the X-ray wavelength, (3) the direct beam center, and (4) the detector swing angle 2θ. In most cases, this information is contained in the header of each diffraction image and is automatically loaded into the program. Otherwise, the user must do so. Identification of suitable diffraction peaks allows for proper indices assignments (*hkl* integer values related to the reciprocal lattice). Spot identification is the first and arguably the most important step in indexing. One or more images are selected and reflections above a certain sigma value are harvested and used for indexing. From a collection of harvested spots, indexing algorithms attempts to fit the locations of the spots to potential unit cell and laue groups based on the spots angular and coordinate location. For each possible laue group identified, attempts are made to compute possible unit cell dimensions consistent with the identified spots. This fitting does include grossly inaccurate fitting profiles, which are easily discernable due to high errors seen in distortion indexes or large %-fit values, displayed by each program, associated with incorrect laue group selection.

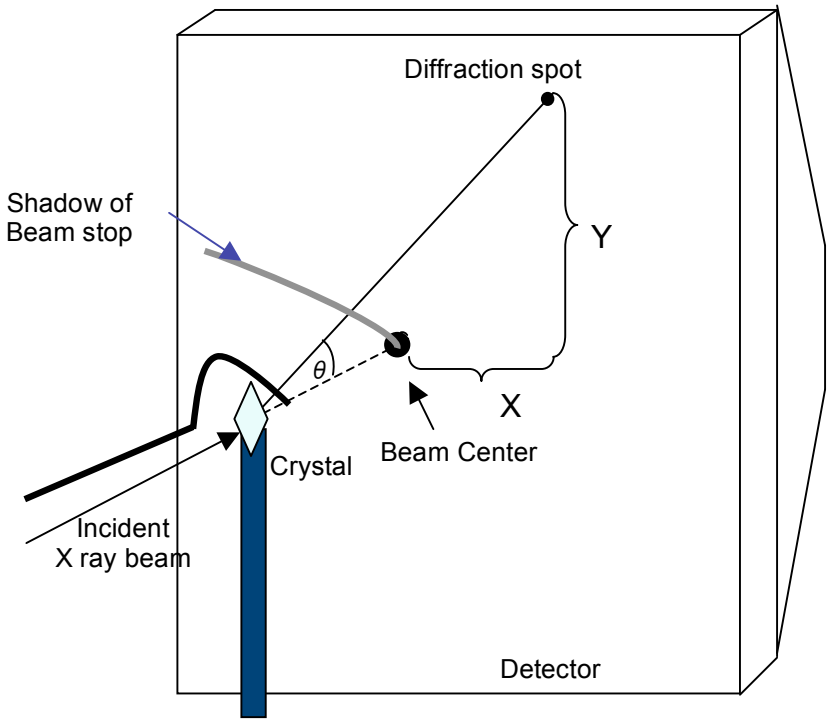

Figure 2.1 Scattering Representation: The "diffract spot" has a X,Y and φ (not shown) coordinate system which used to extrapolated the hkl Miller indices as a measure from the origin (0,0,0). The distances from the origin and between these spots assist in calculating an orientation matrix.

I will briefly discuss the two types of indexing common to the programs studied in this work. These methods are Difference Vectors (originally Diffraction Vectors) eventually named Auto-Indexing and Fast Fourier Transform (FFT). Despite the method of indexing used, the technique of identifying diffraction spots is similar regardless of the program used. The detector is typically divided into a predetermined set of regions, and background estimates from each region are obtained from the mean, median, and mode of the counts within each region. Consequently, each program locates peak positions by identifying pixel counts substantially higher than the surrounding background are considered spots.

The older of these methods termed Diffraction Vectors indexing evolved from Difference Vectors indexing originally developed for small molecule crystallography. The goals of

Difference Vectors indexing is to assign an elementary unit cell and orientation matrix that generates integral indices values associated with individual diffraction reflections. The first step is to choose the position **X** (x,y,z) describing a reciprocal-lattice point paralleled with vector **h** representing the Miller indices(h, k, l) by a reciprocal space unit cell matrix [A], with a,b,c being the unit cell dimensions;

$$A = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix} \qquad \text{eq 2.0}$$

Such that:

$$X = [\phi][A]\mathbf{h} \qquad \text{eq 2.1}$$

This method uses a trial and error approximation of **X** based on three or four sets of indices trials. Choosing a single sample of three reflections; $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$ and $[\phi]$ defining the rotational matrix about the detectors spindle, matrix [A] can be calculated which relates the unit cell to the actual crystal position within the beam. If the crystal was properly oriented such that the detector spindle axis corresponds to the direction of an axis of the unit cell the $[\phi]$ matrix will be unitary. This is easily accomplished by inverting Equation 2.1 yielding:

$$h = [\phi]^{-1}\,[A]^{-1}\,X \qquad \text{eq 2.2}$$

If the values of the chosen indices (h, k, l), which comprise $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ are sufficiently close to integers the A^{-1} matrix is considered commensurate the process of generating the indices of all reflections using a primitive cell is continued. The unit cell created by this method, being primitive, may not disclose the complete symmetry of the lattice. A least squares refinement method based on observed reflections is used to refine the symmetry for the correct lattice identifier. However, if $[A]^{-1}$ is not acceptable a different set of reflections such as $\mathbf{h}_2$ or $\mathbf{h}_3$ are

chosen until an acceptable set of indices are located (41). This process was incredibly time consuming, as multiple **h** vectors would require consideration within the diffraction pattern. In addition, this method was exceptionally subject to erroneous spots generated from cracked crystals, crystal twinning, or doubled matrices which is characteristic of X-rays passing through two crystal forms with different orientations.

The aforementioned procedure serves as a foundation on which all improvements involving indexing techniques are based. The next method utilized for indexing diffraction images is termed Difference Vector indexing or Auto-indexing. This again includes utilizing a random sampling of three non-coplanar vectors **X** (eq. 2.1), and assigning arbitrary **h** indices from the diffraction pattern (42). An orientation matrix [A] (eq. 2.0) and information corresponding to a unit cell can be calculated from these chosen images. Although it is highly unlikely that these values will be an accurate representation of the dimensions correct unit cell (a,b,c.), they must be a sub-cell (a',b',c') of the true unit cell.

Auto-indexing moves a step beyond Difference Vectors indexing by calculating multiple **T** vectors, a dimensionless unit vector of a chosen direction, from lists of lattice vectors used to devise various sub cells such that;

$$\mathbf{T} = u\mathrm{a}' + v\mathrm{b}' + w\mathrm{c}' \qquad \text{eq. 2.3}$$

with *u*, *v*, and *w* representing real space integer diffraction spot contributions. All **T** vectors are tested against the chosen non-coplanar **X** vectors to determine if their product would yield integer values. Granted, perfect integer values are not likely to occur, however values within an acceptable range of error may be satisfied. With the appropriate **X** vectors identified, the proper unit cell and orientation matrix can be calculated. The improvements of this technique over Diffraction Vectors indexing save enormous time and effort required to manually interpret

sample vectors. Instead of generating an orientation matrix for each vectors chosen this method allows for the elimination of errant vectors using the $\boldsymbol{T} \bullet \boldsymbol{X}$ test. Yet this procedure is still subject to the same pitfalls of erroneous spots generation, which limit the Diffraction Vectors indexing technique.

An additional improvement of the Diffraction Vector indexing method was christened Auto-Indexing which offers a means of overcoming the shortcomings of the previous indexing procedures, namely choosing individual vectors for a trial and error form of unit cell determination. The creation of Difference Vectors labeled $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ take the place of the aforementioned **X** vectors. The **X** vectors are again generated in the same fashion as in the Diffraction Vectors analysis using available spots from a single diffraction image. The Difference Vectors are actually the difference between two chosen **X** vectors ($\mathbf{U} = \mathbf{X}_i - \mathbf{X}_j$) (43). Errors associated with Diffraction Vector indexing are reduced in this indexing procedure. From a substantial list of **X** vectors the difference vectors **U** are calculated, and many of these differences will repeatedly occur. Those **X** vectors, which correspond to the frequently replicate differences, are used as basis vectors which define the reciprocal lattice and thus the orientation matrix. This averaging will decrease the errors pertaining to intergerness of indices and lowering the overall error (44).

Computer analysis significantly decreased the amount of time needed to generate the large number of **U** vectors to detect a frequency pattern. However, there was no way to check every possible orientation for vectors which could be generated from the indices within a diffraction images unless the user defines each possible vector orientation, which of course is protein dependent. The results from Auto-indexing were markedly better than Diffraction Vector

indexing, but the use of randomly chosen vectors kept the chances of consistently generating the best orientation matrix and unit cell parameters low.

Difference Vector or Auto-Indexing was constructed from the original techniques used in formulating Diffraction Vector indexing. The next step forward involves the use of a Fast Fourier Transform (FFT) algorithm, which combines a scan of the entire reciprocal space by a 1-D Fourier analysis. Using such a method was originally proposed as early as 1986 (45) but was not realizes until the mid 1990's (46, 47), as part of the DENZO data reduction program, due to technological gaps involving computer memory and processing speed (48).

The fundamental methods ~~seen~~ illustrated in equations 2.0-2.3 (chapter 2) remain in use within the FFT algorithm. The major differences begin with the classification of $\boldsymbol{T} \bullet \boldsymbol{X}$ values, considered ρ, generated during Auto-Indexing. To use the FFT algorithm all possible projections of reciprocal lattice point (**X**) onto the direction of **T** are sampled. The direction of **T** is defined using polar coordinates ranging from $0 < \varphi \leq \pi/2$, $0 < \psi \leq 2\pi$.

Due to this hemispherical range, a base separation of 0.03 radians was used to produce approximately 7,300 equally space possible **T** orientations. For each direction **T**, the values corresponding to the largest 30 identified *k* indices maxima are located, either the h or l index could have been used as well. Using this grouping as a sub-set, the largest maxima for *l* (or k,h) indices are discerned as well. Directions chosen form the remaining vectors yield a linearly independent set of three basis vectors from which a primitive unit cell can be calculated. The three sets of vectors, which produce the best indexing results, will produce an eventual orientation matrix [A].

Unlike previous procedures, which required arbitrary selection of vectors from three reciprocal lattice points, all possible directions and frequencies can be tested using the FFT

indexing technique. This method represents a major advancement in the X-ray structure determination process and significantly reduces the time required to properly index data. The programs studied in this work each use a variation of this procedure, save one. HKL2000 and PROTEUM2 follow this method while a variation of this method titled DPS (Data Processing Suite)(49) indexing is utilized in MOSFLM and d*TREK. Only XDS does not use FFT indexing but instead used the older Difference Vector indexing technique.

This increased indexing accuracy is absolutely necessary when attempting to detect the anomalous signal. This is especially pertinent when dealing with the weak anomalous signal from characteristic of a Sulfur-SAD experiment. Proper differentiation between peak and background, as well as accurately selecting the centroid of each spot, are pivotal for including all portions of the reflections in the peak profile. Taking special care during the indexing process to determine these aspects will directly effect data integration and scaling quality. In the case of weak anomalous scattering signal poor indexing will often result in low or non-existent anomalous differences. Proper spot prediction is absolutely essential to accurately locate the centroid of each predicted spot. Having accomplished this, all symmetry related reflections are predicted and compared with the locations of the initially harvested reflections. In a perfect experiment the observed and predicted reflection locations should be accurately predicted with a certainty of 0.01% or better (50). At the conclusion of initial indexing, the program can further refine the parameters pertinent to spot location, crystal orientation, and detector settings to increase the fit between the observed and predicted spot locations. The refinement techniques involve parameter correlations which are often left to the program to sort out, but in some cases the user is given latitude in restraining portions of the refinement process.

2.2 Refinement:

The refinement process is usually conducted simultaneously with indexing. This portion of processing minimizes the differences between calculated and observed values of the spot positions in X, Y and ω, or by minimizing the "non-intergerness" of the observed reflection indices h, k and l identified during indexing. At the conclusion of indexing, assuming all has gone to plan, the data reduction program should have properly identified the appropriate Bravais lattice type and generated a orientation matrix corresponding to the observed diffraction data. The predicted reflection positions should overlap with observed diffraction spots throughout all oscillation images contained within the data set. There are essentially 3 classes of parameters considered during refinement: (I) crystal related (unit cell, orientation, mosaic spread); (II) detector related (distance and orientation); (III) beam related (position, wavelength, beam divergence). These values are often highly correlated to one another such that refinement of one parameter may affect others, such that the refinement diverges instead of converging. Refinement is commonly conducted after Bravais lattice selection but this is program dependent. Refining initial spot selection using just the searched/harvested spots in Bravais group P1 (triclinic) before actually choosing the lattice type would logically increase the likelihood of properly identifying the proper lattice type and cell dimensions since detector parameters have now been refined. An additional round of indexing after lattice choice can also be quite beneficial to accurately locating the initial approximation of the reflection's centroid, as well as refining any other parameters such as beam center, detector distance and the orientation matrix which would eventually produce a better refined signal-to-noise calculation as well.

2.3 Integration:

After completion of indexing and refinement the crystal/detector parameters and orientation should be well known and refined. The next step in data reduction process is integration. Integration of X-ray diffraction data relies on the results from indexing for accurate predicted peak positions needed to accurately measure the full intensity of each reflection.

The two most popular methods of obtaining the integrated intensity of a reflection are Summation Integration and Profile fitting (51, 52). Of the two methods Summation Integration is the older. Summation Integration considers the intensity of the pixel values within a defined X-Y grid (spot integration box) centered on the predicted spot position and pixel values lying outside the spot integration box to determine the reflection's background. The area of intensity to be integrated calculated is determined by summing the pixels within the spot integration box and subtracting the average background. An intermediate zone is created that encompasses intensity values, which are above pre-determined, background levels but below the values associated with spot intensities. These pixels are simply discarded. The cutoff values for pixel intensity can be defined to eliminate the abyssal zone thereby making the pixels above preset background peak. This method is quite useful for strong reflections but weak reflections found in higher resolution shells may not be accurately measured. This is primarily due to the lack of a substantial difference between the coherent scattering from the atomic components of the protein and incoherent and background scattering from various items such as solvent, air, mounting apparatus, or cryoprotectant.

The second method of integration is the most commonly used and referred to as Profile Fitting. This method divides the detector face into smaller segments from which the habit of

various diffraction spots therein are “learned”. Averaging the profiles of strong spots from a particular segment will produce a standard profile for fitting all diffraction spots in that region.

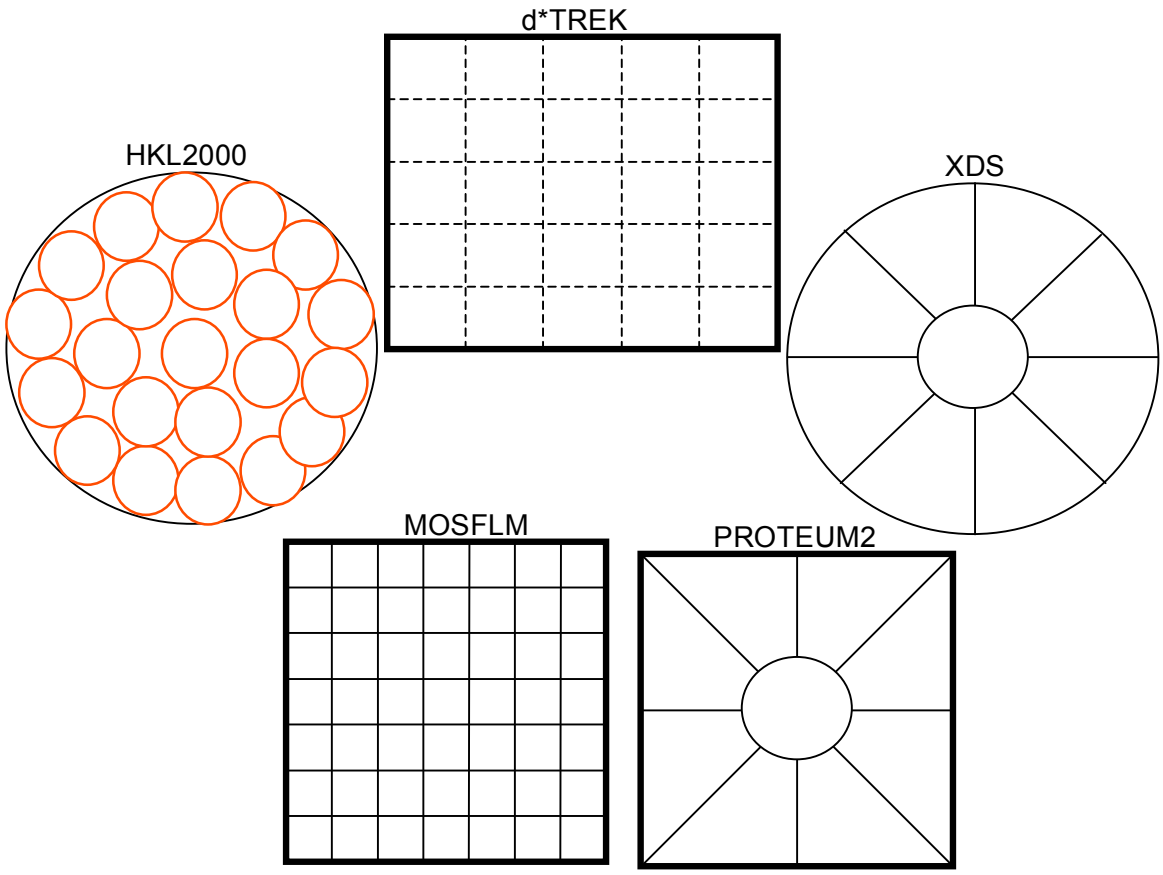

Figure 2.2 Detector divisions used for Profile Fitting: Each program covered in this study divides the detector face differently and averages the strong spots within those regions to create standard profile. MOSFLM, and d*TREK use square sectioning while XDS and PROTEUM2 use triangular patters with a circular origin for detector division. HKL2000 uses a flexible Profile fitting Radius, displayed as a single circle, which can be moved throughout the detector face highlighting reflections to be used within its area for standard profiling.

Consider a detector face covered in pixels, the intensity and location of each pixel is assigned an identifier S_i for peak and B_i as background. The pixels defined as B_i can be considered S_{i+1} in any direction from S_i depending on measure of the pixel values. This will define spots throughout the detector space. For strong reflections analyzing the difference $S_i - B_i$ will highlight the outline or shape of the intensity measurements. Averaging the profiles of the strong spots will produce standard profile, P_i which is normalized to 1; $\sum_i P_i = 1$. The variances between each spot profile used in the creation of standard profile, V_i, are generated to create a best estimate for intensity beneath each profile:

$$I = \sum \frac{P_i(S_i - B_i)}{V_i} \Big/ \sum \frac{P_i^2}{V_i} \qquad \text{eq 2.2}$$

The standard profiles are usually learned using strong reflections spots due to the greater differences between S_i and B_i when defining pixel and background. However, once created these profiles will be used throughout the detector segments for both strong and weak spots alike. Thus, defining the area in which to integrate the diffraction intensities, accounting for any contribution $S_i - B_i > 0$, will out perform summation fitting at high resolution due to the blanket cutoff intensity values necessary for spot identification (51). The process of generating reference profiles and matching observed profiles is repeated through out the data set for each image. As data processing proceeds, continual minimization routines are conducted in reference to the standard curves to account for small shifts in peak positioning to ensure the full intensity of each spot is considered.

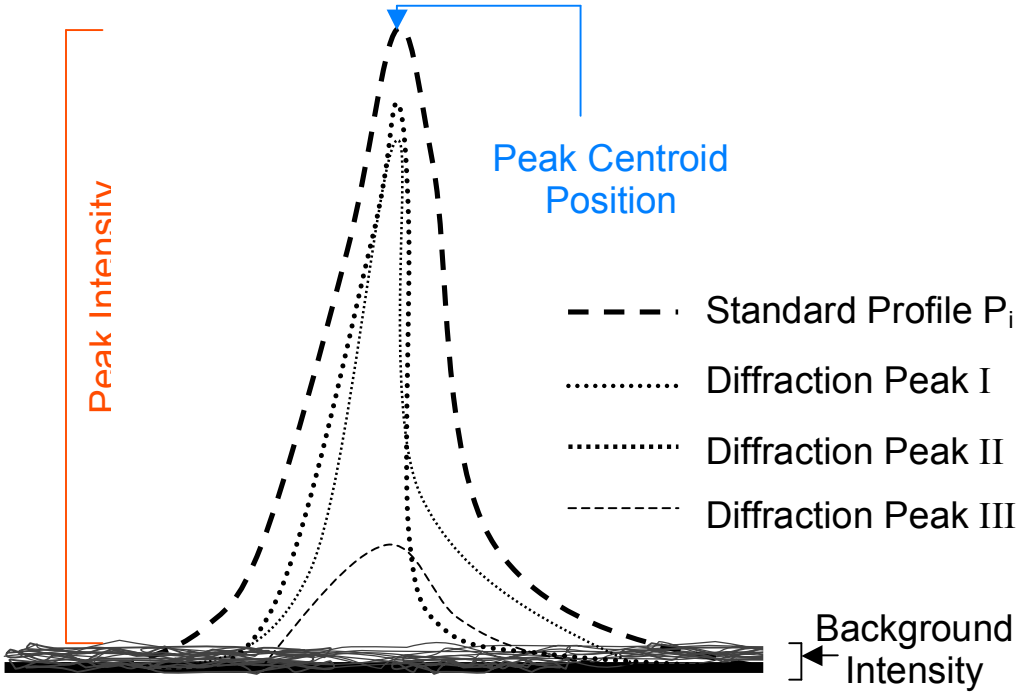

Figure 2.3 Profile fitting: The Standard Profile is generated from a collection of strong spots within an image or collection of images. The spot intensity is not commonly affected by the spot width, although weaker reflections seem narrower than strong ones it is due to the shoulders of the reflections being convoluted by background noise. Although Diffraction Peaks **I** and **II** may have been used to construct the Standard Profile, Diffraction Peak **III** will be similar in width to prevent any offsets in centroid identification for proper integration.

Therefore knowing the precise position of these observed peaks is of the utmost importance, or the intensity calculated from the observed and reference profile fit will be inaccurate (35). Failure to accurately define the centroids of the predicted peaks will result in the formation of a reference profile with smeared or inaccurate shapes. With the knowledge of the strong reflection indices and position, weaker reflections are located based solely on predictions from the information obtained from strong reflections. The average profile created during Profile Fitting indicates the area, which is interrogated for intensity measurements above estimated background measurements for the respective segments of the detector.

Geometrical anamorphosis will cause the spot profiles to vary across the detector surface such that standard profiles for a particular spot constitute an average of surrounding peak profiles by superimposing the observed reflection profiles with high I/σ values. Using this method peaks can be analyzed as partials or full reflections (28, 35). The primary issues with this method of integration are the need of high I/σ reflections and clear differentiation of reflections and background, which is traditionally a problem for unit cells with long cell axis, which will cause reflections to occasionally overlap. This will cause problems in the "learning process" associated with generating standard profiles. Fortunately, the experimenter can adjust the detector distance to combat these problems.

2.4 Scaling

At the conclusion of integration the final step of data reduction commences, entitled scaling. Two major goals are accomplished during scaling. First, scaling corrects for crystal decay and absorption. Intensities from the initial images of the data set are used as control as these images should not suffer from any decay due to radiation. Using these values the intensities throughout the data set scaled. If data collection were perfect there would be no need to scale the calculated intensities from the various images. Scale factors must be applied to each image to ensure any discrepancies are as low as possible between the various intensity measurements for identical or symmetry related reflections.

Second, scaling corrects for errors associated with the experimental setup (2). Consolidating the integrated intensities from symmetry related reflection requires reformatting the integrated intensities such that related observations are merged. For S-SAD it is important that Friedel mates, indices h,k,l and –h,-k,-l, are not merged to retain the effects of anomalous

scattering. The indicator commonly used to describe the intensity disparities, which require scaling, is the R-factor;

$$R = \frac{\sum_j \left| \langle I \rangle - I_j \right|}{\sum_j I_j} \qquad \text{eq. 2.3}$$

<I> is the average intensity and I_j represents the intensity after the application of the scale factor. Acceptable measures for R-value are <5% for excellent data, 6-10% is considered usable, 10-20% presents questionable results and +20% being completely errant and an indicator of serious problems.

Experimental error correction is often considered part of scaling but it primarily corrects erroneous intensity measurements. The factors which account for error in data collection such as absorption, crystal defects, radiation damage, X-ray beam instability, detector defects and other systematic errors are quite difficult to accurately define (40). The creation of systematic error models based on redundant data is the classic approach to detecting aberrant intensity measures within the experiment. Typically, highly redundant data are used to determine a model for removing experimental errors. This modeling approaches attempts to simulate the experimental error by using pre-defined functions. For high quality data, this technique is sufficient. During integration, statistical analysis of the intensity measurements generate error which is defined in terms of standard deviations, σ. Data reduction programs which use these values create error models such that the eventual "goodness of fit" or χ^2 values are calculated as close to identity as possible (53):

$$\chi^2 = \frac{\left(\langle I \rangle - I_j \right)^2}{\sigma^2} \qquad \text{eq. 2.4}$$

From the equation if the $(<I>-I_j)^2$ value is equal to σ^2,a resultant value of one for χ^2 would indicate nearly perfect native (no anomalous) data. The adjusting of σ values alters the error models and thus affects χ values. When dealing with moderate to low-resolution data sets, the aforementioned error adjustments may produce significant hurdles and possible bias. X-ray diffraction experiments involving weak scatterers should receive even more attention pertaining to how error correction is approached.

The culmination of each portion of data reduction is a single file that consists of h,k,l indices and their corresponding intensities and errors, used to determine the phase and trace the best solution of the protein's structure.

Chapter 3

Experimental Design

Each data reduction program has its own approach to indexing, refinement, and integration of the R1, R2 and R1-R2 merged data sets. Except in the case of MOSFLM, which uses SCALA, each data reduction program contains it own scaling algorithm. My original intent was to compare the scaling algorithms from each program against 3DSCALE (40). Formatting issues concerning the conversion of integration files from HKL, XDS, and MOSFLM prevented this comparison. Although it is widely accepted that indexing and integration are the pivotal portions of data reduction I will review the scaling and error correction used by the various programs in relation to 3DSCALE. The scaling algorithms used by HKL2000, d*TREK, PROTEUM2 and SCALA utilize empirical spherical harmonic scaling (ESHS) (54), while XDS and SCALA, if chosen by user option, employ detector scaling (DETS) (55). Both of these methods are extensions of Hamilton, Rollett and Sparks scaling algorithm which applies two constants (**S**, **B**) to all observed reflections within the data set (56). The **S** constant is the scaling factor which essentially places intensities on a common scale throughout the data set, while the **B** term is a isotropic factor intended to correct resolution dependent errors. The operation of the scaling factor, **S**, has not significantly changed throughout scaling algorithm development. Conversely, the isotropic factor, **B**, has been improved upon to deal with continuing error evolution due to advancements in the technology of data collection and X-ray generation. ESHS replaces the **B**-factor with a spherical harmonic function which is defined in detail in *Methods in macromolecular Crystallography* (57). The DETS method modifies the isotropic **B** factor by applying a scaling factor to specific portions of the detectors' surface. The scaling factor used for a specific reflection is part of a weighted average used throughout the corresponding diffraction

image (55). Both ESHS and DETS use a modification of isotropic scaling factor B, but neither method emphasizes the need for experimental error correction. HKL2000, d*TREK, PROTEUM2, SCALA and XDS employ error models which attempt to simulate experimental error based on predefined functions. In an effort to overcome the insufficiencies of error-model correction techniques, a model free approach was proposed by Zheng-Qing Fu (58) referred to as 3DCS (Three dimensional model-free error correction or scaling), which will now be reviewed.

3.1 3DSCALE

3DSCALE is part of a larger data-processing program suite titled SGXPro (40). The 3DCS algorithm is employed by 3DSCALE for scaling integrated intensities generated from various data reduction program. Although this program can read several format templates, neither HKL2000, XDS nor MOSFLM offer a integrated intensity file for external scaling. The goal of 3DSCALE is to correct the experimental error during scaling, while avoiding the bias of traditionally used error-modeling systems. The imprecise description of error modeling employed during data correction is subject to bias and limited by theoretical variations allowed by such a model.

In an effort to overcome the insufficiencies of error-model correction a technique, using a model free approach, titled 3DCS. This algorithm was first implemented into PROSCALE (59)as part of the Bruker-AXS data reduction program PROTEUM and as a portion of the SGXPRO processing suite. This method offers an efficient means to address many of the experimental errors associate with data collection. The concept of correcting experimental errors involved with data collection is extremely convoluted and difficult to express in a single comprehensive formula. A 3 dimensional symbolic function $C(\zeta, \eta, t)$ was devised which includes several factors that influence errors associated with data collection. $C(\zeta, \eta, t)$ is best described as a

culmination of factors, several of which are contain multiple errors such as A(ζ, η, t) – absorption, R(ζ, η, t) – radiation damage, D(ζ, η, t) – detector defects, and X(ζ, η, t) – X-ray source deficiencies are a few of the terms used. These are combined as:

$C(\xi,\eta,t) = A(\xi,\eta,t) * R(\xi,\eta,t) * D(\xi,\eta,t) * X(\xi,\eta,t)$.... and so on.

The complexity of each sub-function contained within C(ζ, η, t) can be illustrated by examining a single component A(ζ, η, t) as an example. Among other influences A(ζ, η, t) analyzes crystal orientation, (air path between the collimator and detector), crystal habit, solvent content, diffraction geometry, molecules which inhabit the crystal itself, and mounting scheme. The three dimensional functions comprising C(ζ, η, t) do not posses any intuitive or theoretical quantities. (ζ, η) serve as diffraction spot coordinates and (t) is the span time required for data collection (58).

The correction function C(ζ, η, t) is applied to each diffraction spot for correction and scaling. This is easiest understood by examining intensities and associated error as follows for the eventual formation of I/σ_I, $I_j = C(\xi,\eta,t) * I_j^0$ and $\sigma_j = C(\xi,\eta,t) * \sigma_j^0 + * I_j^0 * \sigma_c$. The intensities and standard deviations σ_j^0, I_j^0 and σ_j, I_j represent the j^{th} reflection both before and after the error correction is applied.

The formulation of C(ζ, η, t) is determined from a least squares procedure which minimizes a target χ^2 function allowing for specificity of C(ζ, η, t) for the individual data set being analyzed. To further specify the error function C(ζ, η, t), a batch of images are considered as a stack of consecutive frames (τ) divided into sectors (a) which are further divided into small angular radial bins (r). Thusly, the (ζ, η, t) space is segmented into $N_\tau * N_a * N_r$ blocks. An adjustable parameter is allocated to the edges of each block. This implies the number of parameters which must be considered when calculating the error correction for each reflection

contained in a data set is defined by $N_\tau * N_a * N_r$. These values are used during the least squares procedure in the formulation of C(ζ, η, t). This method of error correction defines the amount of "correction" assigned to each reflection from parameters surrounding it. There still exist the possibility of incorrect estimations of the standard deviations within the data, which could prove problematic and skew the resultant correction constants. To combat this 3DSC algorithm uses a statistical cross-validation technique to legitimize the formulation of correction parameters by using a free R_{merge} test (60-62). A subset representing 5% of the data (randomly chosen unique or non-symmetry related reflections) of the total data are "set-aside" and not involved in error correction procedures. The "free" data will be scaled but not corrected. This will provide a test of how well the scaled corrected intensities compare with the scaled uncorrected values. This method of error correction validation is unique to the 3DSC algorithm. The remaining 95% of the data will be evaluated via the least squares refinement to minimizing X^2 values (56). The following formulas are used for this purpose:

$$\chi^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N_i}(I_j - I_i)^2}{\sigma_j^2} \qquad \text{eq 3.1}$$

$$I_i = \frac{\sum_{j=1}^{N_i} W_j I_j}{\sum_{j=1}^{N_i} W_j} \qquad \text{eq 3.11}$$

$$W_j = \frac{1}{\left(E_1\sigma_j^2 + E_2 I_j^2\right)} \qquad \text{eq 3.12}$$

N_i represents the number of unique equivalents for the i^{th} reflection. I_j, σ_j^2 terms signify the intensity and variance, respectively, of a j^{th} reflection after the C(ζ, η, t) correction and scaling.

$$\left\langle \chi^2(E_1,E_2) \right\rangle = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{j=1}^{N_i}\left(I_j - I_i\right)^2}{\sum_{j=1}^{N_i}\left(E_1\sigma_j^2 + E_2 I_i^2\right)} \qquad \text{eq 3.4}$$

The vales of E_1 and E_2 are automatically determined by an additional least-squares approach in an effort to achieve a value of 1.0 for χ^2 analysis (58). This measure of scaling quality is similar to the validation used by SCALEPACK, which is a component of HKL2000. During convergence the $\pm I_i, \sigma_i$ values of each diffraction spot are further scrutinized until the newly corrected and scaled intensities are ready for further structure determination.

A comparative study was conducted between commonly used error correction and scaling programs in order to test their results versus those from Dr. Fu's 3DCS error correction and scaling program. Four scaling and error correction algorithms were used; these were ISOS, ESHS, DETS, and 3DCS. ISOS was implemented by SCALA (used by MOSFLM), ESHS performed by SCALEPACK (HKL2000) and SADABS (PROTEUM2), DETS scaling was conducted by XDS. All algorithms were executed using default settings. Three data sets were used in this comparison: Insulin, CBP (C-terminal domain of a Corrinoid-binding protein), and Pfu631545. The quality of the data sets ranged from high to poor quality to provide a range tests for each scaling algorithm.

Method	R_{merge}	Complete-ness (%)	Redundancy	$I/\sigma(I)$	N_{tr}	TR%
Insulin						
ISOS	0.057 (0.103)	96.9 (83.6)	15.8 (3.8)	50.4 (7.7)	48	94.1
ESHS	0.035 (0.077)	96.7 (81.3)	15.8 (3.7)	57.4 (13.6)	49	96.1
DETS	0.034 (0.089)	96.9 (83.6)	15.8 (3.8)	82.1 (12.6)	49	96.1
3DCS	0.034 (0.074)	96.9 (82.0)	15.8 (3.8)	60.6 (11.5)	49	96.1
CBP						
ISOS	0.053 (0.098)	99.2 (92.1)	7.7 (6.3)	28.4 (15.3)	95	76.0
ESHS	0.046 (0.067)	99.0 (92.0)	7.5 (6.2)	33.5 (17.4)	88	70.4
DETS	0.049 (0.061)	99.2 (92.1)	7.7 (6.3)	37.7 (19.8)	93	74.4
3DCS	0.049 (0.063)	99.2 (92.1)	7.6 (6.4)	34.0 (16.8)	102	81.6
Pfu631545						
ISOS	0.084 (0.281)	99.2 (96.0)	3.5 (3.2)	17.9 (4.1)	126	47.4
ESHS	0.072 (0.276)	99.3 (96.1)	3.6 (3.2)	17.0 (3.2)	147	55.3
DETS	0.077 (0.275)	99.2 (96.0)	3.5 (3.2)	18.5 (4.2)	141	53.0
3DCS	0.070 (0.254)	99.2 (96.1)	3.5 (3.2)	12.2 (3.3)	160	60.2

Figure 3.1: Comparative Analysis of Prominent Error Checking/Scaling Programs: The R_{merge}, redundancy and I/σ_I values illustrate differences in the quatity of the data. The largest differences in output are seen in TR% values. The TR% represents the percentage of total traced amino acids. – adapted from Z.-Q, Fu (2005) Acta Cryst D61, 1643-1648

The final results of Dr. Fu's study were output from SGXPRO as traced maps. As expected, the highest quality data set, Insulin, yielded no real differences between the various methods used. The CBP data set, being of medium quality, produced traceable maps in each case yet the 3DCS algorithm generated the best initial structure model with more complete chains than any of the other programs. However the most interested data set was the Pfu631545. Despite each program tracing reasonably the same percentage of amino acids, only short fragments were built which denotes a substandard structural model except in the case of the 3DCS algorithm. The final structure Pfu631545 was solved using the initial model without the use of molecular replacement. This is the primary reason 3DSCALE was chosen for this study. The 3DCS algorithm out performed all other methods and may serve as an excellent tool for scaling data of medium resolution and less than ideal quality.

3DSCALE is initiated within the SGXPRO program pallet by selecting the Data Reduction button and then choosing 3DSCALE.

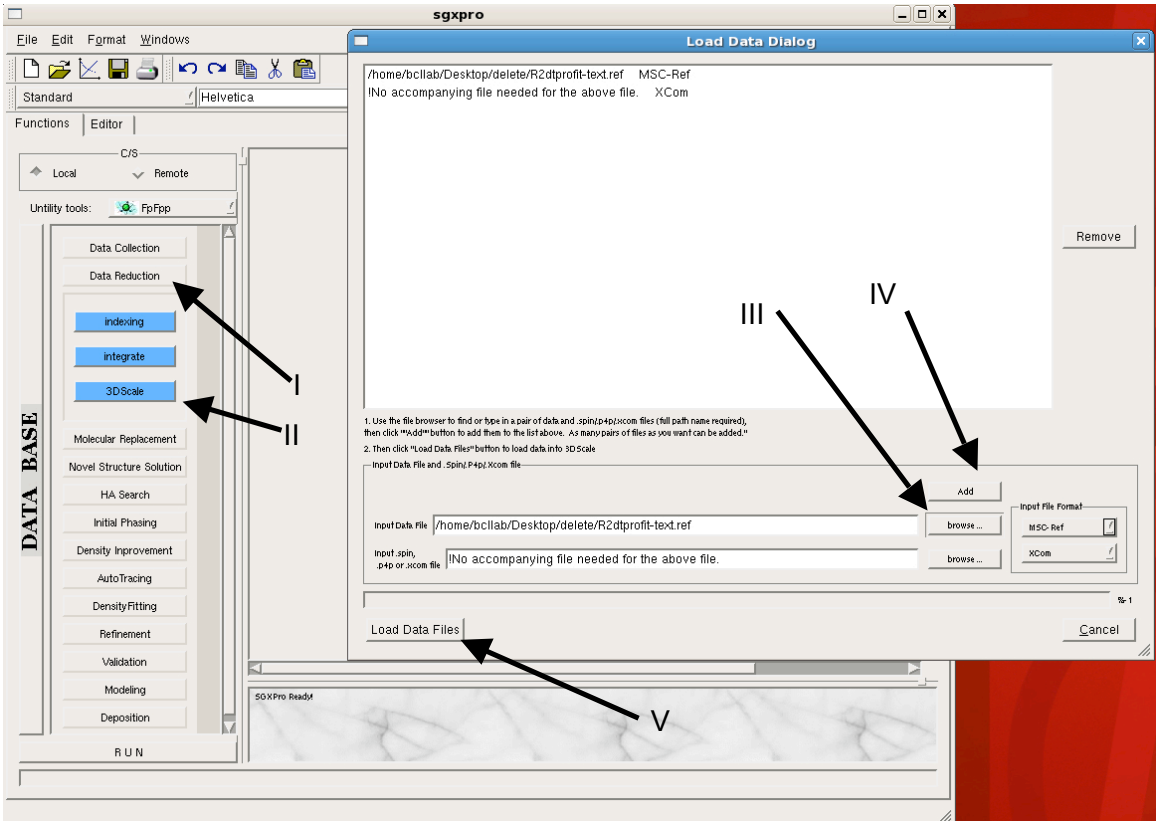

Figure 3.2: The initial/input screens for 3DSCALE input: (1) Data Reduction tab, (2) 3DSCALE selection tab, (3) integrated intensity file, (4) and (5) encompass the initiation buttons for the program.

The resultant integration files from the various programs covered in this work are added then loaded into the program. Initialing the scaling portion of this program requires little input other than the integrated intensity file. By this point in the data reduction process, the space group and resolution limits will have been discerned; the next window displayed by 3DSCALE allows the user to enter this information.

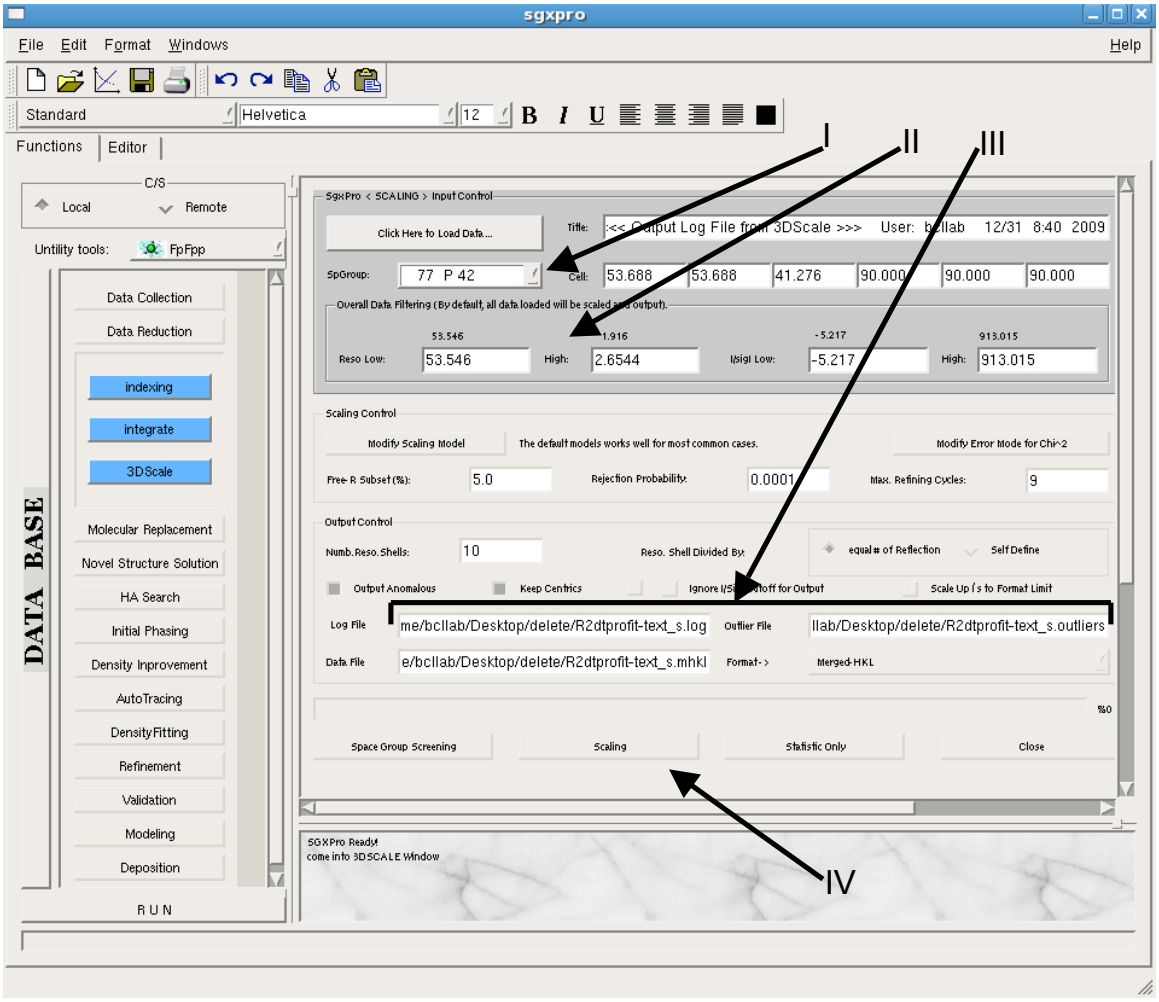

Figure 3.3: 3DSCALE Operations Window: The user will have more information concerning the (I) Space Group, (II) Resolution Range, and (III) output directories for 3DSCALE.

The output files from 3DSCALE include a log file, graphics file, and a .mhkl file (equivalent to a .sca file) the scaled and corrected data output file. The key output values seen in the 3DSCALE log are the Completeness, Redundancy, R_{sym}, I/σ_I and Ras values. Of the programs used in this work, HKL2000, XDS, and MOSFLM do not currently produce integration output files, which can be properly utilized by 3DSCALE. The scaling routines used within the original programming packages will be used except in the case of PROTEUM2 (SADABS) (63) and d*TREK. As for these two programs the original scaling results generated by PROTEUM2

(SADABS) and d*TREK's self contained scaling packages did not generated tracing results of the same quality as those results generated by 3DSCALE.

During the final process of structure solution, the creation of a traced map containing the phases calculated from the scaled intensities and the best fit of either the known amino acid sequence or a alanine traced backbone of the structure is used. For this work, a 95 polyalanine residue chain was used for tracing purposes. The model-building program used in this work (RESOLVE)(64) was also included in the SGXPRO data processing package in the "Novel Structure Solution" module (40).

3.2 SGXPro

The "Novel Structure Solution" module of the SGXPro program pallet was used to determine protein phases and initial chain trace using the scaling results from each program discussed in this work. SGXPro was chosen due the ease of use and provided for an unbiased structure determination process.

The SGXPro is a parallel workflow engine designed to offer easy access to multiple popular crystallographic programs allowing for simultaneous selection of separate paths within the structure determination process (40). In 2001, the time of SGXPro's inception, there was no program, which offered a powerful, upgradable, and user-friendly access to the various programs used during structure determination. Automated structure determination pipelines devised in the mid 2000's such as ElVES (65), AutoSHARP (66), CRANK (67), HKL2MAP (68), SCA2STRUCTURE (69), HKL3000 (70), PHENIX (71), and Auto-Ricksaw (72) lacked a systematic approach for identifying and grouping the programs needed within a single GUI environmen. SGXPro organizes the input and output files of the most popular and efficient programs by interlacing these in a parallel work flow system which will offer the most probable

results in producing a structure with little user interaction during the determination process. SGXPro possesses a simplified flowchart type of communication between the different software packages to allow for several simultaneous methods of structural determination. The input and output files are organized in a user-friendly GUI allowing for quick and efficient location of specific files.

The “Novel Structure Solution” (Figure 3.4) module within SGXPro was chosen as a means of automated structure solution from scaled data within this study due in part to the successes involving tracing initial maps from both moderate and low quality data sets.

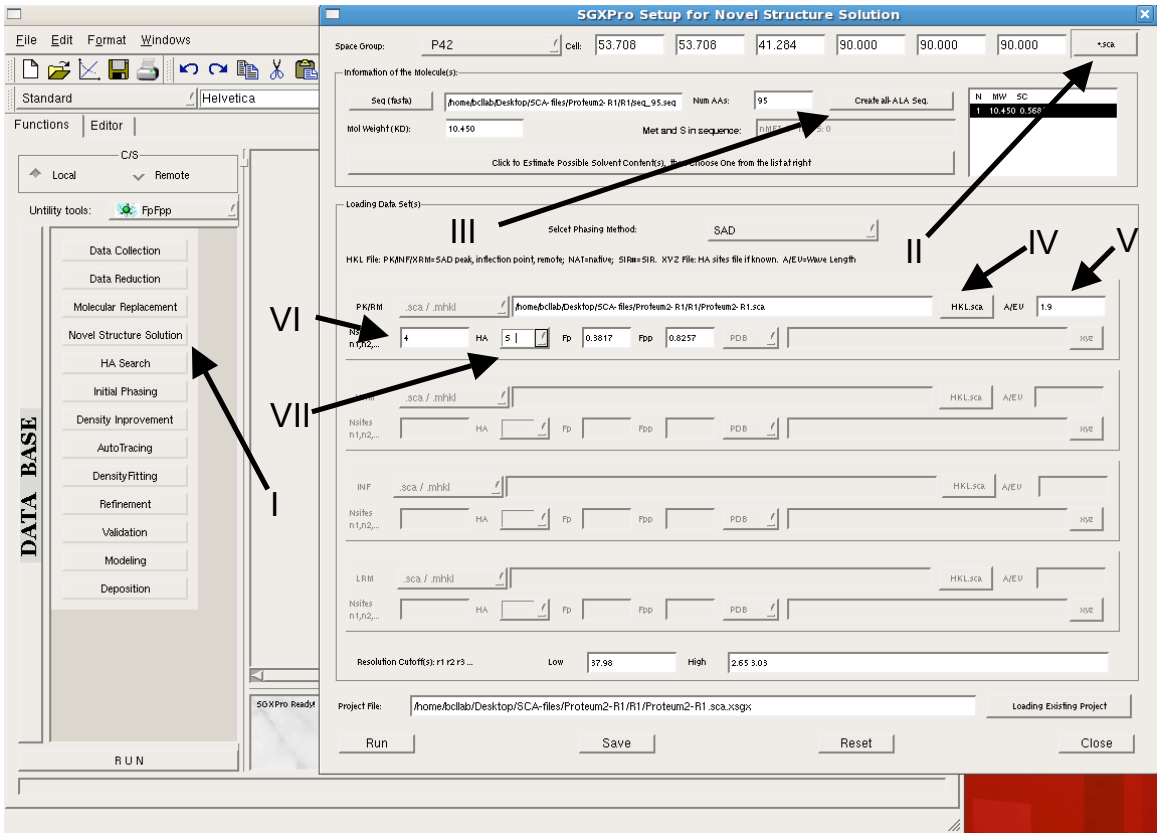

Figure 3.4: The SGXPro interface highlighting the Novel Structure Solution: (I) Opens the Novel Structure Solution GUI, (II) interface for loading the scaled intensities from earlier data reduction, (III) creation of Alanine sequence file for model tracing, (IV) second interface for loading scaled intensities, (V) wavelength identifier, (VI) number of heavy atoms to be searched, and (VII) heavy atom identifier.

In addition, minimal input is needed (scaled intensities, number of amino acids within the protein, wavelength used and type and number of anomalous scatterer). SGXPro also has a proven track record of structure solution by serving as the primary structure determination engine at the University of Georgia, and onsite at SERCAT sector 22 Argonne National Laboratory, Advanced Photon Source. Finally, SGXPro's ease of use provided for multiple modeling attempts for each individual data processing trial. The flow chart used for the Novel Strucutre Solution are as follows: SHELXD(73), ISIR(10), SOLVE(74), RESOLVE(64), CCP4(39) and finally COOT(75).

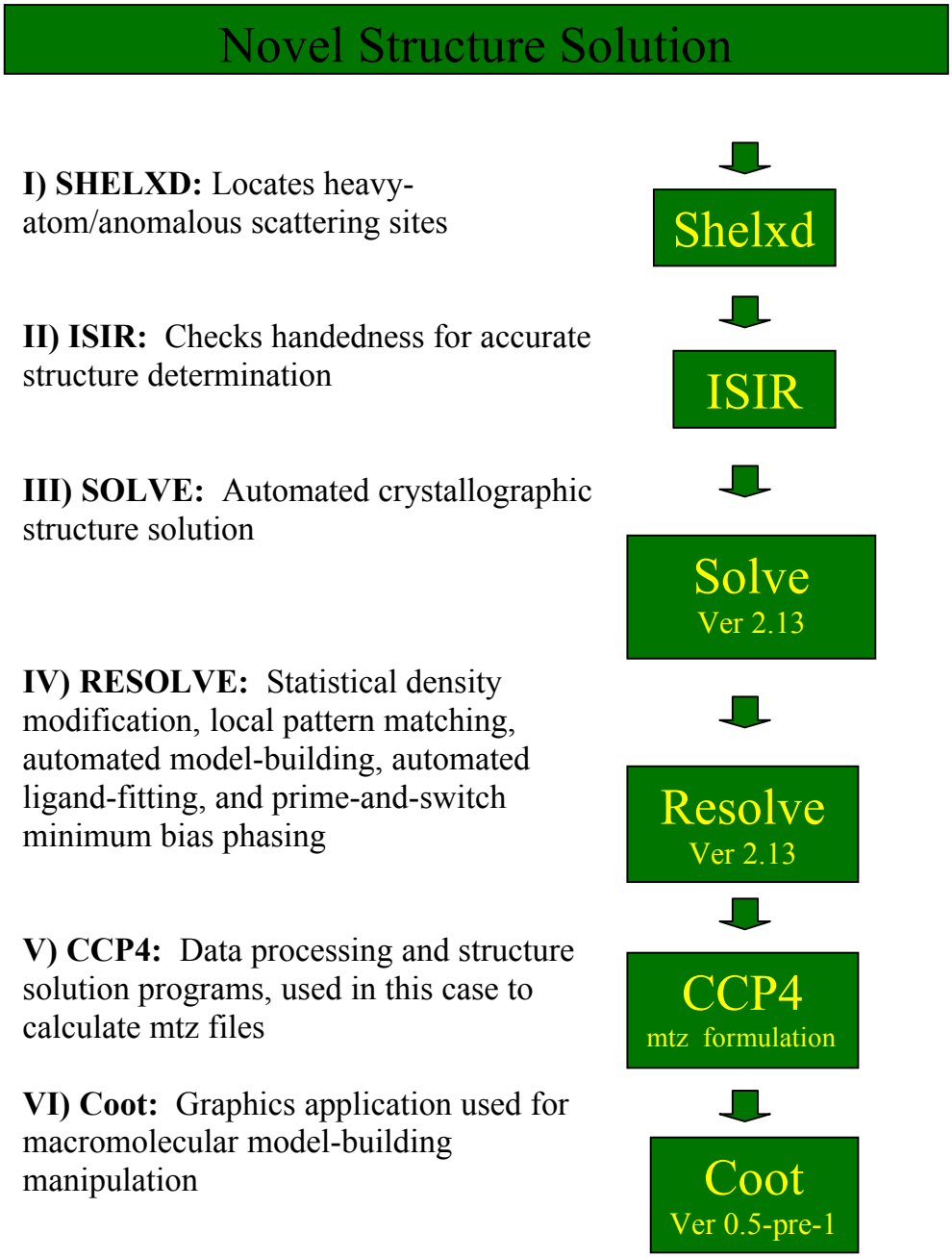

Figure 3.5: Flowchart of Novel Structure Solution: The programs used herein are not related to any of the data reduction program studied in this work.

The output files from each program in the Novel Structure Solution flowchart is automatically passed to the next program. This fact highlights the useful attributes of the SGXPro pallet. After executing the Novel Structure Solution program the five best structural solutions are were placed in processing specific folders for user review.

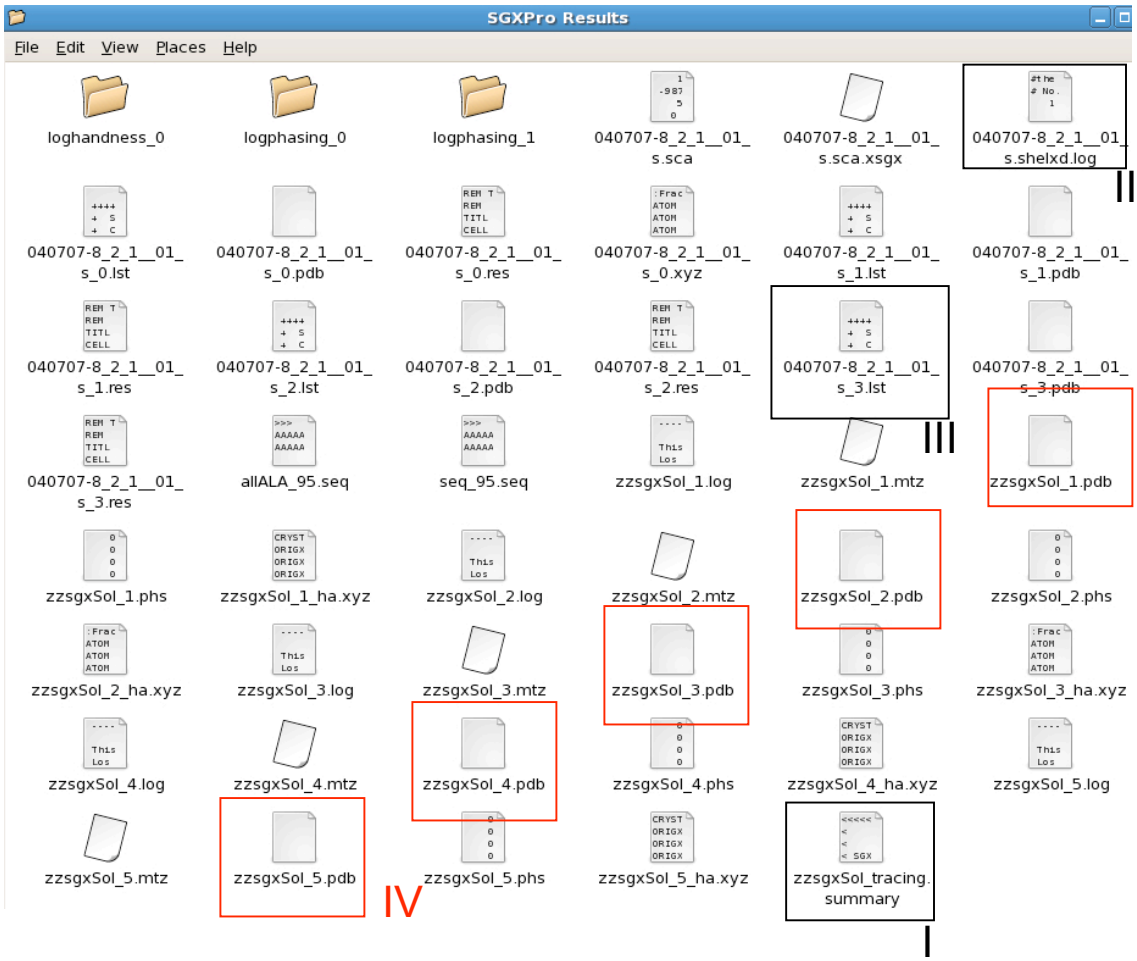

Figure 3.6: Output files from SGXPro: (I) The tracing summary for the five best solutions generated, (II) ShelxD log containing the quality of the programs ability to locate the heavy atom substructure, (III) inter-atomic distances between each of the heavy atoms located during processing. (IV) Structure solution files.

The PDB coordinate files represent the five best trials produced by the program. At the end of the calculation, the program automatically opens COOT (75) displaying the best of the five solutions obtained according to the program. The secondary solutions can be viewed as well by user choice. The solutions are organized and recorded by a solution number within the folder, which contains the scaled intensities used for structure solution. For the general user files of importance are the RESOLVE Summary output file, ShelxD, and ShelxD-lst files.

Summary output file:

No# (i)	NumBuilt (ii)	NumSegs (iii)	Top3Segs (iv)	Model Built (v)
1	85	22	9 6 6	/home/bcllab/Desktop/R1/t1/zzsgxSol_1.pd

Heavy atoms: experimental phasing: /home/bcllab/Desktop/R1/t1/zzsgxSol_1_ha.xyz

ShelxD – Heavy atom locator:

#the log file includes the following columns in order:

# No.	Ntry	cc_all	cc_weak	PATFOM	FILENAME
1	7	43.85	23.09	78.88	/home/bcllab/Desktop/R1/t1/tucker1_3.lst

Heavy Atom peaks/Inter-atomic distances

Peak	x	y	z	self cross-vectors
99.9	0.6266	0.1586	0.4101	21.7
				138.3
89.9	0.6619	-0.0308	0.4685	17.6 10.6
				139.8 68.8
71.4	0.8389	0.2179	0.3296	29.0 12.3 17.3
				68.7 68.5 24.9
52.3	0.6627	-0.1278	0.4771	22.2 15.7 5.2 16.0
				128.1 50.2 101.5 0.0

Table 3.1 Excerpts from SGXPro output files: Summary - (i) solution number (1-5), (ii) number of residues built, (iii) number of segments within the solution (iv), top 3 segments traced, and (v) directory information concerning the corresponding .pdb and .xyz file location. **ShelxD –** Measure of the quality of heavy atom location. **Heavy Atom –** differences between identified heavy atoms sites.

The Summary output displays information concerning the trace amino acids from SGXPro (RESOLVE). The ShelxD file selects the best correlation coefficients CCall and CCweak (76) which measure the agreement between calculated and observed heavy atom peak positions from Paterson Maps. The PATFOM (Patterson figure of merit) measures the consistency between observed and calculated peak positions. The magnitude of these peak positions from the Patterson Map are normalized to a maximum of 99.9 for easy differentiation between probable heavy atom positions (70-90 range) and suspect heavy atom positions (0-60 range), in this case

these are Sulfur positions. The best indicator of properly located heavy atoms are the CCall / CCweak values. Generally, CCall/CCweak values of 30 / 10, respectively indicate accurate heavy atom placement within the model. The PATFOM values are traditionally large, above 120, for properly located heavy atoms. However this often provides false negatives pertaining to viable solutions with PATFOM values below 85 corresponding to solutions of nearly the same quality of those solutions containing PATFOM of 125. Of these two indicators the CCall and CCweak values are relied on more often than the PATFOTM.

The user will find it necessary to examine all five solutions generated, as the initial solution may not necessarily be the best. For example, a solution which RESOLVE traces 89% of the total number of amino acids comprising the protein appears to be a good but residues traced is not nearly as important as the ratio of the number of amino acid segments versus the number of amino acids built. If a solution traces 85 of 95 amino acids, but the number of segments is 22 and the top three chains are of lengths 9, 6, and 6, the model will be quite poor and badly traced when viewed within a molecular graphics application. The relationship between the number of amino acids traced (NumBuilt) and the number of segments (NumSegs) is more important than the total number of amino acids traced within the model alone. The following represents the results from SGXPro “Novel Structure Solution module”;

Traced Alanine protein structure map Test-set.pdb

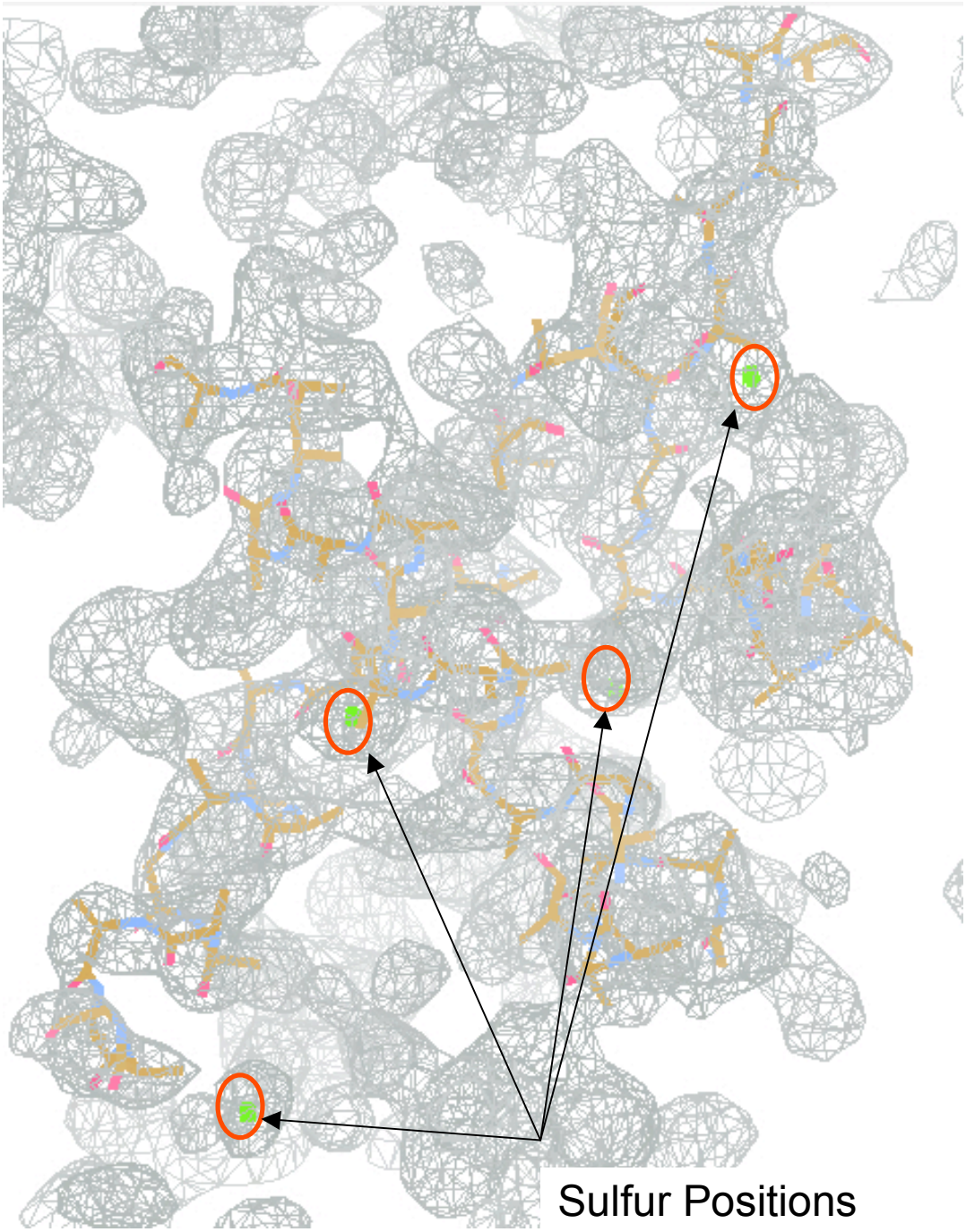

Figure 3.7 Coot output from SGXPro, Test-set.pdb: Corresponding to the statistics in Table 3.1. The tracing results of this program were used as a partial judgment of the quality of each data reduction program.

From this study the best solutions were found to have a ratio of Numbuilt and NumSegs of approximately 1 to 9 or better. This information is contained in the Summary output file;

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
(i)    (ii)     (iii)     (iv)        (v)
-------------------------------------------------------------------
 1     58        5     20 15 14 /home/bcllab/Desktop/R1/t1/Test-set.pdb
Heavy atoms:                   /home/bcllab/Desktop/R1/t1/Test-set_ha.xyz
```

Table 3.2 Excerpts from SGXPro output file, Test-set.pdb: Summary - (i) solution number (1-5), (ii) number of residues built, (iii) number of segments within the solution (iv), top 3 segments traced, and (v) directory information concerning the corresponding .pdb and .xyz file location.

The SGXPro solutions for R1, R2 and R1-R2 merged data sets will be discussed within the data reduction subsections that follow.

Chapter 4

Data Reduction Programs

The extent to which I am able to interpret the approach each data reduction program employs during indexing, refinement, integration and scaling is limited by what their authors were willing to disclose. MOSFLM is the only program within this study which is universally distributed free of charge regardless of academic or industrial use. XDS is free for academic use while requiring purchase and licensure for industry or for-profit organizations. HKL2000 requires purchase for both academic and industrial use, except for the case of temporary licensure allowing the experimenter a 6 months license for processing data collected at a synchrotron source. PROTEUM2 and d*TREK were designed for specific detectors and are usually sold as part of the package involving the purchase of an X-ray detector. Most program authors are reluctant to divulge specific information pertaining to the inner workings of their programs. Nevertheless, I have found a few authors who were generous in sharing their knowledge. I continue to extend my thanks and appreciation to them.

4.1 HKL2000

HKL2000 is the most commonly used data reduction program currently in use (28). My recent investigations into the PDB show HKL/HKL2000 data reduction program as responsible for over ~65% of submitted protein structures. It is well known that its authors are extremely protective concerning the details of the various algorithms used in the program. My efforts here are to provide an overview of the practical and theoretical use of the program.

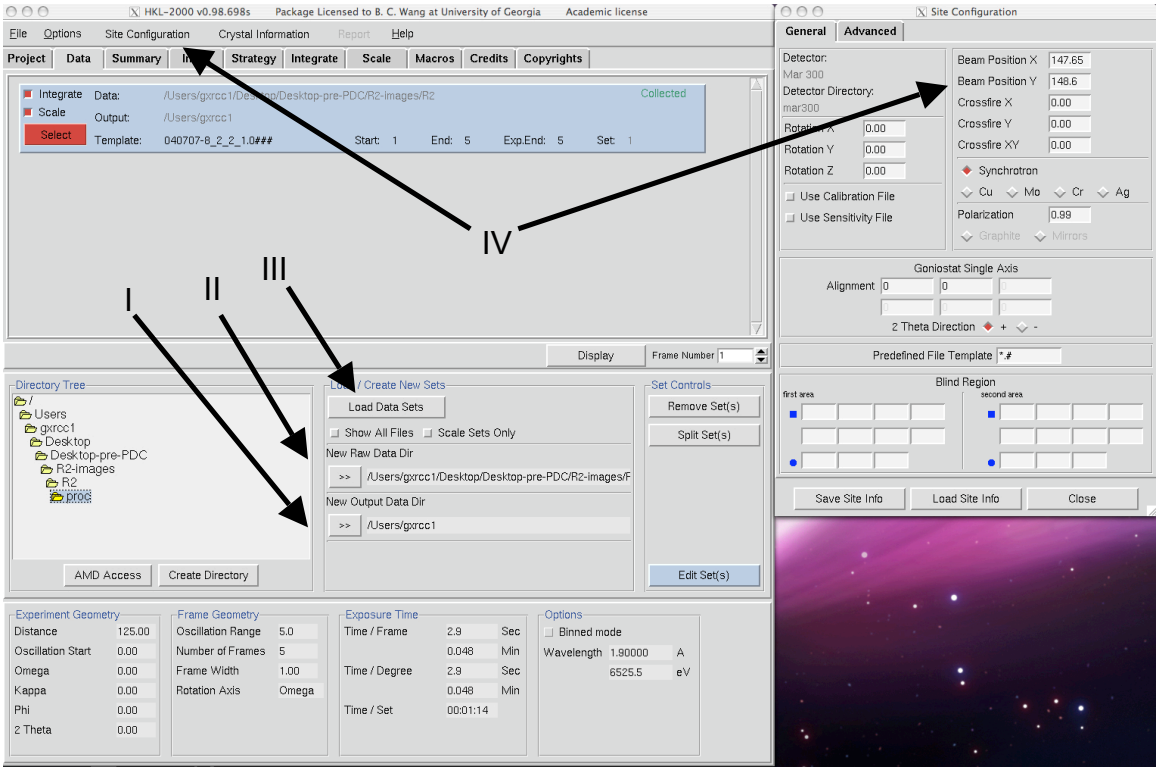

Figure 4.1.1 Initial HKL2000 Screen: After identifying the detector used during data collection to initiate HKL2000 processing (I) the image directory should be identified, (II) an output directory for results, (III) actually loading the image files, (IV) and entering the beam center from the Site Configuration Tab.

Once initiated, a HKL2000 GUI prompts the user to identify the detector type on which the data to be processed was recorded. The location of the data and an output directory for results are identified (Figure 4.1.11; I, II). The user then, loads the images and identifies the beam center. Beam position is input via selection of the System Configuration tab (Figure 4.1.11;III, IV) located in the menu bar

Finding Spots

The spot finding portion of the HKL2000 program suite, commonly known as DENZO, is initiated by selecting the Index tab (Figure 4.1.11; I) and then the Peak Search button (Figure 4.1.11; II).

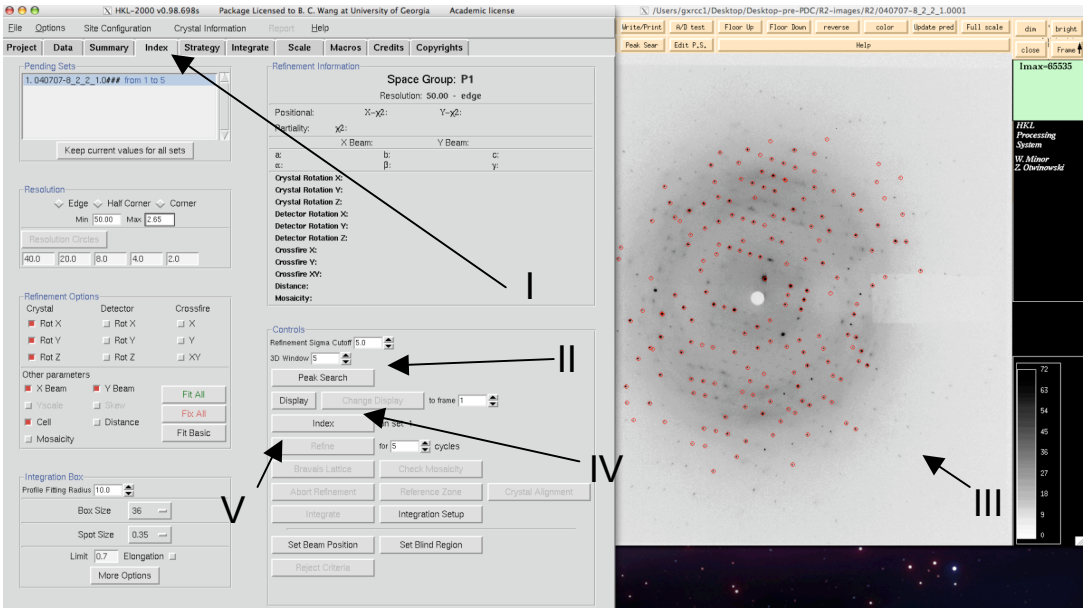

Figure 4.1.2: Indexing Tab/Peak Search from HKL2000 Processing GUI: (I) The Indexing Tab, (II) The Peak Search button initialing both indexing and spot finding process, (III) the initial images detected by HKL2000, and (IV) the Index button to initiate indexing, and (V) button used for refinement (not available until after indexing).

DENZO's default settings uses the first five images from the initial images loaded for indexing. Selecting the Peak Search button initiates a search within the loaded images and chooses an adjustable number of peaks from a single oscillation image (Figure 4.1.2; III). Manual adding or removing spot selection is allowed. It is important that the lunes present in each diffraction image have adequate separation to prevent overlapping. If this is not done, spot locations may not be recorded properly and adversely influence refinement parameters. An overlap of lunes could also imply the crystal contains more than one lattice (twinning), which can affect indexing and the resulting orientation matrix (28). The oscillation range at which the

data was collected should be chosen such that a fluid succession of the diffraction patterns between images is apparent. These requirements being fulfilled, the next step in indexing is mapping the diffraction maxima of the spots identified in the peak search by either the auto "search" or manual inclusion/exclusion.

Indexing/Refinement

DENZO uses the center of the oscillation range as a best estimate for the angles at which each diffraction spot occurs. An Auto Indexing algorithm is employed which searches all reflection positions, found during Peak Search, for all possible indices until the program finds integer values of one index (h, k or l) for all searched reflections. This is paramount to finding one real space direction of the crystal axis (a,b,c). This method is called *real-space indexing*. The placement of one real space vector is tantamount to finding the regularity for the reciprocal lattice in the vector's direction (77). DENZO utilizes the fast Fourier transform method of searching for real space vectors. After Peak searching the user chooses the Index button (Figure 4.2.1; IV), which will conduct an initial indexing pass and open a GUI displaying a list of Bravais Lattices. The best cell choices are displayed for all 14 Bravais lattices accompanied by a distortion index value. This index is a measure of the degree of distortion needed to "strong arm" the cell dimensions of a chosen lattice type to match experimental to predicted peaks. The unit cell parameters a,b,c and α,β,γ generated in this list are not yet refined. DENZO allows the user to choose how to refine several key indexing parameters; as follows:

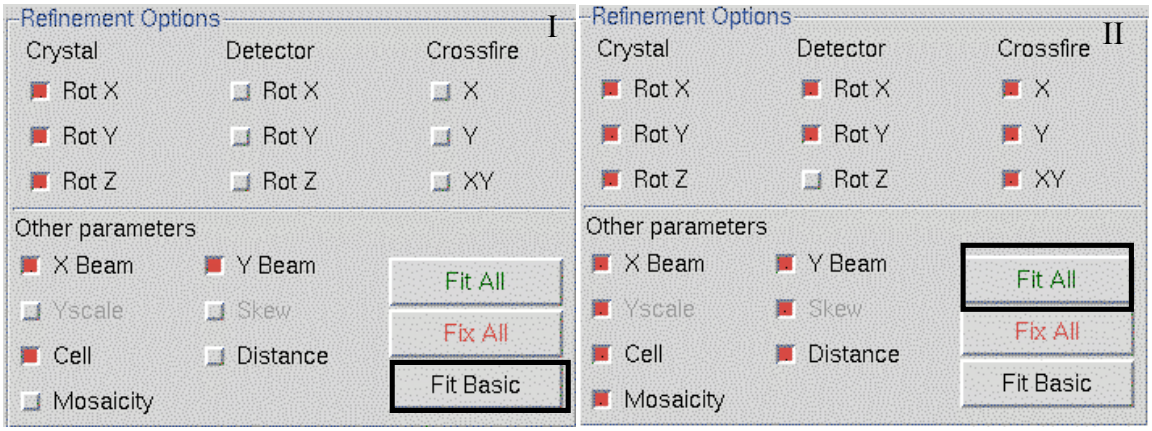

Figure 4.1.3 Fit Parameters for Refinement: (I) Fit Basic selection only selects the options highlighted by red selection square; (II) Fit All selection selects a more complete set of options.

These values serve as options within the Indexing GUI for refinement. Choosing to *Fit All*, *Fix All* and *Fit Basic* will alter the refinement of correlated parameters conducted by the program, although Fix All is not commonly used.

Crystal rot X, Y, Z	Angular deviations from he reference orientation specified by the vertical axis
X, Y Beam	Position of the origin or direct beam center
Cell	Unit cell lengths and, once selected, the Bravais lattice angles
Detector rot X, Y	Corrections for the detector face rotational offset
Crossfire X, Y, XY	Measure of beam divergence effecting the prediction of partial reflections positions
Yscale	Anisotropic correction factor in the pixel dimensions
Skew	Refines the non-orthogonality of the vertical and horizontal scanning directions, this feature is no necessary for CCD detectors and has a value of 0.0
Distance	Detector distance from crystal to beam spot on detector
Mosaicity	The rocking angle, which would describe all spots seen on a single diffraction image; a measure of the order within the unit cell.

Table 4.1.4: Indexing/Refinement options within HKL2000: These values can be refined simultaneously but HKL2000 highly correlates several parameters this refinement process should be carefully considered.

There are different options the user has in addition to the "Fit" refinement settings such as the reference zone, setting blind region and beam position, and setting rejection criteria. In addition to these options, and other more advanced parameters such as the Crossfire and Rotation (Figure 4.1.3; II) all fall beyond the realm of standard user interaction. Only experienced users would properly benefit from attempting to do more than accept the values HKL2000 offers during refinement. The HKL2000 manual advises the users to select the lowest symmetry Bravais Lattices (Primitive triclinic) displayed after initial indexing. Then, conduct refinement using the Fit All designation within the Refinement Options (Figure 4.1.3). Next, the user

chooses the proper Bravais Lattice or that lattice with the highest symmetry and the lowest distortion index if no previous knowledge exists concerning the crystal's space group. After this second selection further refinement should be carried out. It is important to mention the method of refinement covered in the HKL2000 manual for Bravais Lattice selection differs from the method used in this study. I have found it beneficial after indexing (Fig 4.1.2; IV) to accept the default, Primitive Triclinic Bravais Lattice and refine the initial results from HKL2000 using Fit Basic instead of Fit All. It may be tempting to choose the highest symmetry group from the list, but I have found it best to wait until the distortion index is as low as possible before selecting a higer symmetry Bravais lattice. The idea here is to impose as little "stress" as possible for the corresponding fit relating to the Bravais Lattice choice or distortion index. The Primitive Triclinic lattice should be accepted and refined by continually selecting the Refine button (Figure 4.1.2; V) until the χ^2 values become steady. This should lower the degree of distortion for the intended higher symmetry Bravais Lattice that accurately describes the crystal. Next, the user should select the Fit All tab and repeat the aforementioned refinement technique while maintaining the Primitive Triclinic Bravais lattice selection. After the full refinement has converged, the user selects the Bravais Lattice tab and chooses the lattice which either corresponds to the known space group of the crystal or the highest symmetry space group identified having a low distortion value. The overall results of integration and scaling are improved using this method as opposed to accepting recommendations from the HKL2000 manual (77).

For the AF1382 test case, both methods highlighted above yield primitive tetragonal as the highest Bravais Lattice choice with the lowest distortion index. However, the level of the distortion index does differ between the two methods.

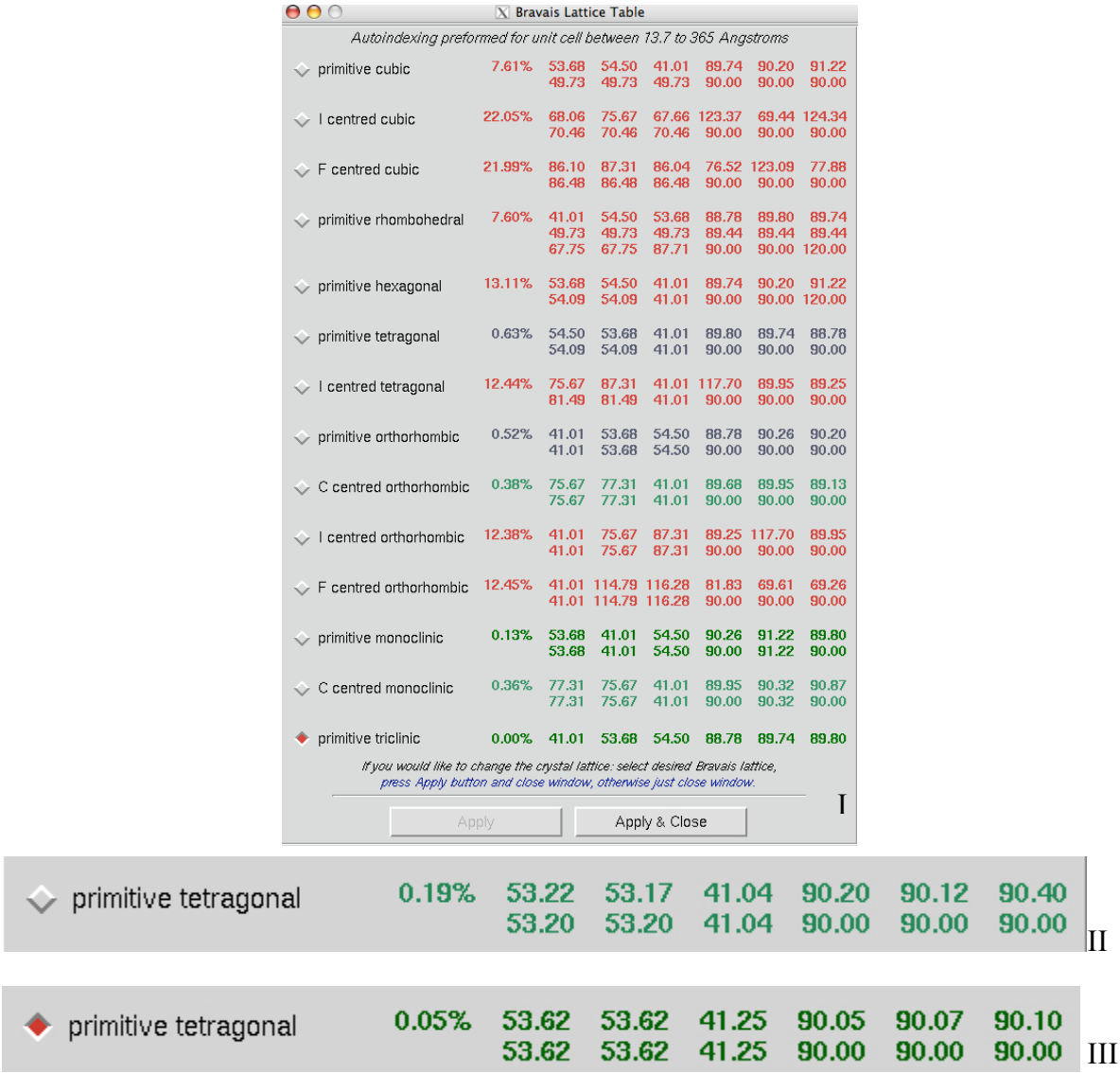

Figure 4.1.4: Revised Refinement Procedure: (I) The Bravais Lattice selection using Fit Basic parameters initially calculated by HKL2000, (II) Result of using the method used within the HKL2000 manual (III) Employing the method used in this study which yields a significantly lower distortion index for Primitive Tetragonal system.

The user should always check the indexing by inspecting the observed and predicted spot positions on the images used for indexing. The predicted spot positions from HKL2000 should match the experimental spot positions; this is highlighted using a color scheme for quality of fit (Figure 4.1.6). Green circles depict the subset of spots chosen during auto indexing, yellow are

the calculated locations of spots according to the index, and red spots are rejections. Regardless of the color of the circles, the location of the circles should readily agree with the observed diffraction spots.

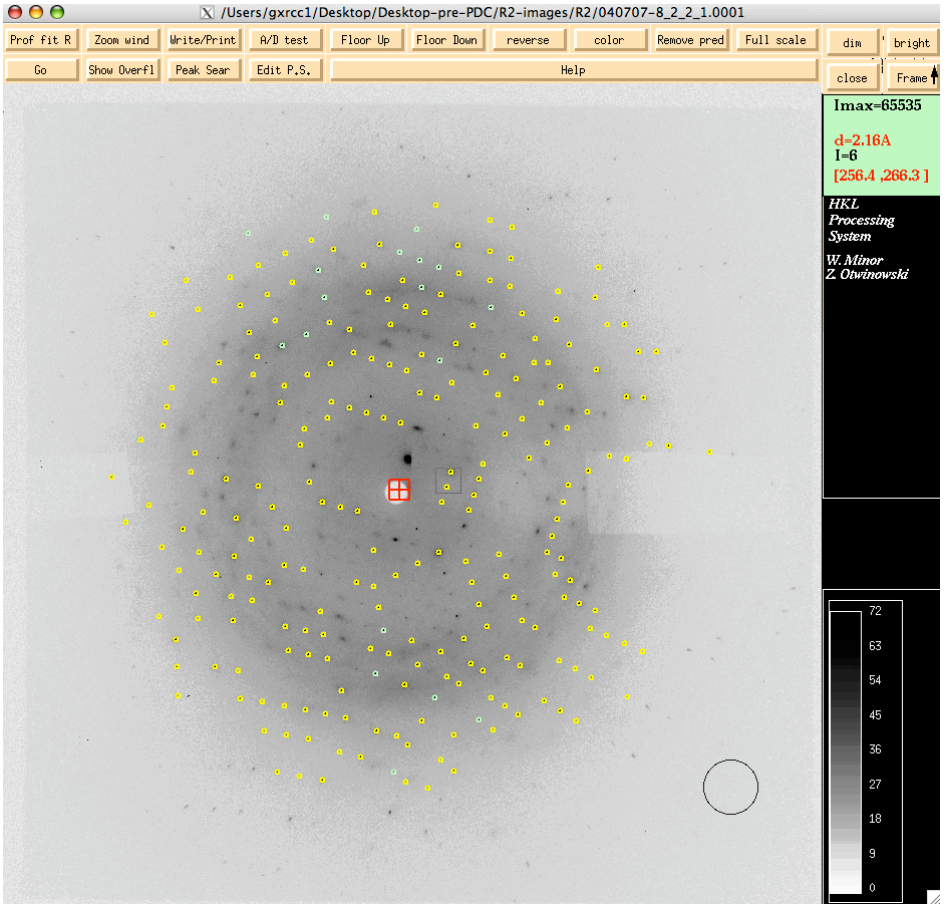

Figure 4.1.5: Peak Search Window: Agreement between the colored spots, predicted by HKL, and actual diffraction spot.

The calculated mosaicity values should also be inspected to ensure these are reasonable (0º-perfect and >2º-extremly high) or integration will likely fail (Figure 4.1.8, III). The higher the mosaicity the more elongated and wider the diffraction spots; this creates added difficulty during spot habit identification.

The size of the box and spot chosen by HKL2000 are by default 36 and 0.35, respectively, it is unclear what the dimensional measure of these values represents. The program

author claims this to be the best ratio of box to spot size. I have found it necessary to select the spot and corresponding box size such that the diffraction spots are fully encompassed. The spot fitting module of HKL2000 also allows substantial freedom in choosing the global shape of the spot profiles to include elongated spots similar to an ellipse.

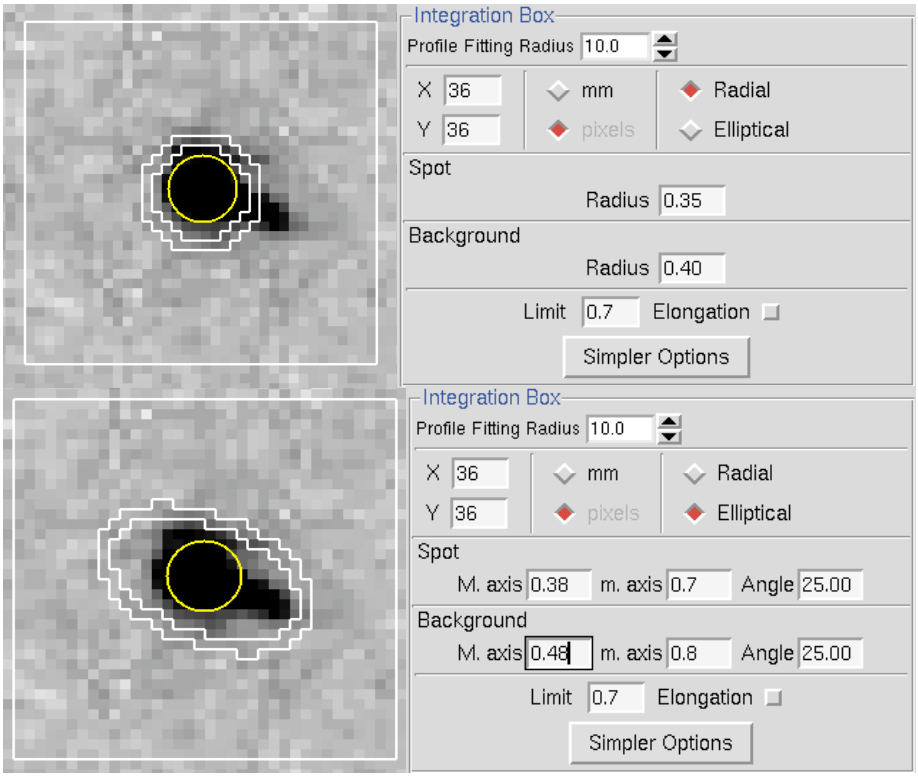

Figure 4.1.6a: Spot Selection in HKL2000: The user can specify the habit of the spot selection to best fit the diffraction spots.

In theory this would be ideal but the user needs to be very careful in the selection of these two ranges. This is because the values of the spot and box must satisfy every spot within the data set. Since parallax causes elongation of reflections at high resolution it is usually best to refrain from using an elliptical spot shape. The spot selection should encompass the entire diffraction spot using the inner diameter of the two circles. The area between the inner and outer circles encompasses a "no man's land" which should contain none of the diffraction spot. The

area between the outer circle and perimeter of the box should contain no diffraction intensity and is used for calculating the background.

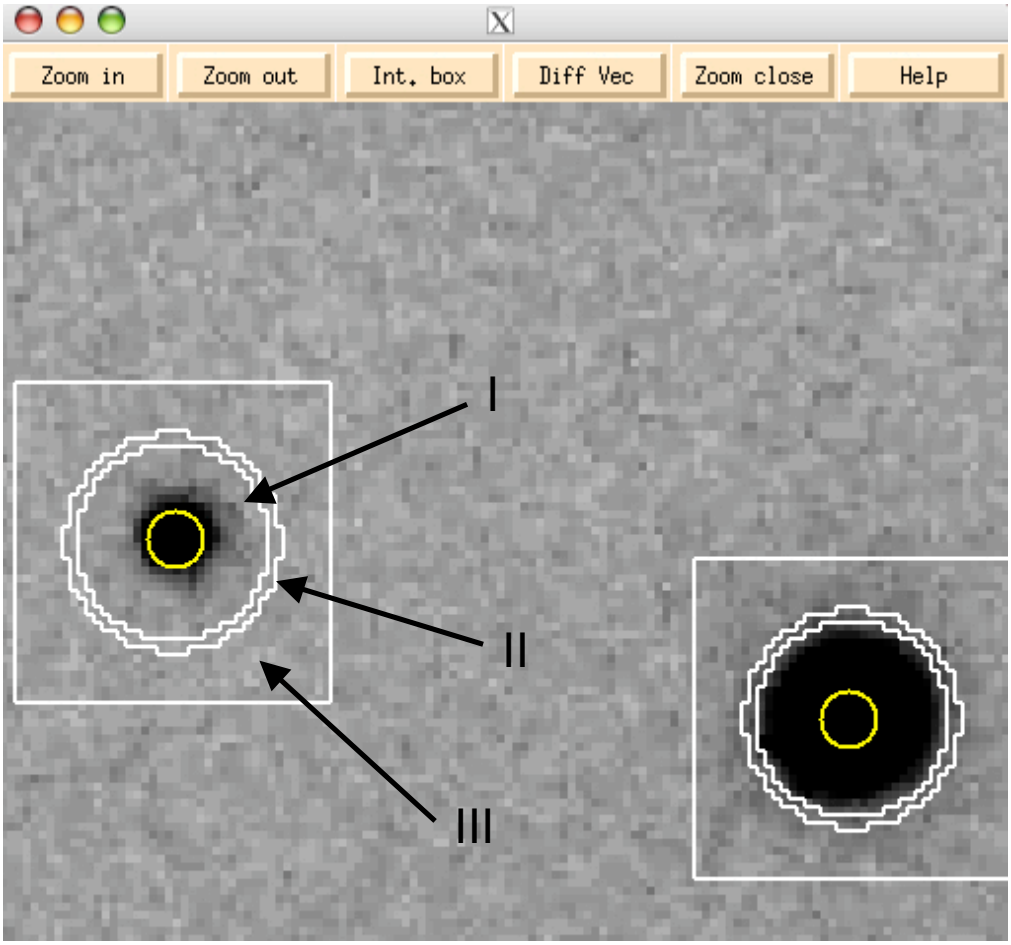

Figure 4.1.6b: The HKL2000 spot and box measurements: The area(s) within the integration box are used during integration to discern spot, and background. (I) the area which encompasses the peak, (II) the area between the two circles is considered "no man's land" and no signal is recorded (III) the area between the outer circle and the perimeter of the integration box is used to detect the background intensity.

The concept of spot and box selection is shared in some manner by all the programs covered in this study. The major differences being the method by which the data reduction programs determine the size of the spot and box. Since the spot area must encompass the entire diffraction spot it would seem the larger the spot the more range offered for choosing the center

of the reflection. Likewise, a smaller spot area would allow for a more precise choice of the reflection centroid (78). Neither of the aforementioned scenarios will satisfy every spot from the variety of sizes found in diffraction experiments (Fig. 4.1.7a,b). It is also imperative that the box size does not overlap a spot from a neighboring reflection or the background calculation will be unreliable. The negative issues of using this method of spot and background differentiation are associated with the universal choices the user must make. HKL2000 will accept the spot and box sizes chosen by the user for all spots throughout the data set. This will allow the spot size to barely encompass some spots while dwarfing (Figure 4.1.7a,b) others leaving the user in the same position as mentioned earlier. Also the aforementioned distortion index generated by HKL2000 is seldom 0% for the initial lattice system chosen, thus the predicted reflection positions calculated by Denzo will be slightly askew from the experimentally observed reflection locations. This slight shift in spot location, coupled with the blanket spot and box size used by HKL2000, may cause an offset in the spot centroid position that is used for integration. This is especially true when considering the weak anomalous signal which can easily be lost due to these factors. This is easily seen when parallax is considered as diffraction spots elongate in relation to the radial distance from the beam center. The change in spot shape due to parallax or high mosaicity can be extreme in some cases, and a blanket spot shape selection may not collect all of the diffraction spot. Finally, any numerical values in the portion of the Index/Refinement GUI from HKL2000 that are red in color should be considered before continuing with data reduction. The χ^2 values which should be ~1 and green in color ideally, but orange is acceptable for integration.

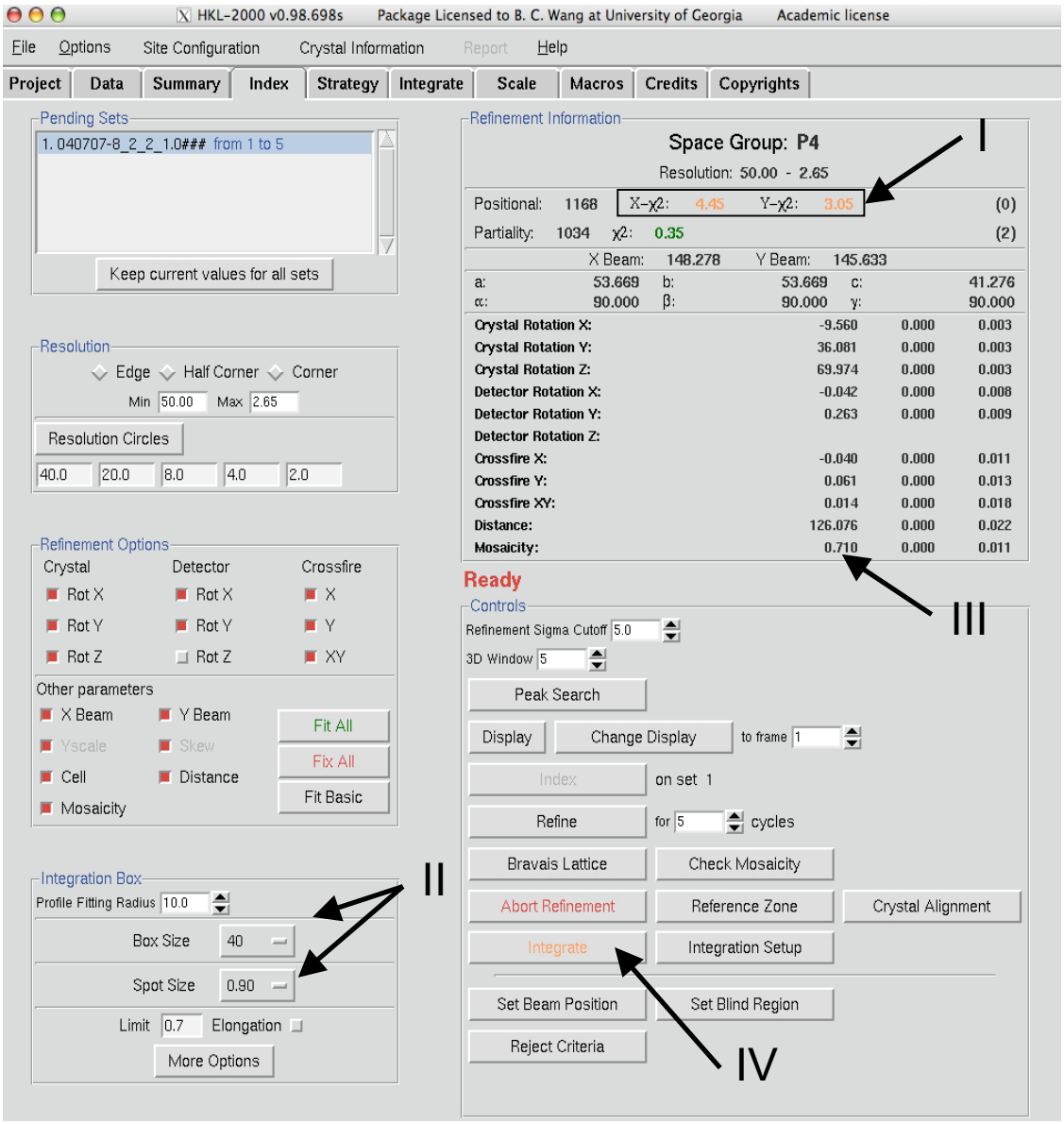

Figure 4.1.7: Final Refinement before Integration: (I) the χ^2 values associated with refinement, (II) Spot and Box size necessary for accurate inclusion of both large and small diffraction spots within the data set, and (III) the calculated Mosaicity from the indexing algorithm.

Problems involving indexing most commonly arise from incorrect beam position or unknown spindle directions. Double checking the beam center and ensuring the correct and current HKL2000 site.def file, which describes the experimental setup, for the detector being used (79).

The automatic indexing method can choose spots from a single oscillation image to formulate indexing parameters for the entire data set (77). The only requirement is that the oscillation range be small enough that the lunes (diffraction rings seen on oscillation images) are easily resolved to prevent mis-indexing. By utilizing just one image during indexing, Denzo will identify all spot predictions for the complete data set for the purpose of integration. If there is crystal slippage or any deformity of the lattice it may not be apparent from the first frame or even in the first five frames which represent 0.8 – 1.4% of typical 360 degrees data set. Most data reduction programs use a varying range of images throughout the data set for indexing. The default image range used by HKL2000, 5 images, assumes the entire data set can properly be defined from such a small percentage of images. If the data is robust such a small range may be fine but in the case of medium quality data this could prove problematic.

The next step in the data reduction process is Integration. This is accomplished by selecting the Integrate button, which will be colored as a reflection of the quality of the χ^2 values (Figure 4.1.8; I, IV) generated during indexing.

Integration

During integration, every oscillation frame is processed independently generating a .x file for each image. This file describes the reflections from the frame in 2-D space and normalized by the background values calculated from the size of the box selected from each reflection (28).

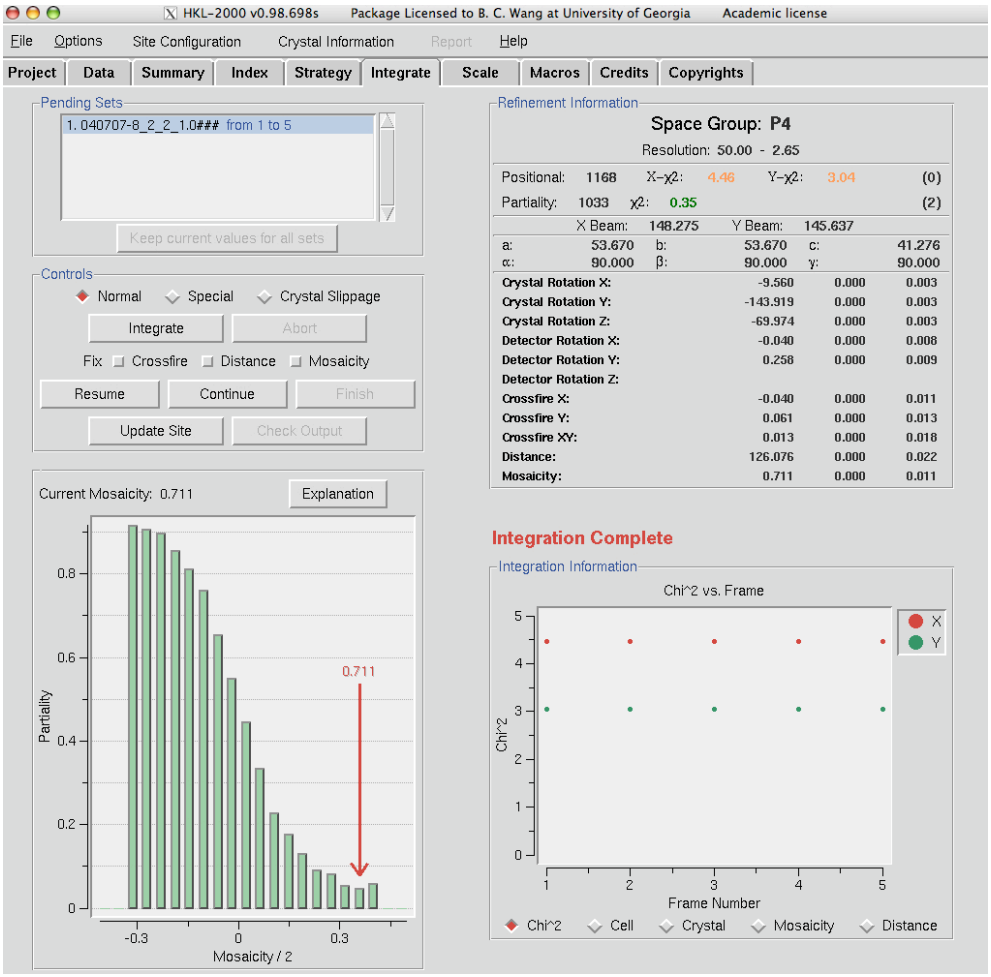

Figure 4.1.8: The Integration GUI within HKL2000: This window displays the values associated with the refinable parameters corresponding to the frames within the data set.

DENZO uses 2D profile fitting generally described in for integration. The use of profile fitting necessitates precise peak estimation due to the identified peaks being used to create the profiles to be used in the profile fitting process. DENZO averages the observed profiles by superimposing these peaks atop one another. If the peaks are not properly predicted, resultant profiles will be displaced or broadened inducing additional error.

The fitting process used in HKL2000 suggest estimates of the average spot profile is defined by the observed profile M_i minus the predicted profile P_i. The variance of M_i is a function of the expected signal in a pixel represented as V_i, the formula is as follows:

$$\sum_i \left[\frac{2(M_i - P_i)}{V_i} \right] \qquad \text{eq. 1.0}$$

The index *i* represents all pixels within a 2D profile (28). This method of profile fitting increases the accuracy of each diffraction spot by decreasing the statistical error, but the fitting method will produce systematic errors throughout the data set if initial spot prediction is inaccurate. The method of background approximation within DENZO relies on the assumption that the background is a linear function of the detector coordinates. Certain scenarios will cause HKL2000 to exclude pixels or diffraction spots from processing. The reasons for this exclusion are most commonly associated with three cases. First, the most common reason for rejections is ~~are~~ overloads. Overloaded pixels are ignored during the intensity calculation process during profile fitting, so overloads should be avoided. The most commonly overloaded pixels that occur are within 4Å of the beam stop. This resolution also coincides with key areas of phasing information, due to atomic scattering power decreasing with increasing scattering angle. Thus, a large number of overloads within this region may render a data set useless. The second most common reason for spot exclusion is spot overlap. Overlapping of spots will cause DENZO to ignore pixels, which occupy spot regions from neighboring spots, to prevent overlapping spots from convoluting the data set. The third most common reason for spot exclusion is dead pixels. Some pixels will simply have no measurement recorded, and will not be used if flagged as such.

During integration HKL2000 monitors the changes in detector distance, mosaicity, χ^2 values, Crystal parameters, and Cell value variations for every frame. Changes in χ^2 values reflect an error estimate, instability in the crystal slippage, icing problems, and inconsistent ~~G~~ goniostat movement. Occasionally the refinement can be unstable because of a high correlation

among some parameters. High correlation makes it possible for the errors in one parameter to compensate partially for the errors in other parameters.

There are other, often subtle, ways in which errors in predicting spot positions can lead to serious integration errors. Errors in the prediction of spot positions also affect the statistical error (precision) of the summed intensities in reference to the initial indexing and formulation of the orientation matrix from which these quantities are derived.

Scaling

Due to the lack of a single file containing the un-scaled integrated intensities from HKL2000, for use in 3DSCALE, no option remained but to use the scaling program contained within HKL2000 entitled SCALEPACK. The SCALEPACK error model uses expected or predicted error without taking observed variations into account. Manipulating the parameters which compose the error model in relation to each other should reduce the inherent bias associated with spot intensities below average level within the data set (46). Error correction via modeling is shared by each data reduction program studied in this work. The large number of components contributing to the correction factor applied to each batch of images during SCALEPACK processing is monitored by a goodness of fit term used as a means of monitoring the statistical influences of Bayesian scaling termed χ^2. Because a modeling system is used errors involved with the formulation of correction factors may be large. Of course ideal data would need no correction but in the real world case used in this study manual adjustment to the error model were necessary for Dr. Fu's successful phasing trial.

The scaling portion of the HKL2000 GUI is rather straightforward. However if the results are not favorable, optimizing the output will require an advanced user with an understanding of

how the program functions to make the necessary adjustments. At the conclusion of integration, and in the case of anomalous SAD data processing the following options should be highlighted within the scaling window; Scale Restrain – to limit the amount of variance between the scale factors between batches of images, B Restrain – limits the B factors from consecutive batches, Anomalous –merges +++ and --- reflections separately (i.e. do not enforce Freidel's law), Ignore Overloads – ignores saturated reflections, and Write Rejection File – stores the reflections which exceed the HKL2000 criteria for rejections (Figure 4.1.10; I).

An additional consideration during scaling is Global Refinement. HKL does recognize that the use of a single image to refine the unit cell parameters may be imprecise if the predicted peak positions do not correspond well to the observed positions. The concept of Global Refinement is shared amongst each program studied in this work. This is a clean up procedure in which a separate refinement is conducted for each image using set unit cell parameters for the entire data set. This will remove errors inherent to a batch of images in which the unit cell, mosaicity and orientation angle were poorly determined. The importance of this action is imperative to determining which reflections are partial or full. The difference being all programs except HKL2000 includes this type of processing at the conclusion of integration. The recommended choice of "Small Slippage with a Imperfect Goniostat" is advised in the HKL

2000 manual and was used in this study (Figure 4.1.9; II).

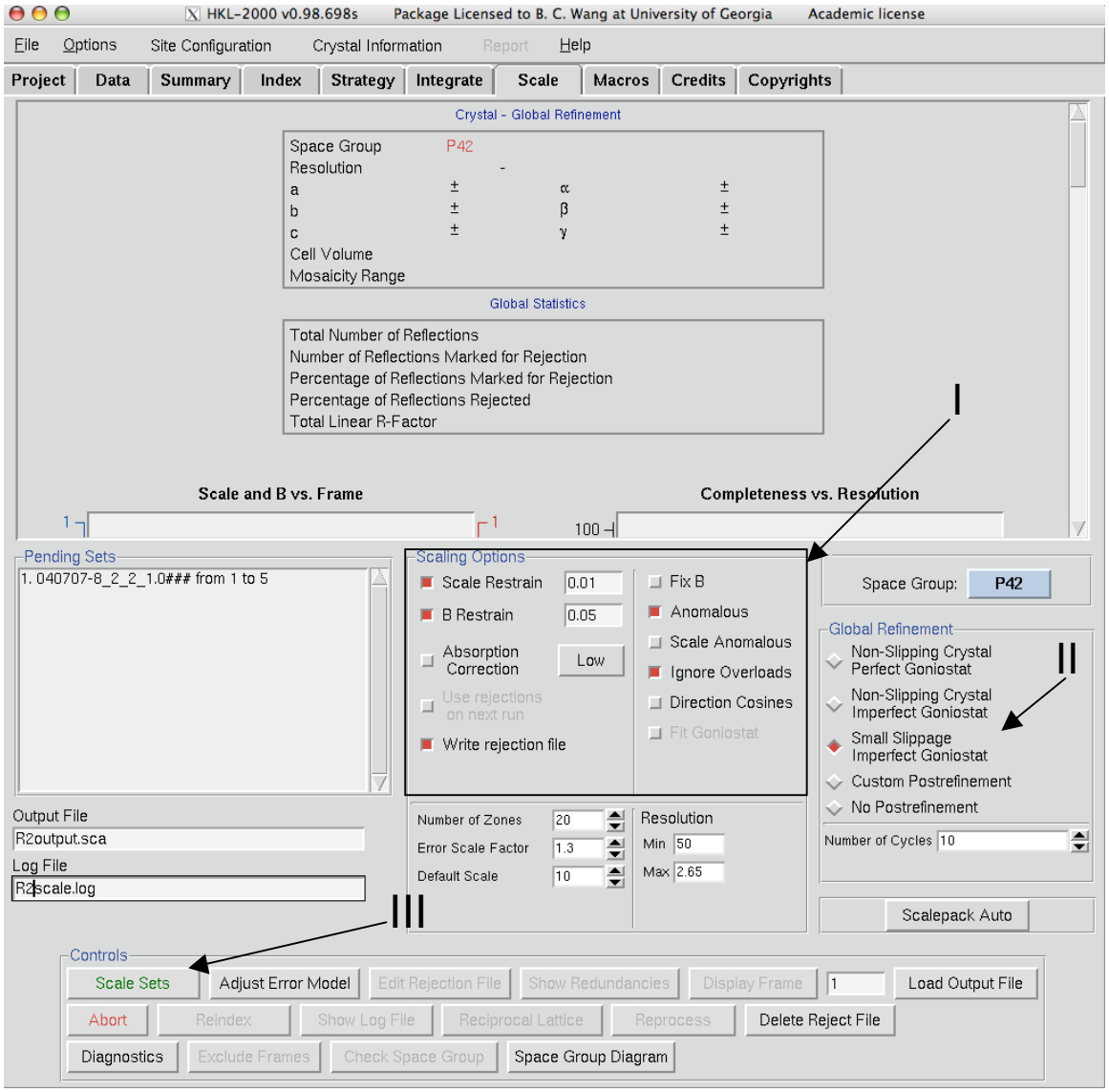

Figure 4.1.9: HKL2000 Scaling interface: (I) Options used when scaling S-SAD data sets, (II) Global Refinement factors which increase the accuracy of the scaling algorithm, and (III) Scale Sets , used for initiating scaling.

The scaling process is executed by selecting the Scale Sets tab (Figure 4.1.10, III) at the bottom of the page. At the conclusion of the initial scaling process the option Write rejection file should be changed to Use the rejection file on next run, instead of creating a new rejection file. Changing this scaling option and reinitiating the scaling routine will remove the reflections, which were initially rejected by HKL2000 and rescales the data. It is also good practice to

change the name of the scaling file and scaling log output from HKL2000 for later comparison as it will be overwritten.

This data was collected and initially processed at SERCAT and at the University of Georgia in 2007 yielding reasonable individual scaling results, but poor results when the data were merged using the HKL2000 GUI. The individual tracing results for the R1 and R2 data sets using the SGXPro "Novel Structure Solution" were inconsistent with a viable structural solution as well. The issues concerning processing of this data were discussed in an earlier section. My original efforts using HKL2000 for processing each of the R1 and R2 360° data sets and the 720° R1-R2 data set, were likewise unsuccessful. The scaling and tracing results of this effort are as follows (JTS-1).

R-factors -

$$R_{merge} = \sum_{h} \sum_{i} \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

R_{merge} is the R-factor used by HKL2000 to relate differences in symmetry related reflections. Thus R_{merge} is a measure of the accuracy of the data. The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection.

R1- Scaling Statistics using SCALEPACK

R1 – HKL2000

Summary of reflections intensities and **R-factors by shells**

Shell limit	Lower Angstrom	Upper Angstrom	Average I	Average error	Norm. stat.	Linear Chi**2	Square R-fac	R-fac
50.00	7.19		3277.3	31.7	11.0	1.834	0.028	0.039
7.19	5.71		823.5	8.7	4.9	1.551	0.031	0.034
5.71	4.99		1073.2	10.9	6.2	1.851	0.033	0.032
4.99	4.53		1879.8	18.2	9.4	1.587	0.032	0.037
4.53	4.21		1374.2	13.5	7.7	1.666	0.035	0.038
4.21	3.96		1235.0	13.4	7.9	1.680	0.037	0.040
3.96	3.76		1099.6	11.9	7.5	1.761	0.041	0.044
3.76	3.60		825.6	10.1	7.0	1.794	0.046	0.050
3.60	3.46		706.9	9.3	6.7	1.646	0.049	0.050
3.46	3.34		602.3	8.7	6.6	1.594	0.051	0.051
3.34	3.23		386.6	6.8	5.8	1.421	0.061	0.062
3.23	3.14		344.0	6.8	5.9	1.277	0.062	0.063
3.14	3.06		291.8	6.4	5.7	1.102	0.067	0.063
3.06	2.98		233.6	6.0	5.5	1.020	0.077	0.074
2.98	2.92		176.8	6.0	5.7	1.157	0.100	0.095
2.92	2.85		187.4	6.1	5.8	1.003	0.092	0.087
2.85	2.80		129.9	6.0	5.8	0.954	0.120	0.116
2.80	2.74		110.6	6.0	5.8	0.969	0.132	0.123
2.74	2.70		100.3	6.0	5.9	1.107	0.143	0.138
2.70	2.65		81.4	6.5	6.4	1.171	0.169	0.163
All reflections			761.4	10.1	6.7	1.424	**0.042**	0.040

I/sigI = **75.39(12.5)**

I/Sigma in resolution shells: **Completeness**

Lower limit	Upper limit	% of of reflections with I / Sigma less than 0	1	2	3	5	10	20	>20	total
50.00	7.19	0.5	1.0	1.0	1.0	1.6	2.6	3.1	95.9	99.0
All hkl		0.6	2.7	5.1	7.2	10.4	16.4	26.5	73.3	**99.8**

Average **Redundancy Per Shell**

Lower limit	Upper limit	
50.00	7.19	13.4
7.19	5.71	14.3
5.71	4.99	14.6
4.99	4.53	14.5
4.53	4.21	14.6
4.21	3.96	14.6
3.96	3.76	14.5
3.76	3.60	14.4
3.60	3.46	14.4
3.46	3.34	14.2
3.34	3.23	14.1
3.23	3.14	14.0
3.14	3.06	14.0
3.06	2.98	14.2
2.98	2.92	13.4
2.92	2.85	13.8
2.85	2.80	13.2
2.80	2.74	12.9
2.74	2.70	12.5
2.70	2.65	10.5
All hkl		13.8

Total Reflections used: 114871

Table 4.1.2: Scaling results from HKL2000 GUI processing: R1 processing yields acceptable R-factor results for continued processing.

R1- Automated tracing results using SGXPro

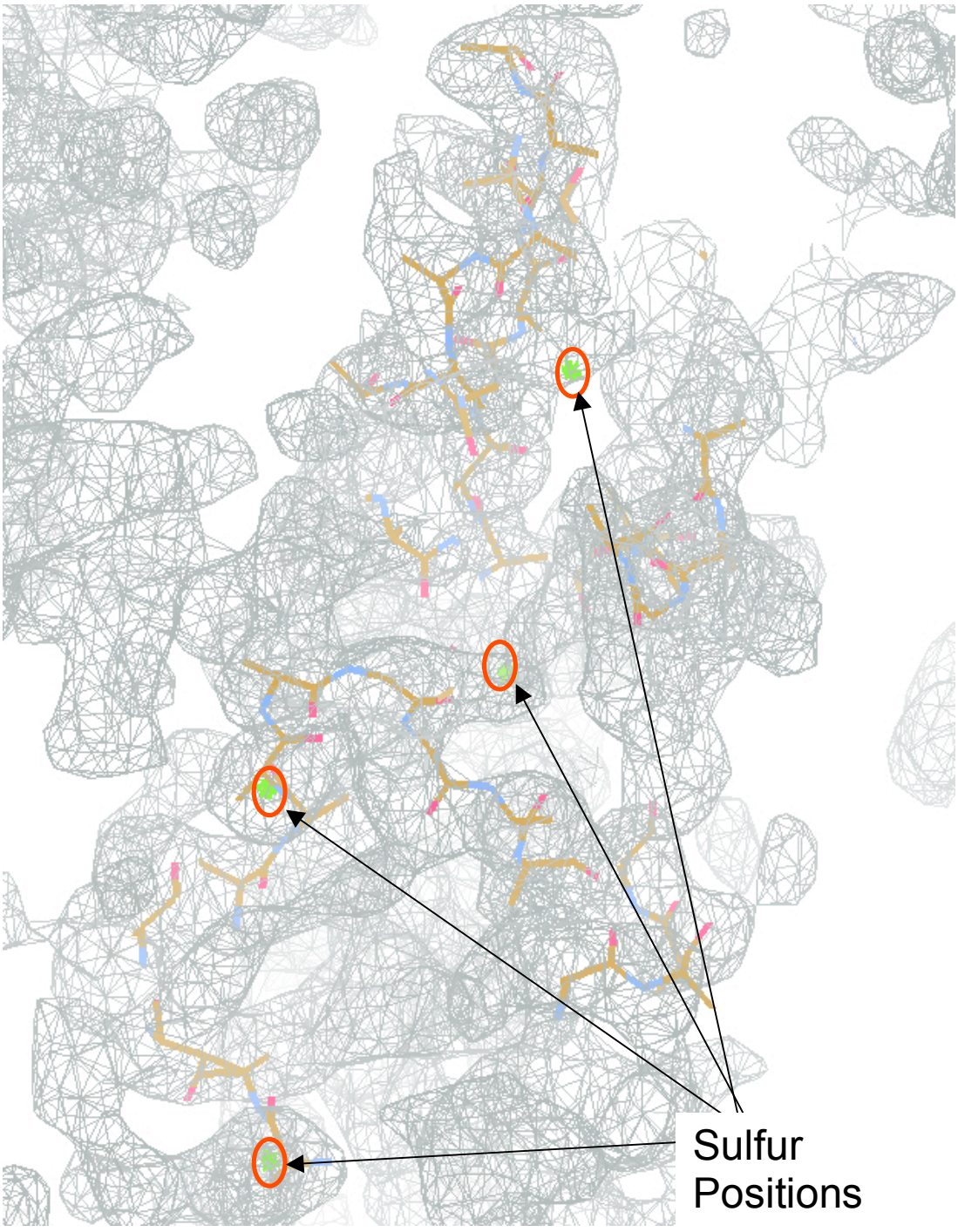

Figure 4.1.10: JTS-1 Tracing result from HKL2000 processing of R1 data set: All four Sulfur positions were correctly identified in Red, with 52% of the total amino acids traced. The trace from Resolve did not agree with the refined model.

R1- Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------  -----------
 2      49       6      12 9 9  /Aug_HKL/sgx/t4/zzsgxSol_2.pdb
Heavy atoms:                    /AUG HKL/sgx/t4/zzsgxSol_2_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 1 helices with 6 amino acids long
#Beta-sheets: No discernable beta sheet(s
CC_ALL/CC_WEAK    42.48/19.07
PATFOM      46.15
```

Table 4.1.3: JTS-1 R1 Tracing results from SGXPro Novel Structure Solution: R1 data set processing yielded acceptable CC-ALL/Weak and PATFOM but poor phases and tracing statistics. The RMSD values of the Sulfur position were however very accurate.

R2- Scaling Statistics using SCALEPACK

R2 – HKL2000

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
50.00	7.19		8078.2	100.3	46.0	2.888	0.038	0.050
7.19	5.71		2246.8	22.5	11.3	2.097	0.033	0.036
5.71	4.99		3012.4	28.6	14.4	2.475	0.035	0.037
4.99	4.53		5292.7	48.7	22.4	1.981	0.033	0.039
4.53	4.21		3877.7	35.7	17.9	2.238	0.036	0.040
4.21	3.96		3489.0	34.9	17.5	2.315	0.038	0.042
3.96	3.76		3082.7	30.3	16.6	2.313	0.040	0.043
3.76	3.60		2334.5	24.6	14.7	2.441	0.044	0.048
3.60	3.46		2012.0	22.2	13.7	2.367	0.045	0.047
3.46	3.34		1695.8	20.0	13.1	2.107	0.046	0.047
3.34	3.23		1092.8	14.3	10.7	2.026	0.053	0.058
3.23	3.14		985.7	13.7	10.5	1.862	0.052	0.053
3.14	3.06		827.6	12.3	9.7	1.622	0.055	0.060
3.06	2.98		693.5	11.3	9.2	1.399	0.057	0.057
2.98	2.92		507.8	10.2	8.9	1.288	0.068	0.068
2.92	2.85		533.8	10.4	9.1	1.277	0.065	0.066
2.85	2.80		406.4	9.7	8.8	0.967	0.068	0.071
2.80	2.74		327.6	9.3	8.7	0.876	0.074	0.076
2.74	2.70		301.1	9.1	8.5	0.859	0.078	0.077
2.70	2.65		252.9	9.4	8.9	0.960	0.091	0.089
All reflections			2082.1	24.2	14.2	1.830	**0.041**	0.044

I/sigI = **86.04(26.9)**

I/Sigma in resolution shells: Completeness

Lower limit	Upper limit	% of of reflections with I / Sigma less than 0	1	2	3	5	10	20	>20	total
50.00	7.19	0.0	0.0	1.0	1.0	1.0	2.1	2.6	94.8	97.4
All hkl		0.2	1.0	2.1	3.2	4.7	8.6	15.6	84.2	**99.9**

Average Redundancy Per Shell

Lower limit	Upper limit		Lower limit	Upper limit	
50.00	7.19	12.3	3.34	3.23	14.3
7.19	5.71	14.4	3.23	3.14	14.3
5.71	4.99	14.5	3.14	3.06	14.3
4.99	4.53	14.5	3.06	2.98	14.5
4.53	4.21	14.6	2.98	2.92	14.2
4.21	3.96	14.6	2.92	2.85	14.3
3.96	3.76	14.4	2.85	2.80	14.1
3.76	3.60	14.5	2.80	2.74	13.8
3.60	3.46	14.5	2.74	2.70	14.0
3.46	3.34	14.3	2.70	2.65	12.6
			All hkl		14.1

Total Reflections used: 113271

Table 4.1.4: JTS-1 R2 Scaling results from HKL2000 GUI processing: R2 processing yields acceptable R-fac results for continued processing.

R2- Automated tracing results using SGXPro

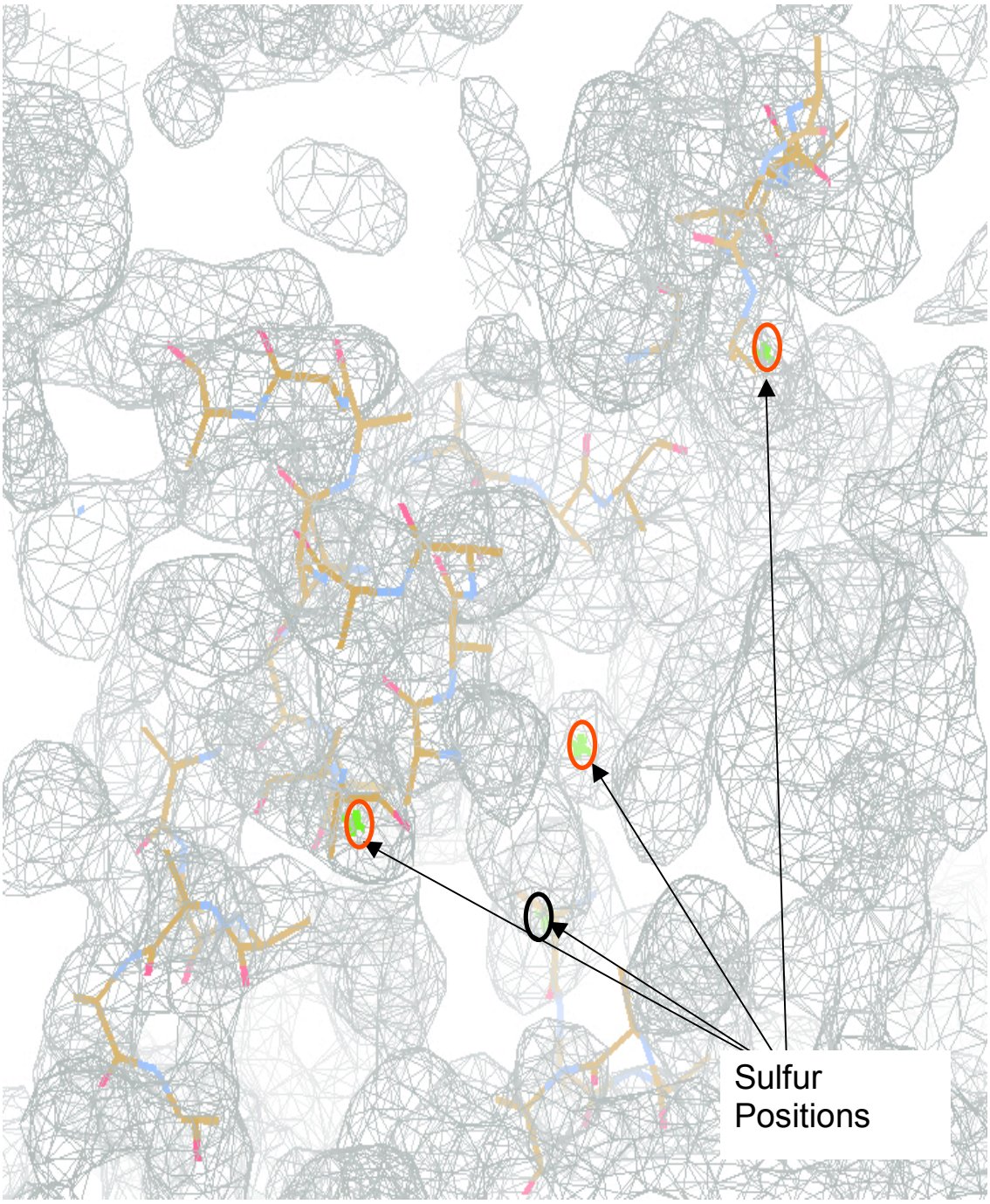

Figure 4.1.11: JTS-1 Tracing result from HKL2000 processing of R2 data set: Three of the four Sulfur positions were correctly identified in Red while an errant Sulfur position identified by SGXPro is circled in black. A total of 46% of amino acids were traced. The trace from Resolve did not agree with the refined model.

R2 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------  -----------
 2      44       5      14 10 9  /Aug_HKL/R2/t3/zzsgxSol_2.pdb
Heavy atoms:                    /Aug_HKL/R2/t3/zzsgxSol_2_ha.xyz
Sulfur atoms found:     3
#Alpha-helix: 2 helices 10 and 9 amino acids long
#Beta-sheets: 1 beta sheet 8 amino acids long
CC_ALL/CC_WEAK    33.48/16.81
PATFOM      59.89
```

Table 4.1.5: JTS-1 R2 Tracing results from SGXPro Novel Structure Solution: R2 data set processing yielded acceptable CC-ALL/Weak and PATFOM but poor phases and tracing statistics. The RMSD values were mediocre for the positions of the Sulfur atoms.

R1-R2 Merged-Scaling Statistics using SCALEPACK

R1-R2 – HKL2000

Shell Lower limit	Upper limit	Average Angstrom	Average I	Average error	Norm. stat.	Linear Chi**2	Square R-fac	R-fac
50.00	4.95	1721.0	19.4	3.5	117.666	0.350	0.419	
4.95	3.93	1544.2	16.7	3.8	106.700	0.383	0.486	
3.93	3.44	769.6	8.5	2.7	115.199	0.395	0.511	
3.44	3.12	356.0	4.4	2.0	95.658	0.439	0.642	
3.12	2.90	178.3	2.5	1.5	62.092	0.450	0.737	
2.90	2.73	95.7	1.8	1.3	38.890	0.493	0.852	
2.73	2.59	60.4	1.5	1.3	20.915	0.521	0.838	
2.59	2.48	46.2	1.5	1.3	14.137	0.516	0.834	
2.48	2.38	27.5	1.4	1.3	8.750	0.542	0.000	
2.38	2.30	18.0	1.7	1.6	5.450	0.518	0.895	
All reflections		492.6	6.1	2.1	64.548	**0.388**	0.458	

I/sigI = **80.75(10.59)**

*

Shell Lower limit	Upper limit	I/Sigma in resolution shells: Completeness % of reflections with I / Sigma less than 0	1	2	3	5	10	20	>20	total
50.00	4.95	0.2	0.4	0.4	0.4	0.5	0.5	1.1	98.4	99.5
All hkl		0.4	1.4	2.9	4.5	7.4	13.3	23.6	76.2	**99.8**

Shell Lower limit	Upper limit	Average Redundancy Per Shell
50.00	4.95	27.6
4.95	3.93	26.4
3.93	3.44	26.4
3.44	3.12	26.8
3.12	2.90	27.3
2.90	2.73	27.1
2.73	2.59	26.0
2.59	2.48	24.2
2.48	2.38	19.5
2.38	2.30	13.1
All hkl		**24.5**

Total Reflections used: 359491

Table 4.1.6: JTS-1 R1–R2 Scaling results from HKL2000 GUI processing: Processing merged R1 and R2 data sets results in erroneous scaling. R-factor values above 20% are considered errant and a clear indication of incorrect data processing.

No tracing attempts could be conducted due to the high errors associated with the merged processing of the R1-R2 merged data sets from the HKL2000 GUI (Table 4.1.6; *). The HKL2000 GUI is unable to properly scale these data sets due to the likely need of reindexing one of the data sets in respect to the other. This will be discussed in more detail during the HKL2000-Reinvestigation section. Currently the HKL2000 GUI does contain a reindexing option.

In an attempt to increase my familiarity with HKL2000 and understand how Dr. Zing-Qing Fu was able to successfully process and merge the data using HKL2000 I reprocessed the

data using my revised indexing technique (discussed earlier) and Dr. Fu's script based scaling procedure, see below. I was eventually able to surpass the quality of the original phases generated by Dr. Fu's work through the implementation of my revised indexing technique. This will be discussed later.

4.1.1 HKL2000-Reinvestigation

As mentioned in earlier sections, the HKL2000 GUI based processing of the data collected from the AF1382 was initially conducted at SERCAT, and again at the University of Georgia, was ultimately unproductive, until intervention by Dr. Zing-Quing Fu (29). The original scaling of the R1 and R2 data sets attempted to utilize data up to 2.3A resolution. The resulting linear R_{sym} values from the SCALEPACK output illustrated a significant increase between the resolution ranges of 2.73 and 2.48A. In addition a dramatic decrease in redundancy was also noted at resolutions above 2.48A. These are clearly indicative questionable data quality within these resolution ranges. Dr. Fu noted these discrepancies in the original processing and eventually determined the optimal resolution, which generated the best phases from the automated structure solution routine with SGXPro to be 2.65A. Limiting the data to this resolution combats the errant R_{merge} values seen from his original results.

I consulted with Dr. Fu concerning his indexing procedures which were identical to those reviewed in the HKL2000 manual (77). I employed my indexing procedure, mentioned earlier in section 3.1. Indexing, in addition to consulting the HKL2000 manual for several tools which improved indexing and ultimately the phases and traced map.

Lattice	Metric tensor distortion index	Best cell (symmetrized) Best cell(without symmetry restrains)					
(1)primitive tetragonal	**0.99%**	51.70	52.97	41.02	90.09	89.80	88.05
		52.33	52.33	41.02	90.00	90.00	90.00
(2)primitive tetragonal	**0.17%**	53.66	53.43	41.04	89.99	89.72	89.84
		53.54	53.54	41.04	90.00	90.00	90.00
(3)primitive tetragonal	**0.05%**	53.46	53.41	41.03	89.96	89.96	89.91
		53.43	53.43	41.03	90.00	90.00	90.00

Table 4.1.7: Indexing results: (1) Dr. Fu's R2 data sets final indexing results prior to integration (R1 was not available), (2) and (3) Improved indexing results used in the processing I conducted.

In an effort to improve indexing, I considered the Reference Zone (Figure 4.1.12;I,II) values which are assigned at random during the initial stages of indexing. According to the HKL2000 manual, if the Crystal Rotation Y: values fall between 45-135° degrees the user needs to choose a Reference Zone having Crystal Rotation Y values outside this range. This is necessary because this range can yield irregular correlations between the refinable parameters thereby yielding erroneous or suboptimal results during refinement of the unit cell and detector parameters.

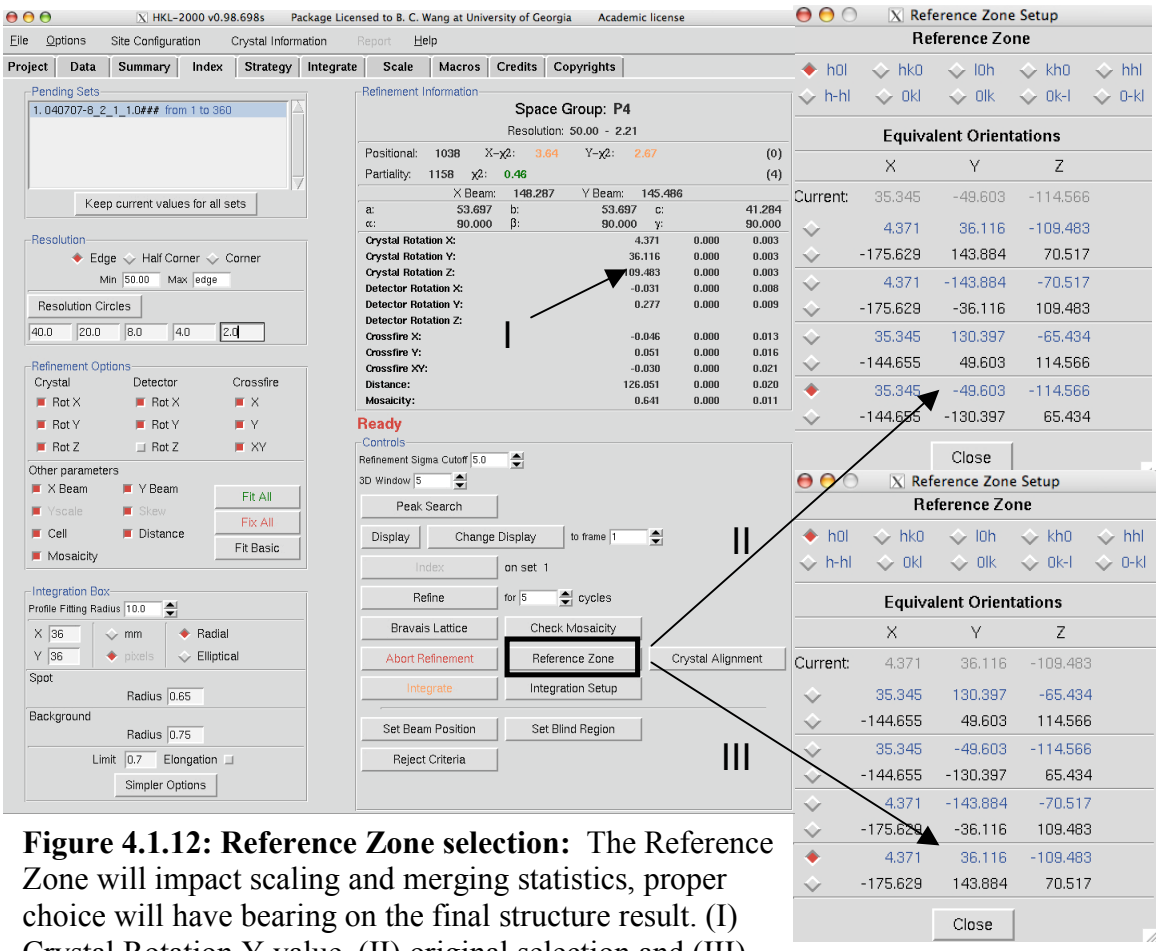

Figure 4.1.12: Reference Zone selection: The Reference Zone will impact scaling and merging statistics, proper choice will have bearing on the final structure result. (I) Crystal Rotation Y value, (II) original selection and (III) newly chosen selection

It is important to note that the R1 and R2 data sets were not collected simultaneously. The crystal was removed from the goniostat after the R1 data set was collected. The R2 data set was collected some time later after remounting the crystal. Therefore, the Reference Zone values between the data sets will not be identical. I found that Reference Zone optimization should be conducted after identifying the proper Bravais Lattice, but before the final refinement process. It is important to note that if the refinement process "explodes", which corresponds to χ^2 values becoming red, the program should be restarted and the process repeated. If the Reference Zones Rot Y values remain within the 45-135° range, merging the two data sets is still possible but could be very problematic (77).

Lastly, I chose to mimic the spot and box sizes chosen by Dr. Fu to determine if my indexing procedure, along with the proper choice of Reference Zone, had any effect on the overall solution generated.

When scaling multiple data sets using the HKL2000 GUI, problems may occur if the indices between the data sets differ. In this study, the R1 and R2 data sets have a $P4_2$ space group which is commonly referred to as polar. A polar space group does not contain symmetry operators that fix the directional origin of the until cell. The SCALEPACK scaling routine within the HKL2000 GUI is known to have problems merging two data sets with no fixed origin. During my attempts at reproducing Dr. Fu's results, I concentrated on improving the indexing and corresponding integration while retaining the use of the proven scaling scripts used by Dr. Fu. Dr. Fu designed these scaling scripts for the purpose of command line execution of the SCALEPACK algorithm within HKL2000. Within the scripts, Dr. Fu accounted for the necessary reindexing of the data to combat the errors associated with the polar characteristics of this data. During our discussions concerning the creation of these scripts, Dr. Fu stressed that a large amount of time and effort were necessary to fully optimize the various error ranges and settings within the scripts. His efforts far exceeded the basic guidelines covered in the HKL 2000 manual. My special thanks to Dr. Fu for his time, patience, and tutoring involving my instruction for future creation of such scripts.

Command line based scaling of HKL2000 integrated .x files from this study required two scaling scripts and thusly is a two-step process. The initial script (Script-1) scales the data while allowing for the creation of an error rejection file. The second script (Script-2) reads and removes those reflections noted in the error rejection file and rescales the data a second time. Dr. Fu's reindexing efforts achieved the realignment of the unit cell axes between the two data sets.

It is necessary to perform this function using script scaling as the HKL2000 GUI does not contain a reindexing option. The output files from the execution of Script-2 were saved within the same directory containing the results from Script-1 but it is good practice to change the names of the files to prevent overwriting results. Execution of the scaling scripts are accomplished using the following command:

>scalepack < *scriptname*.in > *outputname*.log

Scaling script-1 developed for HKL2000 scaling

```
print user interface
scalepack log file '/Users/bcllab1/Desktop/with-lorentz-ugadefSite-R1_R2/REINDEX-
Project2a/tucker1.log'      !Output log file from scaling
resolution 50.00 2.30
number of zones 10                  !Number of resolution shells for statistics
estimated error                     !Estimated error for each resolution shell
0.05 0.05 0.05 0.05 0.05
0.05 0.05 0.05 0.05 0.05
error scale factor 1.3              !Multiplicative factor applied to input σ
default scale 10                    !Reduces the output intensities by this factor
rejection probability 0.0001        !Expected outlier fraction in the data
reference film 7                    !Frame or batch used as reference for scaling
                                      and B refinement
scale restrain 0.01                 !Scale and b factors between adjacent frames
                                      cannot vary by more than this value (1%)
Absorption zo3                      !Not in Manual
Lorentz                             !Not in Manual
space group P42                     !Space group designation
output file '/Users/bcllab1/Desktop/
with-lorentz-ugadefSite-R1_R2/
REINDEX-Project2a/tucker1.sca'      !Output file containing h k l  and I σI
Anomalous                           !Flag for keeping Bijovets (I+ and I−) separate in output file
ignore overloads                    !Ignore the saturated reflections
intensity bins                      !Not in Manual
add partials 1 to 360 361 to 720    !Partials in derivative will be summed
fit crystal cell 1 to 360 361 to 720 !Specifies cell values fit all frames during
                                       postrefinement
fit crystal mosaicity 1 to 5 6 to 10 11 to 15   !Specifies mosaicity value fit to batches
16 to 20 21 to 25 26 to 30                        during postrefinement
31 to 35 36 to 40 41 to 45
46 to 50 51 to 55 56 to 60
61 to 65 66 to 70 71 to 75
76 to 80 81 to 85 86 to 90
91 to 95 96 to 100 101 to 105
106 to 110 111 to 115 116 to 120
121 to 125 126 to 130 131 to 135
136 to 140 141 to 145 146 to 150
151 to 155 156 to 160 161 to 165
166 to 170 171 to 175 176 to 180
181 to 185 186 to 190 191 to 195
196 to 200 201 to 205 206 to 210
211 to 215 216 to 220 221 to 225
226 to 230 231 to 235 236 to 240
241 to 245 246 to 250 251 to 255
```

```
256 to 260 261 to 265 266 to 270
271 to 275 276 to 280 281 to 285
286 to 290 291 to 295 296 to 300
301 to 305 306 to 310 311 to 315
316 to 320 321 to 325 326 to 330
331 to 335 336 to 340 341 to 345
346 to 350 351 to 355 356 to 360
361 to 365 366 to 370 371 to 375
376 to 380 381 to 385 386 to 390
391 to 395 396 to 400 401 to 405
406 to 410 411 to 415 416 to 420
421 to 425 426 to 430 431 to 435
436 to 440 441 to 445 446 to 450
451 to 455 456 to 460 461 to 465
466 to 470 471 to 475 476 to 480
481 to 485 486 to 490 491 to 495
496 to 500 501 to 505 506 to 510
511 to 515 516 to 520 521 to 525
526 to 530 531 to 535 536 to 540
541 to 545 546 to 550 551 to 555
556 to 560 561 to 565 566 to 570
571 to 575 576 to 580 581 to 585
586 to 590 591 to 595 596 to 600
601 to 605 606 to 610 611 to 615
616 to 620 621 to 625 626 to 630
631 to 635 636 to 640 641 to 645
646 to 650 651 to 655 656 to 660
661 to 665 666 to 670 671 to 675
676 to 680 681 to 685 686 to 690
691 to 695 696 to 700 701 to 705
706 to 710 711 to 715 716 to 720
fit batch rotx 1 to 360 361 to 720   !Crystal orientation of spindle axis angle
fit batch roty 1 to 360 361 to 720    will be refined for each batch or frames
                                      separately during postrefinement
postrefine 10                        !Number of cycles of postrefinement
write anomalous rejection file       !"write rejection file" is within manual and
                                      concerns placing -observations with greater than 90%  chance of being
                                      outliers in a reject file.
format denzo_ip                 ! Format of the input intensity data
sector 1 to 360                 ! These values will be substituted in for the ### identifier
FILE 1 '/Users/bcllab1/Desktop/ok/compare/Project_3a/R1_try2-SameAsAlbert/040707-
8_2_1_1_0###.x'
HKL MATRIX 0 1 0
           1 0 0
           0 0 -1
sector 1 to 360
FILE 361 '/Users/bcllab1/Desktop/ok/compare/Project_2a/R2-try3-SameasAlbert/040707-
8_2_2_1_0###.x'
;
```

Table 4.1.8: HKL2000 Scaling script: Script-1 creates a reflection rejection file and re-indexes the h,k and l indices to ensure the unit cell axis are equivalent. Script-2 performs the same function in addition to utilizing the rejection file generated from Script-1.

Script-2 is identical to Script-1 with the exception of allowing the use of the rejection file created during the scaling procedure. This rejection file highlights outliers, which can be later removed from the data to avoid the addition of error/noise that would further dilute the

anomalous signal. In this case the "*write anomalous rejection file*" is replaced by the command phrase "*@reject*".

The use of script based data processing is not a novel idea. This was previously the only means of executing the each of the programs studied in this work, prior to GUI development. The two scripts used in this study for executing the SCALEPACK algorithm within HKL2000 are more complex than any prescribed within the HKL2000 manual. It would have been impossible to generate these files with out the extensive effort and experience of a individual experienced with DENZO/SCALEPACK/HK2000 processing.

The actual significance of these bold red terms is unclear because the authors included over 500 "flags" or keywords without definitions in the original programs. After speaking with an HKL2000 author concerning the use of these terms, he was unable to offer any resolution pertaining to their use other than to refer the HKL2000 manual. These terms were not contained in either the DENZO, SCALEPACK, HKL or HKL2000 manuals. I next spoke with Dr. Fu concerning these terms. His use of the terms harkened back to the early days of data processing with Denzo, and their use was part of his various attempts to optimize the scaling of the merged data sets. To the best of Dr. Fu's knowledge, the *Air absorption zo3* term is a "ZO-flag" used in include the error effects of air absorption. Lorentz is a correction term applied to profile fitted intensities for reflections nearest the beam center, which accounts for reciprocal lattice points passing through the Ewald sphere for different spans of time such that the intensity of these spots, can be corrected by the following:

$$I = \frac{I_o}{L^*} \qquad \text{eq. 1.0}$$

I is the corrected profile fitting intensity, while I_o is the observed intensity and L^* is the Lorentz kinematical factor (note the * as it is common for the Lorentz factor to be accompanied by a

polarization correction as well, it is unclear if the original use of this term included a polarizing factor within HKL2000. Finally the *intensity bins* term is commonly used as a method of grouping the widely varying range of spot intensities.

At the completion of indexing and integration, the .x files generated from my improved indexing technique and reference zone analyses were scaled using scripts identical to those developed by Dr. Fu. The scaling results from both Dr. Fu's and my "Index Optimizing efforts" were imported into the Novel Structure Solution portion of the SGXPro processing suite for a comparative analysis of our techniques. Dr. Fu's scaling and structural results are as follows (ZQF):

R1-R2 - Scaling Statistics using SCALEPACK scripts

R1-R2 HKL2000

Shell limit	Lower Angstrom	Upper	Average I	Average error	Norm. stat.	Linear Chi**2	Square R-fac	R-fac
50.00	4.95		1947.6	22.3	4.6	1.917	0.045	0.057
4.95	3.93		1602.1	16.2	4.5	1.810	0.046	0.054
3.93	3.44		850.4	9.0	3.2	2.202	0.056	0.064
3.44	3.12		412.0	4.8	2.2	2.235	0.064	0.069
3.12	2.90		213.1	2.9	1.7	2.068	0.078	0.076
2.90	2.73		115.5	2.0	1.5	2.129	0.112	0.108
2.73	2.59		72.8	1.8	1.5	2.138	0.156	0.153
2.59	2.48		53.9	1.7	1.5	2.193	0.190	0.186
2.48	2.38		32.1	1.7	1.6	2.296	0.272	0.266
2.38	2.30		21.3	2.0	1.9	2.234	0.322	0.320
All reflections			544.8	6.6	2.4	2.106	**0.056**	0.058

I/σ_I = **82.5(10.7)**

Shell Lower limit	Upper limit	I/Sigma in resolution shells: % of reflections with I / Sigma less than 0	1	2	3	5	10	20	>20	total
50.00	4.95	0.2	0.4	0.4	0.4	0.5	0.7	1.4	98.1	99.5
All hkl		0.9	2.4	4.6	6.7	10.2	16.3	26.0	73.8	**99.8**

Shell Lower limit	Upper limit	Average Redundancy Per Shell
50.00	4.95	27.4
4.95	3.93	28.8
3.93	3.44	28.4
3.44	3.12	28.3
3.12	2.90	28.1
2.90	2.73	27.6
2.73	2.59	26.2
2.59	2.48	24.2
2.48	2.38	19.3
2.38	2.30	13.0
All hkl		**25.2**

Total number of reflections used 359491

Table 4.1.9: ZQF Merging statistics from 720 degrees of data: R1-R2 Merged processing. Dr. Fu's results: Note the increase in the linear R_{sym} within the 2.73 – 2.48A resolution shells and the decrease in redundancy beyond 2.48A, which is an apparent result of crystal decay. All other parameters were acceptable for further processing.

Dr. Fu's R1-R2 Merged Data Automated tracing results using SGXPro

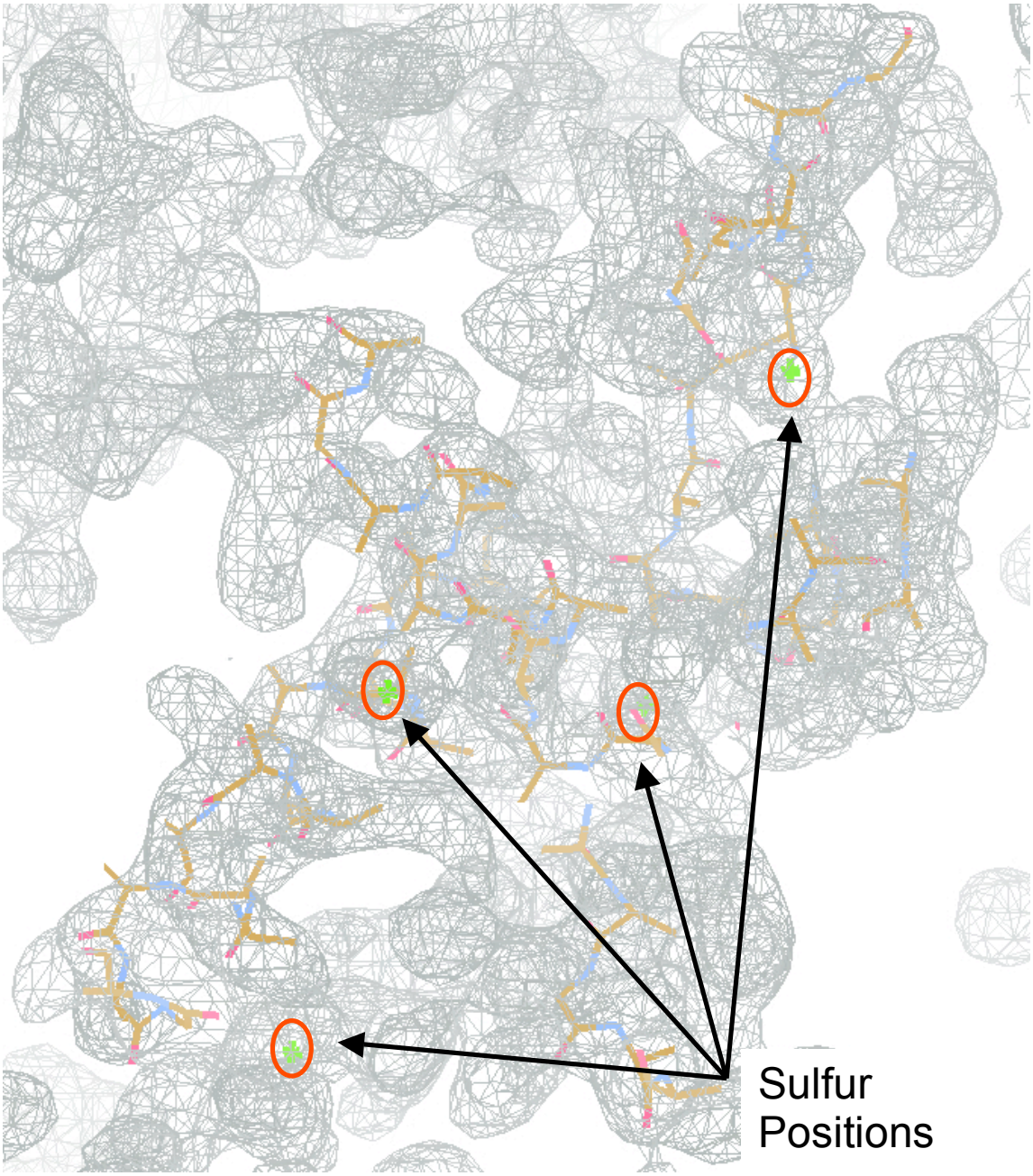

Figure 4.1.13: ZQF Tracing result from HKL2000 processing of R1-R2 Merged data set: All four Sulfur positions were correctly identified, with 59% of the total Amino Acids traced. The trace from RESOLVE agreed well with the refined model.

R1-R2 Heavy Atom/Tracing statistics using SGXPro

```
No# NumBuilt NumSegs Top3Segs  Model Built
 --- -----   ----    --------   -----------
 1     56      6     18 14 12   /Desktop/Albert_Fu/resolve/t3/zzsgxSol_2.pdb
Heavy atoms:                  /Desktop/Albert_Fu/resolve/t3/zzsgxSol_2_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 3 helices 18,14, and 4 amino acids long
#Beta-sheets: 1 beta sheet(s) 12 amino acids long
CC_ALL/CC_WEAK    42.26/23.07
PATFOM      76.41
```

Table 4.1.10: ZQF Tracing results from SGXPro Novel Structure Solution: R1-R2 Merged processing yielded excellent CC-ALL/Weak and PATFOM values with highly desirable phases and tracing statistics.

My efforts involving the improved indexing technique and Reference Zone optimization produced a higher quality scaling result and traced solution than those from Dr. Fu's work. My scaling and structural results are as follows (JTS-2):

R1-R2 - Scaling Statistics using SCALEPACK scripts

R1-R2 HKL2000

Shell limit	Lower Angstrom	Upper	Average I	Average error	Norm. stat.	Linear Chi**2	Square R-fac	R-fac
	50.00	4.95	1952.3	22.3	4.6	2.056	0.046	0.059
	4.95	3.93	1608.4	16.3	4.5	1.950	0.047	0.056
	3.93	3.44	852.9	9.0	3.1	2.386	0.058	0.067
	3.44	3.12	413.5	4.8	2.2	2.382	0.066	0.072
	3.12	2.90	213.6	2.9	1.7	2.191	0.080	0.079
	2.90	2.73	115.9	2.0	1.5	2.229	0.114	0.110
	2.73	2.59	72.9	1.8	1.4	2.235	0.158	0.157
	2.59	2.48	54.1	1.7	1.5	2.262	0.192	0.188
	2.48	2.38	32.2	1.7	1.6	2.355	0.273	0.268
	2.38	2.30	21.4	2.0	1.9	2.278	0.324	0.323
All reflections			546.4	6.6	2.4	2.224	**0.058**	0.060

I/σ_I = **82.8(10.7)**

Shell Lower limit	Upper limit	% of reflections with I / Sigma less than 0	1	2	3	5	10	20	>20	total
50.00	4.95	0.2	0.4	0.4	0.4	0.5	0.7	1.4	98.1	99.5
All hkl		0.9	2.4	4.4	6.6	10.1	16.2	26.0	73.8	**99.8**

Redundancy Per Shell Lower limit	Upper limit			Lower limit	Upper limit	
50.00	4.95	27.6		3.12	2.90	28.2
4.95	3.93	29.0		2.90	2.73	27.7
3.93	3.44	28.7		2.73	2.59	26.3
3.44	3.12	28.5		2.59	2.48	24.3
				2.48	2.38	19.4
				2.38	2.30	13.0
				All hkl		**25.4**

Total number of reflections used: 357815

Table 4.1.11: JTS-2 Merging statistics from 720 degrees of data: R1-R2 Merged processing. Note very small differences between the scaling statistics when comparing the Optimized indexing scaling results and those generated by Dr. Fu's method, namely a small increase in $\mathbf{I/\sigma_I}$ fewer reflections used.

Optimized R1-R2 Automated tracing results using SGXPro

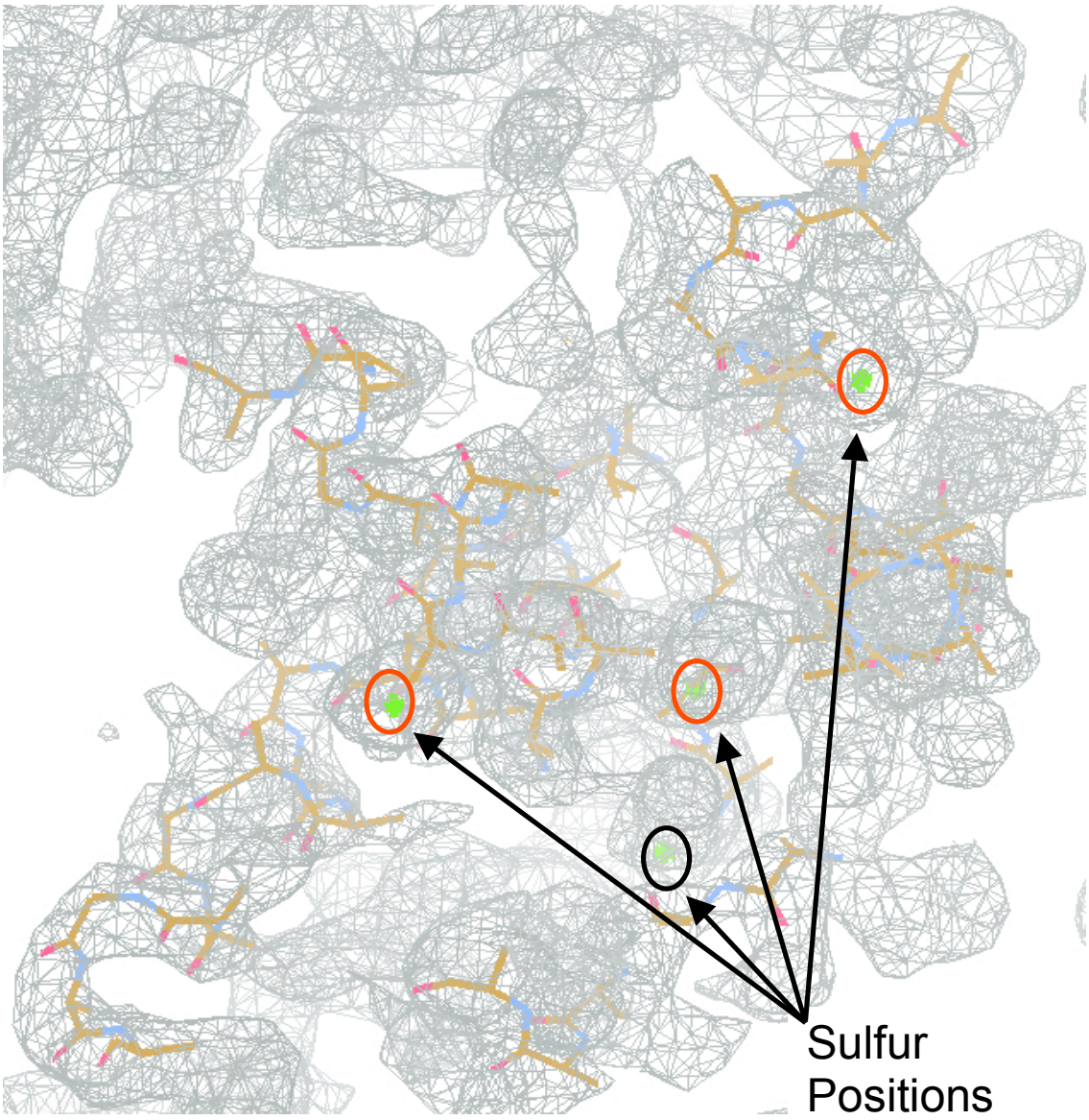

Figure 4.1.14: JTS-2 Tracing result from HKL2000 processing of R1-R2 Merged data set: Three of four Sulfur positions were correctly identified, with 73% of the total Amino Acids traced. The trace from RESOLVE agreed well with the refined model.

R1-R2 Heavy Atom/Tracing statistics using SGXPro

```
No# NumBuilt NumSegs  Top3Segs  Model Built
 ---  -----  ------  ---------
 1     69      7      18 12 11   /Aug_HKL/Project_1a/t1/zzsgxSol_ 1.pdb
Heavy atoms:                    /Aug_HKL/Project_1a/t1/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 4 helices 14,12,11,and 7 amino acids long
#Beta-sheets: 1 beta sheet(s)9 amino acids long
CC_ALL/CC_WEAK    43.9/23.13
PATFOM      85.35
```

Table 4.1.12: JTS-2 Tracing results from SGXPro Novel Structure Solution: R1-R2 Merged processing yielded excellent CC-ALL/Weak but a low PATFOM values. The phases and tracing statistics were excellent as well.

When comparing Dr. Fu's results and my own it is debatable which method produced a better traced map. Dr. Fu's results positively identified four Sulfur positions. My results located three, which corresponded with Dr. Fu's Sulfur positions and a single errant Sulfur position. In addition the tracing quality of the JTS-3 map is higher than the map produced by Dr. Fu. These results lend to the conclusion that the differences between Dr. Fu's and my own processing procedures are miniscule but deal strictly with the quality of indexing. During this process, I also developed a deeper understanding of the SCALEPACK algorithm and scaling scripts. It is my opinion that a novice user would find this scaling script impossible to devise, and an intermediate user would be hard pressed to incorporate the appropriate keywords and isolate the various numerical ranges to mimic Dr. Fu's efforts.

4.2 d*TREK

d*TREK examines diffraction data from two-dimensional X-ray detectors acquired during rotational crystallographic diffraction experiments (35). The d*TREK GUI employs both automated and manual processing schemes depending on the users understanding and ability to operate the program. For novice users, the automated path would be more beneficial. Accepting default values within d*TREK will allow the program to choose parameters it deems best for all values except identification of unit cell parameters and orientation angles. As seen in most data processing, the visual tools reflecting the quality of refinement are subjectively assessing spot alignment between predicted and observed locations. Therefore altering settings without a proficient grasp of the effects do not usually result in recognizable error flags until the conclusion of scaling, if at all. Each process within d*TREK does create a log file listing the statistical analysis of each result. For the novice user, searching through a log file can be tedious without proper instruction. Of course, auto strategy requires minimum input from the user such as image location, swing angle, beam center, goniometer direction and optional space group entry. The program will then automatically finds spots, index, refine and determine a strategy using default settings.

d*TREK is, for the sake of file organization, initiated using the *dtdisplay* within the command line interface while within the directory containing the diffraction images of interest. The first GUI window allows the user to identify the images to be studied.

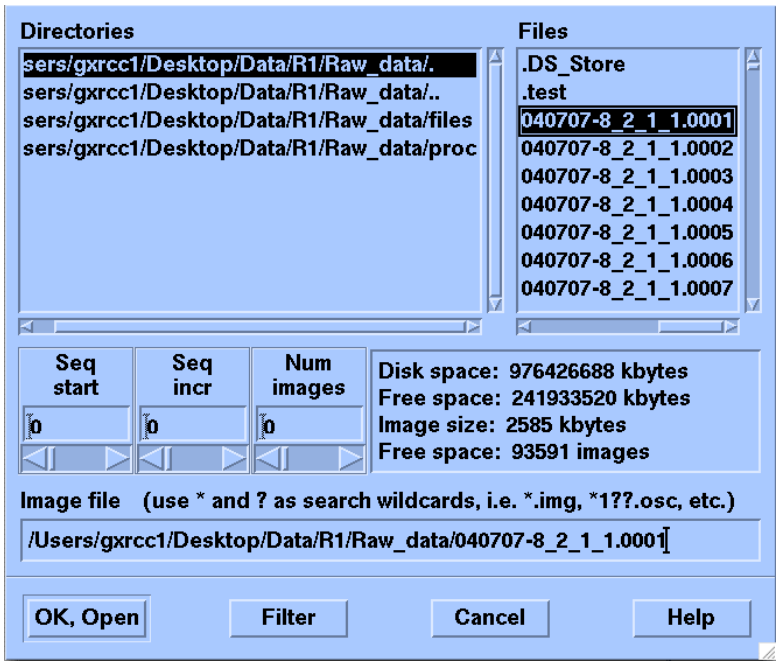

Figure 4.2.1: Initial d*TREK GUI interface: Identifies the appropriate images for processing.

After selecting the first of the images to be studied, a GUI displaying the diffraction image appears alongside experimental information collected from the header of each image. Also this window will remain open throughout processing as well as the display window highlighting those diffraction spots selected during indexing. The position of predicted spots are calculated during the prediction subroutine, and refined every four images representing the batch sizes throughout integration. The user can make a quantitative decision on the accuracy of the observed to predicted positions during processing as a check for correct indexing and refinement.

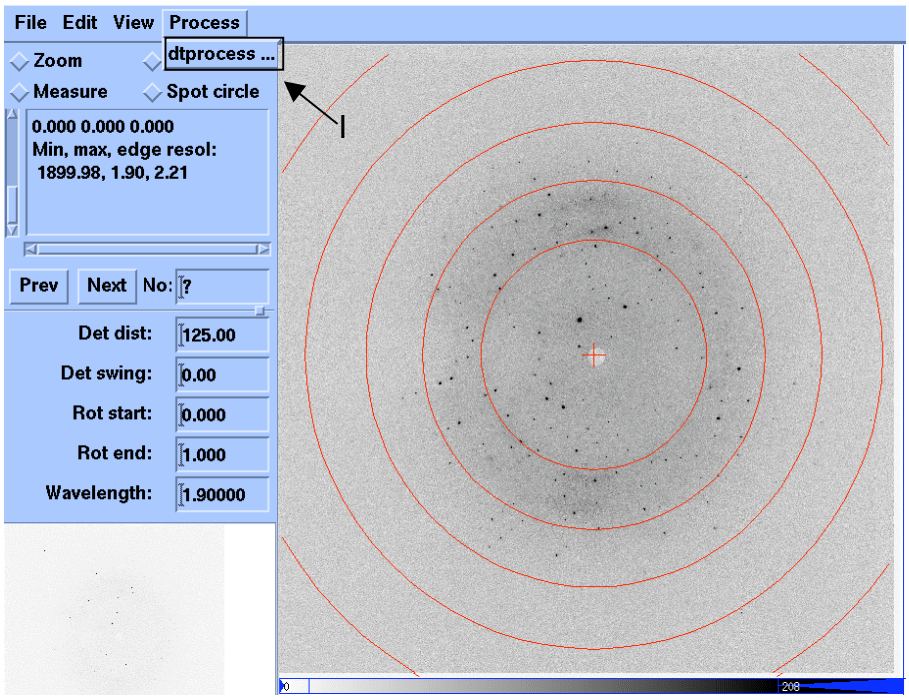

Figure 4.2.2: dtdisplay Window: This window allows the user to consider both the image characteristics as well as various parameters including the beam center, wavelength, detector distance. (I) selecting dtprocess from the Process menu tab initiates processing.

If necessary, the experimental information may be edited to ensure the program is correctly recognizing the header information. If the default information is not correct this could indicate problems with the image files.

Selecting dtprocess initiates the various sub-routines that are separately run by the dtprocess GUI. Information is recognized and transferred from one routine to another via the name of the output file generated within the processing directory. The user may wish to shield part of the image from processing or analyze a particularly concerning section of the image. These and other functions are located under the View tab. Once satisfied, the dtprocess button is selected (Figure 4.2.2; I) which initiates processing (80).

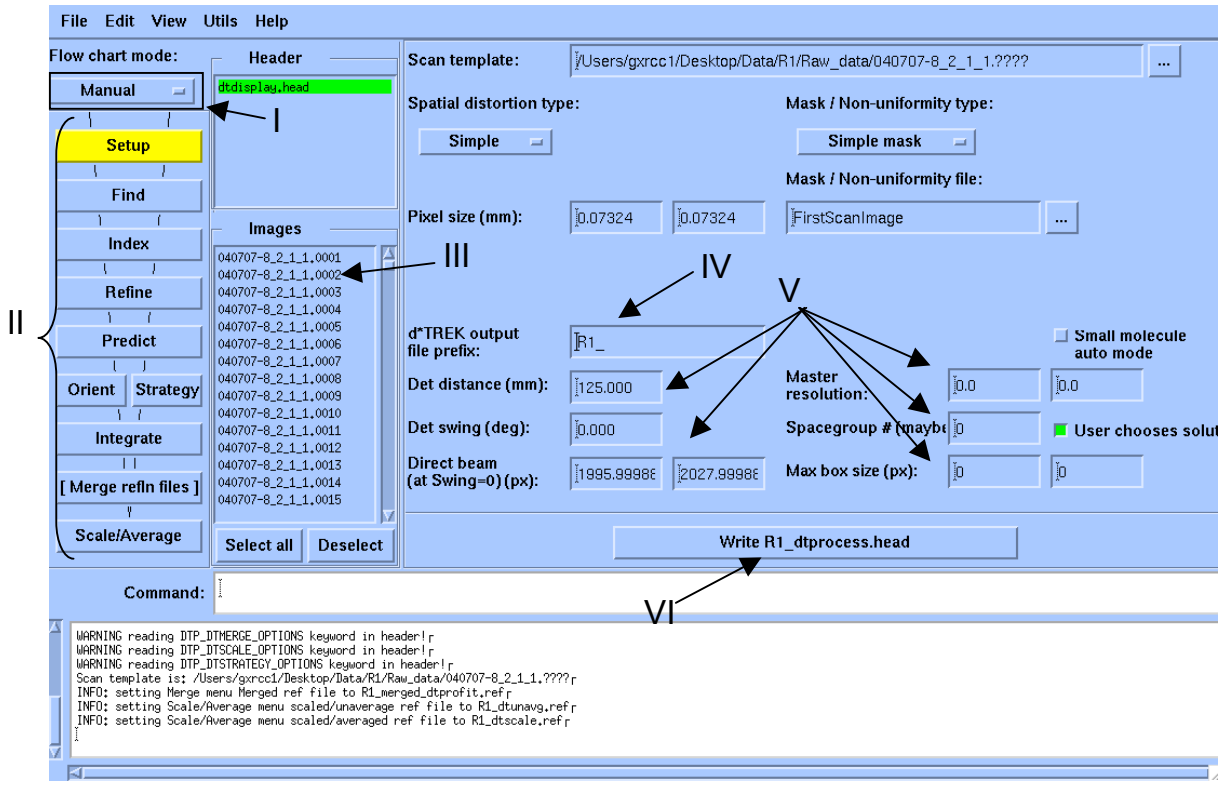

Figure 4.2.3: dtprocess Window: Main processing widow containing: (I) Choice of Manual or various Auto-Indexing flow chart mode will determine the path of data reduction chosen by the user, (II) Processing path used by d*TREK, (III) Images to be processed, (IV) Suffix identifier for output files, (V) User and image defined parameters which are in part read from the image headers and may be edited depending on prior knowledge concerning the crystal and diffraction images, (VI) Writes the above information to a file and initiates processing.

The user should ensure the sequentially labeled images have been recognized and any information read from the header is correct. If the user has prior knowledge of the experimental parameters such as space group, resolution range, beam center and goniometer direction, these should be entered or confirmed. The user may choose to allow d*TREK to automatically process the data using default values which is generally best for the novice user. I will utilize the manual option from the flow chart menu (Figure 4.2.3, I) to explain options within the program. During processing, the program will pause after the completion of every step in the flowchart allowing for manipulation of default parameters. If the information within the Settings portion of the

flowchart is considered correct the Write *prefix*_dtprocess.head (Figure 4.2.3;VI) button is selected and processing begins.

Find Spots

The spot finding portion of processing, entitled dtfind, has simple and advanced options for adjusting parameters.

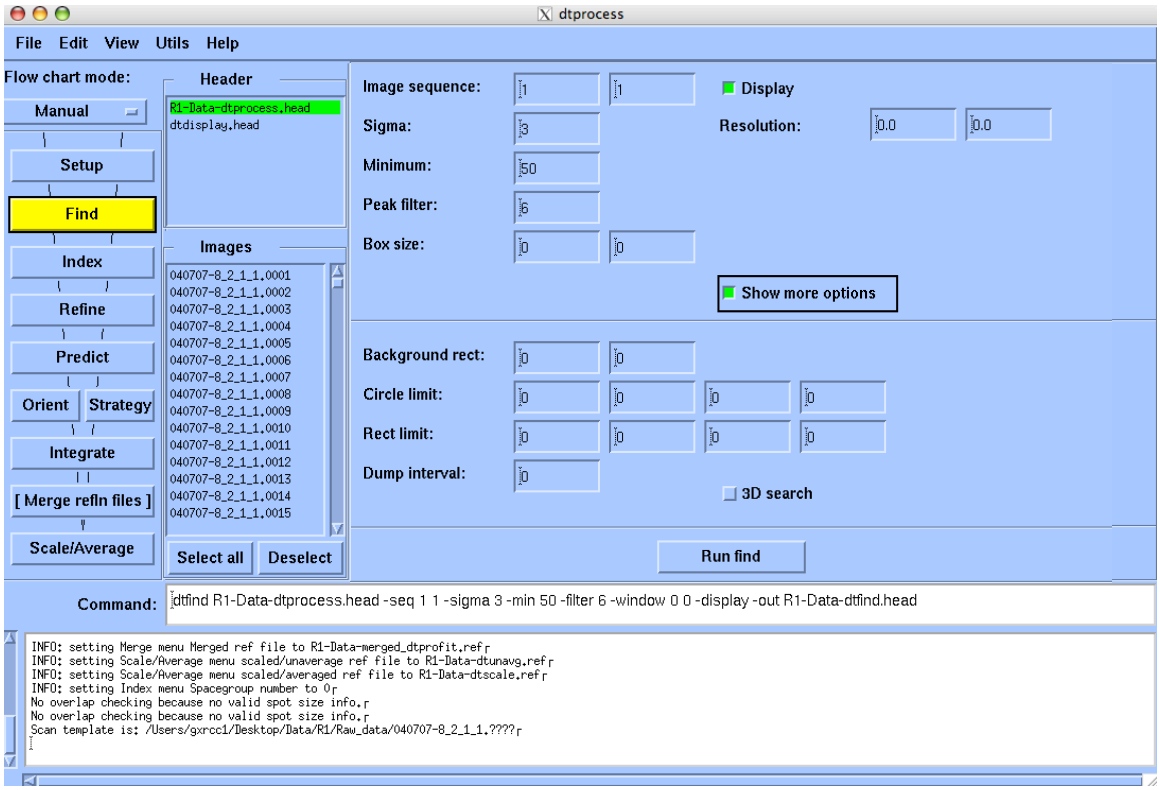

Figure 4.2.4: dtfind GUI: This portion allows the user to select the frames from the list of uploaded diffraction images, σ-value minimums-for spot cutoff, resolution ranges and the box size for both spot inclusion and background calculations.

If there are no changes to the default values within this window, the program selects the values for each parameter based on defaults. Although the image range contains only 1 value, the program automatically selects four images using the default spot criteria of 3σ, with minimum pixel strength of 20 and a peak or size filter of 6. The user does have the option of altering the number included within the image range: σ or minimum intensity values, resolution

limits/minima, Peak filter, and Box size. Lowering the peak filter will increase the number of spots chosen while an incorrectly Box size could result in inaccurately identifying the spot centroid location. There are more advanced options, which can be altered if the user is more familiar with the terms and program. These options include Background Rect, Circle limit, Rect Limit and Dump interval. None of these parameters are mentioned within the d*TREK manual (81).

Indexing

The diffraction spots determined during the dtfind subroutine are input into the indexing algorithm of d*TREK, dtindex.

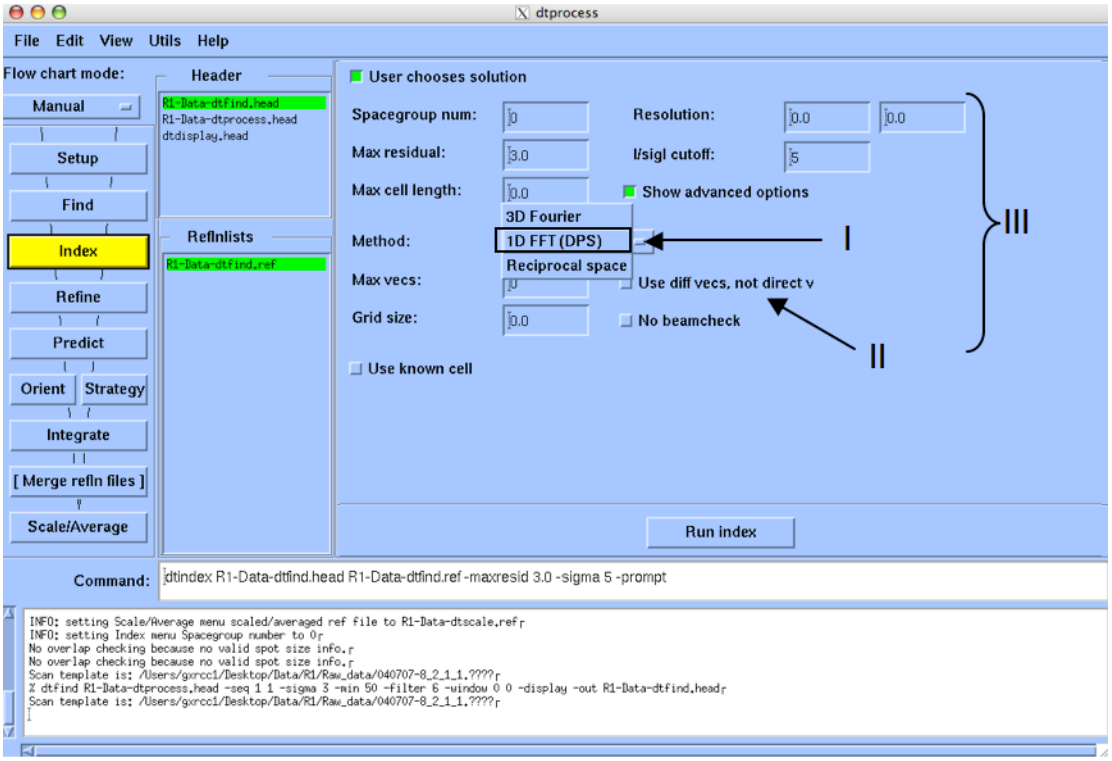

Figure 4.2.5: dtindex GUI: (I) The user has the option to chose the indexing method to be used during processing and (II) the use of difference or diffraction vectors if necessary. (III) The values within this window are normally altered during secondary rounds of data reduction as a means of improvement due to the need of knowledge pertaining to the crystal and its substructure.

The default indexing method (DPS) is shared my many of the programs covered in this work and discussed in detail in Chapter 2, section 2.1. If a change is made to this method, additional decisions such as type of vector (diffraction or difference) and I/σ_I cutoffs will be necessary to consider as well. This is generally avoided, as the 1 dimensional FFT method is the predecessor to both 3 dimensional Fourier and Reciprocal space indexing. There are a few more advanced options such as beam checking and unit cell transformations but these will usually not be exploited by an average user and can result in mis-indexing.

After selecting Run Index, the user will be questioned twice concerning the choice of lattice and unit cell parameter solutions.

Soln num	LeastSq resid(%)	Spgrp num*	Cent type	Bravais type Cell volume	a alpha	b beta	c gamma
7	0.611	75	P	tetragonal 115985	53.228 90.000	53.228 90.000	40.938 90.000
9	0.345	21	C	orthorhombic 231968	75.130 90.000	75.421 90.000	40.938 90.000
11	0.558	16	P	orthorhombic 115985	40.938 90.000	53.156 90.000	53.299 90.000
12	0.335	5	C	monoclinic 232587	75.620 90.000	75.132 90.112	40.938 90.000
13	0.240	3	P	monoclinic 115984	53.156 90.000	40.938 90.221	53.299 90.000
14	0.000	1	P	triclinic 115983	40.938 90.221	53.156 90.079	53.299 90.112

Num	Integer residual	a alpha	b beta	c gamma	Rot1	Rot2	Rot3
1	0.001	53.228 90.000	53.228 90.000	40.938 90.000	116.726	53.773	5.880
2	0.001	53.228 90.000	53.228 90.000	40.938 90.000	-116.726	-53.773	-174.120
3	0.001	53.228 90.000	53.228 90.000	40.938 90.000	-63.274	53.773	5.880
4	0.001	53.228 90.000	53.228 90.000	40.938 90.000	63.274	-53.773	-174.120
5	0.001	53.228 90.000	53.228 90.000	40.938 90.000	108.236	-31.861	-26.091
6	0.001	53.228 90.000	53.228 90.000	40.938 90.000	-108.236	31.861	153.909
7	0.001	53.228 90.000	53.228 90.000	40.938 90.000	-71.764	-31.861	-26.091

Table 4.2.1: Excerpt from indexing log files: The user is left to choose (I) the Bravis type and (II) the unit cell angles and orientation angles to use for the remainder of data reduction.

The first choice is considered the best solution for the Laude group from a list of probable solutions, the Least Squares Residual column represents the irregularity between the calculated cell and lowest symmetry triclinic cell. The lower this value the better, however d*TREK list solution based on highest symmetry and the first solution is likely to contain the correct space group for the data. Recall, the option does exist to enter the space group number if it is known in the Settings portion of processing. If the indexing routine does not list an acceptable lattice, there are a few options for the user. The first and easiest is to abort the indexing and check the beam center and detector distance to ensure they were correctly read into the program. The value for the maximum residual allowed per solution only allows values for the residual less or equal to the maximum value to be displayed. Second, the user will be prompted to choose the best solution for the cell lengths and orientation angles. To preclude the dtindex engine from attempting random cell lengths, the default indexing limits are set to the 1D DPS FFT algorithm which narrows the available choices. The chosen crystal orientation and lattice group will be recorded and passed to the next step of data reduction.

It should be noted that additional options exist for manipulating the dtindex portion of processing such as; maximum number of lattice vectors to be used, which will effect the time an accuracy of the programs results. The methodology of indexing can be changed from FFT to direct methods. This is not advised unless the user is quite certain the crystal is of very high quality and the size of the grid used in constructing the direct space cosine Fourier map. Finally, setting limitations on the cell angles and lengths should be ignored without prior knowledge concerning the unit cell of the protein in question.

Refinement

The Refinement section of d*TREK uses global macros to determine the items selected. This is similar to HKL2000 use of Fit All and Fit Basic. The prescribed method is similar to the recommended method in HKL2000(77). The advised fitting method involves not using Fit All parameters initially, but instead to choose Fit Most option. The first round of refinement log file should be discarded and repeated by selecting the first 10 images, and if desired, an additional set of images 90 degrees away from the initial 10. Next, utilize the Fit All parameters for a final refinement.

```
================================================================
Refinement residuals
rmsResid (A-1) =    0.00266
rmsResid (mm)  =    0.1158  GOOD, less than or equal to 2 pixels (0.146 mm).
rmsResid (Deg) =    0.5119  GOOD, less than 2/3rds the image rot width.
================================================================
```

Table 4.2.2: Excerpt from dtrefine results: The rmsResid measurements represent the root-mean-square residuals for the observed reflection centroids in millimeters and degrees away from predicted positions.

I have found the default values used during automated processing yield the same result regardless of the image numbers selected as described earlier. The user can visually monitor the prediction of spots to ensure the Refinement process function properly by selecting Run Predict in the next section. The predicted spots will appear just as the originally indexed spot on the dtdisplay window illustrating the diffraction patterns from the data. If this prediction seems to poorly fit the observed spots, then the Refinement process should be repeated.

Integration

d*TREK is one of three programs within this study which employees three-dimensional integration of reflections. The reflection fitting algorithm within dtintegrate first locates each diffraction spot in first two dimensions X and Y (in reference to the detector face), and then includes the third dimension from the rotation angle, ω. Data reduction programs such as

HKL2000 or MOSFLM utilize two-dimensional integration programs which sum the integrated contributions from adjacent images in a subsequent step referred to as post-refinement. Although the integration routine used in d*TREK is not specifically related to the Kabsch integration techniques (55) used in XDS or PROTEUM2, all three programs collect the intensities which may span over several images in ω contribute to a single reflection. These multiple reflections are then collected and integrated, as a full reflection(82). There are further consequences of 2D integration compared to the 3 dimensional counterparts. The most pertinent being proper accuracy involving the location reflection centroids in both X and Y without incorporating partial peak effects on true peak position. Treating partial reflections as independent diffraction spots will cause integration engines to assigned centroid positions within each spot. This assignment may not reflect the true centroid of the entire spot, which could span over multiple frames.

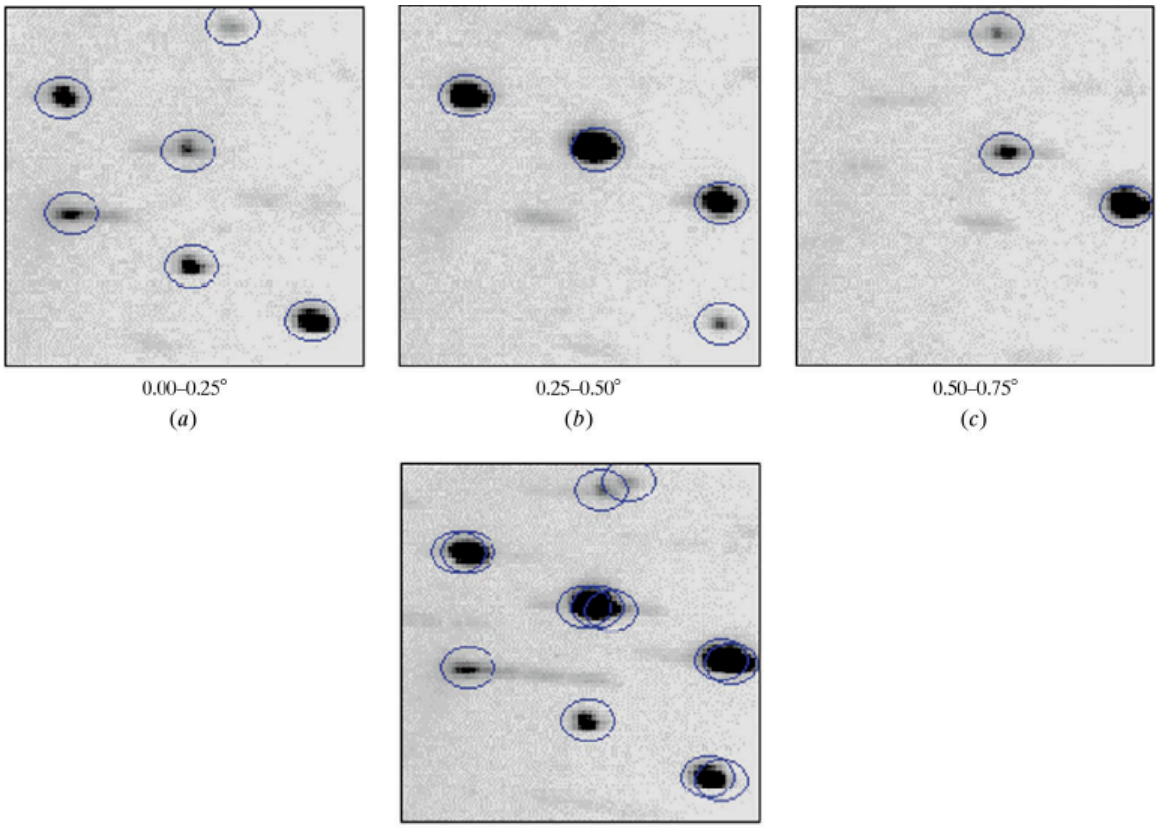

Figure 4.2.6: Importance of 3 dimensional scaling: Spot mosaicity determines the number of partials within the data set in respect to the incremental angle used during data collection. In the 2D case the angular component, ω, is neglected until a subsequent step to integration known as post-refinement. Thus during 2D integration misidentification of "true" centroids will in theory result in small loss in the overall intensity measured - *Adapted from Pflugrath, 1999*

Consequently this is one of the primary reasons I believe data which contains weak anomalous signals should be treated with three-dimensional integration tools due to the possibility of signal loss inherent to improper centroid identification.

A unique aspect of d*TREK is the lack of predetermined spot selection size and shape. Unlike HKL2000, which chooses a standard profile for spot/background differentiation, each spot is defined independently despite its conformation. The background calculation spans the x, y, and ω. This dimensional analysis of the spots are referred to as the shoebox defined by x and y in 2D and the third dimension through successive images in ω. This could minimize the errors

mentioned in chapter 3, section 3.1. Spot and shoebox size can be defined by the user, if necessary, and is usually 3-4 times the size of the spots in question. If left to default, the maximum box size is used by d*TREK and spot shapes are chosen automatically. The background and signal differentiation within the shoebox are the same as those used by HKL2000.

The dtintegrate GUI offers the user has the option of integrating the whole or a portion of the data set by independently choosing images. The resolution range can be truncated or left to defaults, which integrate over the entire surface of the image.

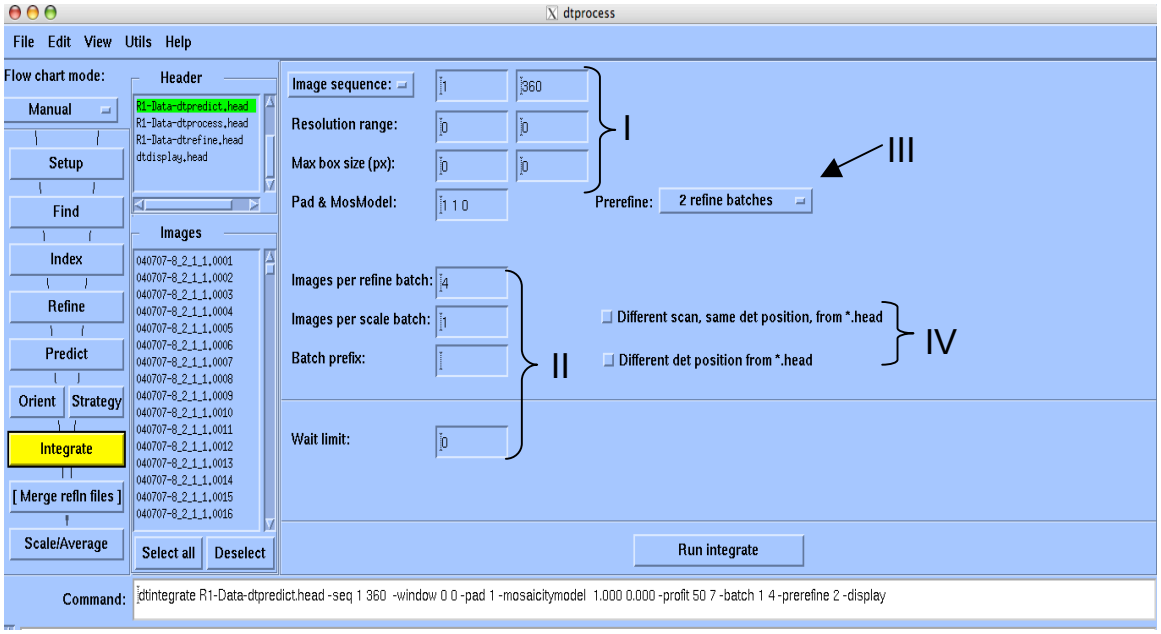

Figure 4.2.7: dtintegrate GUI: Integration options include group (I) through (IV). Group (I) should only be considered after carefully inspecting the diffraction frames Group (II) containing refinement and scaling batch information not commonly altered unless working with small molecules. Option (III) and (IV) refer to refinement batch size and if the experimenter has changed the detector positions for certain frames within a data set, these options are not discussed in the d*TREK manual and should be left to defaults.

The *Pad* and *MosModel* can assist the user in processing frames that qualify as wide or thin slice data as well as modifying the refined mosaicity. The common user seldom changes

these options. If dealing with small molecules, there are other parameters for which the default values may be insufficient; this however falls outside the focus of this study.

Refinement of image batches is simultaneously conducted during integration. A refinement routine is conducted before every batch is integrated; this is defined within the Images per refine batch option within the dtintegrate GUI. The results from each refinement cycle are monitored by RMS values contained in an auto-updating log file that is displayed within the GUI. In addition a visual inspection of the predicted versus observed spots, location can be conducted during integration of each batch of images in the dtdisplay GUI. At the end of integration a summary is displayed from which the user can ascertain the quality of integration. The d*TREK manual does offer suggestions based on typical problems within integration results. In addition d*TREK does contain a GUI titled dtplot which offers several plots which may assist the user in obtaining the reasons for problems during integration. It is important, however, to make sure to change the viewing options to 'absolute' view so that the actual intensity values are plotted versus the standard deviations. The most common solutions to integration problems involve the need for resolution restriction, spot/box choice, and refinement parameters.

Scaling

The d*TREK scaling engine was found to produce lower quality results than 3DSCALE, and thusly relegated scaling duties to 3DSCALE. The integration output file required for 3DSCALE is dtprofit.ref; however, this file must be converted to the proper format beforehand. The following command line code is used for this conversion:

```
>dtreflnmerge dtprofit.ref newdtprofit-text.ref –text
>vi newdtprofit-text.ref
```

Within an editing window the first line of the newdtprofit-text.ref file should hold four different numerical groupings. After the last of these groupings a space followed by the detector dimensions on which the data was collected should be inserted as:

8 25 1 4 3000 3000**
CRYSTAL_MOSAICITY=0.6427 0.0000 0.0000;
CRYSTAL_ORIENT_ANGLES= -63.1350 53.6476 5.6573;
CRYSTAL_SPACEGROUP= 77;
CRYSTAL_UNIT_CELL= 53.7872 53.7872 41.3292 90.0000 90.0000 90.0000;

Table 4.2.3 Excerpt from dtprofit-text.ref file: The first line of this file is edited to contain the detector identification 3000 3000 ** that is specific to the Marr 300 CCD detector used on the SERCAT 22ID line.

Following this addition the file should be saved and is ready to be recognized as a 3DSCALE input file. The scaling and structural results using 3DSCALE and SGXPro from d*TREK processing of data sets R1, R2, and R1-R2 is as follows;

R-factors -

$$R_{sym} = \frac{\sum_{h}\sum_{i}\left|I_{hi} - I_{mean}\right|}{\sum_{hi} I_{hi}}$$

R_{sym} is the R-factor chosen 3DSCALE to relate differences in symmetry related reflections. This is a measure of the accuracy of the data. The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection.

R1- Scaling Statistics using 3DSCALE

R1-d*TREK

Res.Shell		nRefObs	nRefExp	nRefCen	Compl%	Redund
37.75						
to	5.78	351	355	4887	98.87	13.76
to	4.57	701	712	9992	98.46	14.07
to	3.98	1053	1068	15150	98.60	14.17
to	3.61	1404	1421	20267	98.80	14.15
to	3.35	1755	1776	25302	98.82	14.06
to	3.14	2106	2143	30266	98.27	13.92
to	2.98	2457	2494	35146	98.52	13.76
to	2.86	2808	2855	39929	98.35	13.55
to	2.74	3159	3206	44714	98.53	13.37
to	2.65	3514	3561	49078	**98.68**	**13.15**

Res.Shell		Rsym---Shell		Rfree	nRfree	<i/sigi>--Shell		<I/sigI>--Shell		<Chi^2>
37.75										
to	5.78	0.0374	0.0374	0.0368	19	18.10	18.10	66.38	66.38	1.77
to	4.57	0.0410	0.0461	0.0446	40	17.42	16.76	64.78	63.18	1.80
to	3.98	0.0444	0.0538	0.0484	59	16.69	15.28	62.41	57.70	1.88
to	3.61	0.0492	0.0773	0.0524	77	15.97	13.78	59.73	51.67	2.09
to	3.35	0.0526	0.0951	0.0571	97	15.22	12.15	56.75	44.81	2.27
to	3.14	0.0544	0.1009	0.0592	115	14.47	10.44	53.53	37.40	2.42
to	2.98	0.0558	0.1184	0.0604	131	13.77	9.23	50.51	32.39	2.64
to	2.86	0.0568	0.1406	0.0611	144	13.12	7.96	47.57	27.04	2.91
to	2.74	0.0576	0.2008	0.0617	156	12.47	6.51	44.79	22.40	3.16
to	2.65	**0.0583**	**0.2358**	0.0622	173	11.88	5.64	**42.14**	**18.51**	3.32

<<< RAS >>>

Res.Shell		<AnoI/SigI>a --shell		<AnoI/SigI>c --shell		---Ras---- --shell	
37.75							
to	5.78	3.12	3.12	1.98	1.98	1.57	1.57
to	4.57	2.86	2.63	2.02	2.07	1.42	1.27
to	3.98	2.66	2.28	2.10	2.26	1.26	1.01
to	3.61	2.76	3.06	2.23	2.60	1.24	1.18
to	3.35	2.78	2.86	2.35	2.85	1.19	1.00
to	3.14	2.83	3.06	2.44	2.93	1.16	1.05
to	2.98	2.95	3.60	2.57	3.36	1.15	1.07
to	2.86	3.08	3.96	2.71	3.78	1.13	1.05
to	2.74	3.18	3.96	2.86	4.13	1.11	0.96
to	2.65	3.30	4.34	2.96	3.95	1.11	1.10

< Number of Reflections Used for Scaling and Output >:

49882 reflections read in

Excluded for Scaling		Excluded for Output		
0		0		on the marked-off frames
0		0		low resolution cutoffs 37.781
0		0		high resolution cutoffs 2.650
0		0		low I/SigI cutoffs -4.531
0		0		high I/SigI cutoffs 617.464
0		3507		rejected as outliers
3	3	0	0	single reflections
2306	173	0	0	randomly selected for Rfree subset

Observed	Unique	Observed	Unique	
47395	3338	46197	3514	total reflections used scaling/output (excluding lattice-center-related extinctions)

Table 4.2.4: R1 Scaling results from d*TREK processing using 3DSCALE: R1 processing yields acceptable R_{sym} results for continued processing.

R1 - Automated tracing results using SGXPro

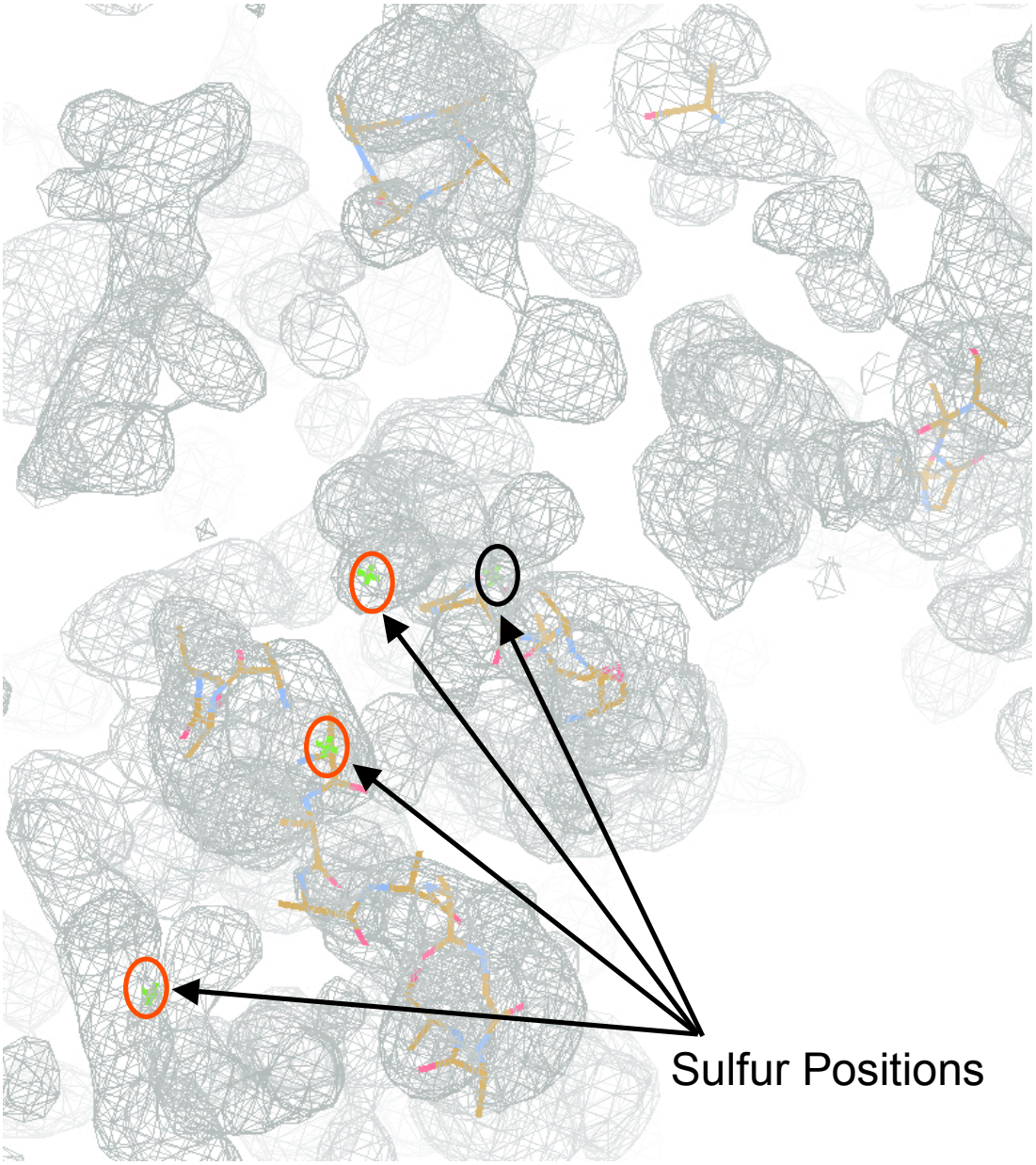

Figure 4.2.8: Tracing result from d*TREK processing of R1 data set: Three of the four Sulfur positions were correctly identified in Red while an errant Sulfur position identified by SGXPro is circled in black. A total of 48% of amino acids were traced. The trace from RESOLVE did not agree with the refined model.

R1 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------  -----------
 1      49       6      16 12 7 /Tucker_Dtrek/R1/SGXPro/t1/zzsgxSol_1.pdb
Heavy atoms:                    /Tucker_Dtrek/R1/SGXPro/t1/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 2 helices 6 and 4 amino acids long.
#Beta-sheets: no discernable beta sheet.
CC_ALL/CC_WEAK    27.8/8.1
PATFOM      44.8
```

Table 4.2.5: R1 Tracing results from SGXPro Novel Structure Solution: R1 processing yielded poor CC-ALL/Weak and PATFOM values. The phases and tracing statistics were substandard as well.

R2- Scaling Statistics using 3DSCALE

R2-d*TREK

```
Res.Shell nRefObs nRefExp nRefCen Compl% Redund
    37.85
to   5.76     350     361    4703  96.95  13.34
to   4.55     700     714    9825  98.04  13.93
to   3.97    1050    1068   14945  98.31  14.10
to   3.60    1400    1425   20076  98.25  14.17
to   3.34    1750    1775   25149  98.59  14.16
to   3.14    2100    2140   30201  98.13  14.12
to   2.98    2450    2490   35137  98.39  14.02
to   2.85    2800    2845   40005  98.42  13.92
to   2.74    3150    3196   44760  98.56  13.79
to   2.65    3506    3552   49158  98.70  13.54

Res.Shell     Rsym---Shell   Rfree    nRfree <i/sigi>--Shell <I/sigI>--Shell <Chi^2>
    37.85
to   5.76    0.0415  0.0415  0.0459     19    19.90  19.90    71.62  71.62    2.13
to   4.55    0.0411  0.0405  0.0448     41    19.68  19.47    72.64  73.65    2.15
to   3.97    0.0431  0.0480  0.0480     56    19.20  18.28    71.37  68.85    2.35
to   3.60    0.0465  0.0626  0.0509     80    18.68  17.14    69.70  64.67    2.57
to   3.34    0.0490  0.0742  0.0524    101    17.99  15.24    67.21  57.28    2.76
to   3.14    0.0511  0.0938  0.0532    112    17.23  13.35    64.29  49.68    3.02
to   2.98    0.0528  0.1100  0.0545    127    16.48  11.79    61.23  42.83    3.32
to   2.85    0.0541  0.1331  0.0552    138    15.76  10.40    58.27  37.58    3.57
to   2.74    0.0551  0.1674  0.0556    152    15.10   9.29    55.39  32.33    3.90
to   2.65    0.0558  0.1918  0.0562    173    14.52   8.26    52.49  26.82    4.18

<<< RAS >>>
Res.Shell <AnoI/SigI>a <AnoI/SigI>c  ---Ras----
    37.85      --shell      --shell     --shell
to   5.76   3.56  3.56   2.13  2.13  1.67  1.67
to   4.55   3.24  2.96   2.15  2.19  1.51  1.35
to   3.97   3.08  2.77   2.27  2.49  1.36  1.11
to   3.60   3.08  3.08   2.40  2.80  1.28  1.10
to   3.34   3.05  2.96   2.52  2.97  1.21  1.00
to   3.14   3.11  3.36   2.66  3.37  1.17  1.00
to   2.98   3.16  3.48   2.80  3.62  1.13  0.96
to   2.85   3.25  3.87   2.93  3.83  1.11  1.01
to   2.74   3.43  4.79   3.08  4.33  1.11  1.11
to   2.65   3.60  5.11   3.23  4.73  1.12  1.08

< Number of Reflections Used for Scaling and Output >:
   ------------------------------------------------
                        49796  reflections read in
   ----------------------   ----------------------
    Excluded for Scaling      Excluded for Output
           0                        0               on the marked-off frames
           0                        0               low  resolution cutoffs     37.886
           0                        0               high resolution cutoffs      2.650
           0                        0               low  I/SigI cutoffs         -3.959
           0                        0               high I/SigI cutoffs        910.650
           0                     2176               rejected as outliers
           3           3            0           0   single reflections
        2412         173            0           0   randomly selected for Rfree subset
   ======================   ======================
    Observed      Unique     Observed      Unique    total reflections used scaling/output
       47241        3330        47480        3506   (excluding lattice-center-related
extinctions)
```

Table 4.2.6: R2 Scaling results from d*TREK processing using 3DSCALE: R2 processing yields acceptable R_{sym} results for continued processing.

R2 - Automated tracing results using SGXPro

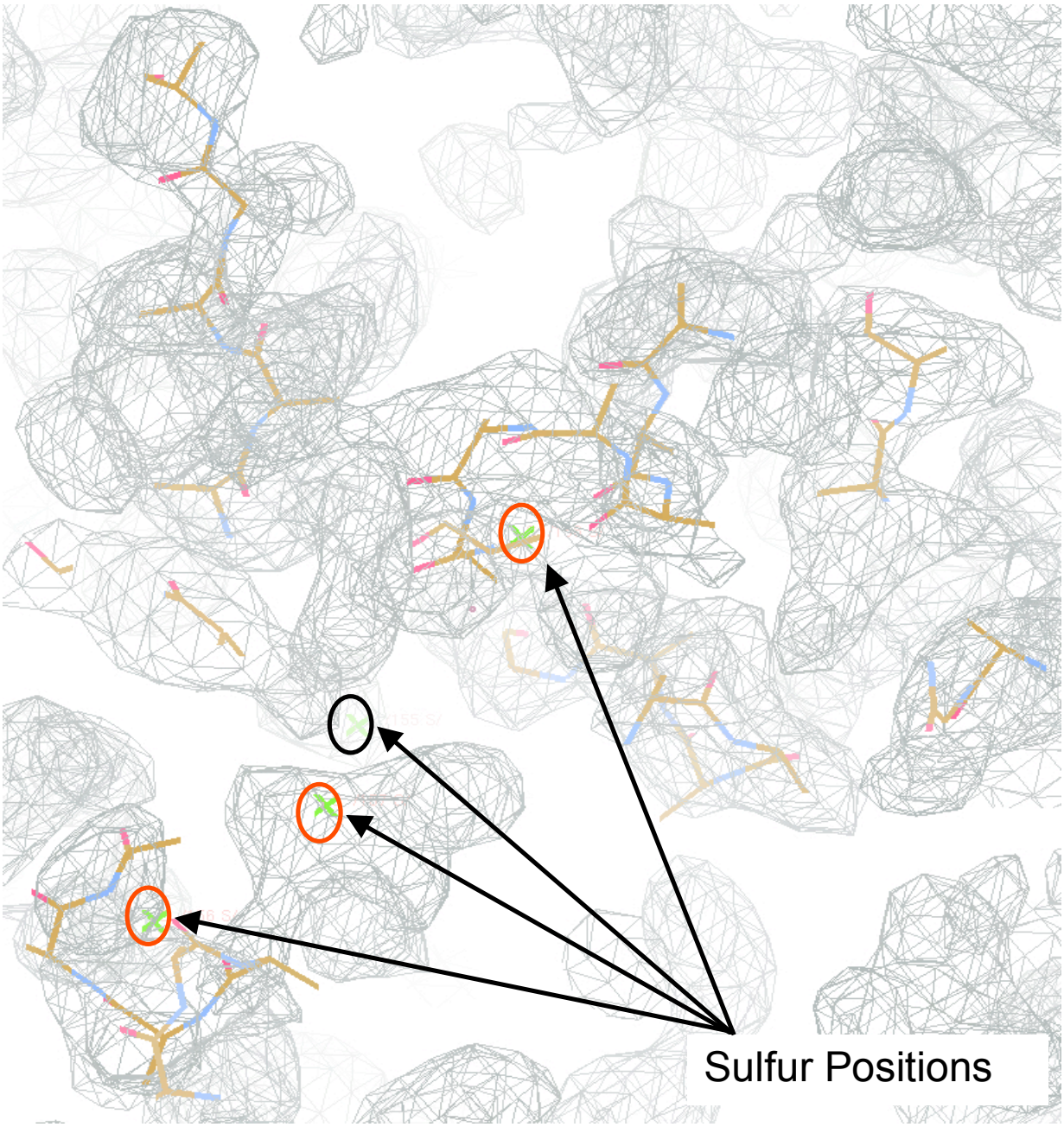

Figure 4.2.9: Tracing result from d*TREK processing of R2 data set: Three of the four Sulfur positions were correctly identified in Red, while an errant Sulfur position identified by SGXPro is circled in black. A total of 61% of the total amino acids traced. The trace from RESOLVE did not agree with the refined model.

R2 - Heavy Atom/Tracing statistics using SGXPro

No#	NumBuilt	NumSegs	Top3Segs	Model Built
1	54	9	11 8 7	/Desktop/Tucker_Dtrek/R2/SGXPro/t2/zzsgxSol_1.pdb

Heavy atoms: /Desktop/Tucker_Dtrek/R2/SGXPro/t2/zzsgxSol_1_ha.xyz

```
Sulfur atoms found:     3
#Alpha-helix: 1 helices with 6 amino acids long
#Beta-sheets: no discernable beta sheet
CC_ALL/CC_WEAK    22.3/7.19
PATFOM      31.35
```

Table 4.2.7: R2 Tracing results from SGXPro Novel Structure Solution: R2 processing yielded poor CC-ALL/Weak and PATFOM values. The phases and tracing statistics were substandard as well.

R1-R2 - Scaling Statistics using 3DSCALE

R1-R2 d*TREK

Res.Shell	nRefObs	nRefExp	nRefCen	Compl%	Redund
37.75					
to 5.77	351	355	9170	98.87	25.74
to 4.57	701	711	19165	98.59	26.97
to 3.98	1053	1064	29209	98.97	27.31
to 3.61	1404	1417	39294	99.08	27.49
to 3.35	1755	1771	49255	99.10	27.45
to 3.14	2106	2126	59146	99.06	27.34
to 2.98	2457	2477	68891	99.19	27.14
to 2.86	2808	2831	78540	99.19	26.90
to 2.74	3159	3183	88116	99.25	26.62
to 2.65	3517	3541	96815	**99.32**	**26.14**

Res.Shell	Rsym	---Shell	Rfree	nRfree	<i/sigi>	--Shell	<I/sigI>	--Shell	<Chi^2>
37.75									
to 5.77	0.0472	0.0472	0.0477	17	17.06	17.06	85.57	85.57	2.26
to 4.57	0.0478	0.0484	0.0491	38	16.71	16.40	86.13	86.69	2.22
to 3.98	0.0500	0.0556	0.0508	53	16.24	15.33	84.31	80.70	2.30
to 3.61	0.0542	0.0757	0.0535	73	15.75	14.33	82.24	76.01	2.47
to 3.35	0.0576	0.0940	0.0562	91	15.19	12.91	79.32	67.63	2.67
to 3.14	0.0598	0.1097	0.0586	108	14.56	11.33	75.87	58.66	2.91
to 2.98	0.0615	0.1263	0.0604	126	13.95	10.11	72.37	51.32	3.14
to 2.86	0.0628	0.1491	0.0614	140	13.36	8.96	68.92	44.79	3.38
to 2.74	0.0638	0.2006	0.0621	153	12.79	7.73	65.46	37.80	3.67
to 2.65	**0.0646**	**0.2273**	0.0627	172	12.28	6.85	**61.99**	**31.37**	3.90

<<< RAS >>>

Res.Shell	<AnoI/SigI>a	--shell		<AnoI/SigI>c	--shell	
37.75						
to 5.77	3.98	3.98	2.17	2.17	1.83	1.83
to 4.57	3.59	3.24	2.13	2.12	1.68	1.53
to 3.98	3.36	2.93	2.19	2.31	1.53	1.26
to 3.61	3.42	3.61	2.30	2.62	1.49	1.38
to 3.35	3.42	3.39	2.42	2.89	1.41	1.17
to 3.14	3.45	3.61	2.55	3.18	1.35	1.14
to 2.98	3.52	3.89	2.68	3.47	1.31	1.12
to 2.86	3.61	4.23	2.83	3.89	1.27	1.09
to 2.74	3.71	4.52	3.01	4.48	1.23	1.01
to 2.65	3.90	5.49	3.16	4.59	1.23	1.20

< Number of Reflections Used for Scaling and Output >:

97158 reflections read in

Excluded for Scaling		Excluded for Output		
0		0		on the marked-off frames
0		0		in the marked-off batches
0		0		low resolution cutoffs 37.886
0		0		high resolution cutoffs 2.650
0		0		low I/SigI cutoffs -4.531
0		0		high I/SigI cutoffs 910.650
0		5215		rejected as outliers
4	4	0	0	single reflections
4566	172	0	0	randomly selected for Rfree subset
Observed	Unique	Observed	Unique	total reflections used scaling/output
92588	3341	91943	3517	(excluding lattice-center-related extinctions)

Table 4.2.8: R1-R2 Scaling results from d*TREK processing using 3DSCALE: R1-R2 Merged processing yields acceptable R_{sym} results for continued processing.

R1-R2 Merged Automated tracing results using SGXPro

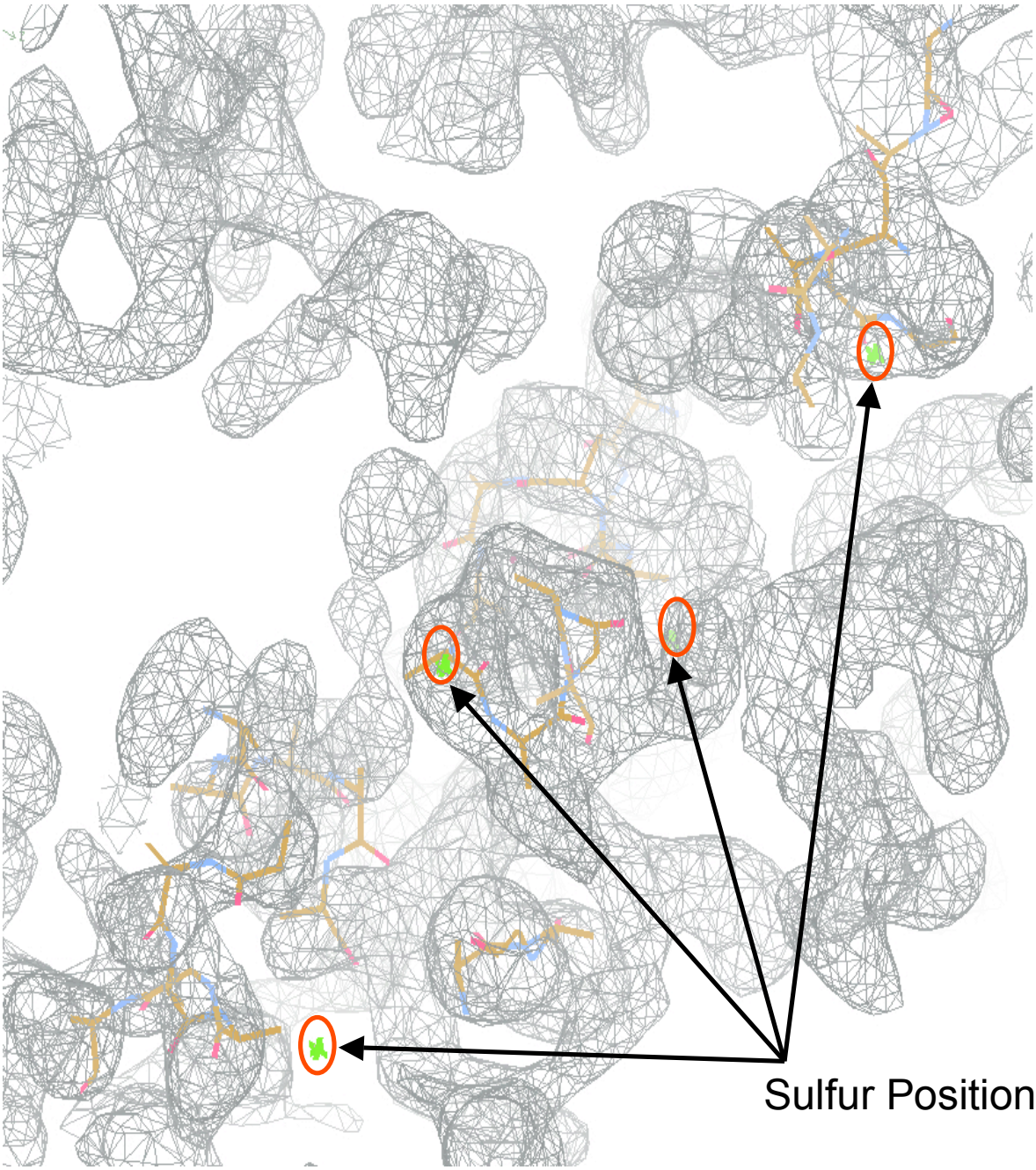

Figure 4.2.10: Tracing result from d*TREK processing of R1-R2 Merged data set: Four Sulfur positions were correctly identified in Red, with 51% of the total Amino Acids traced. The trace from RESOLVE did not agree with the refined model.

R1-R2 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------  -----------
 1      48       7      12 8 7  /Merge_R12_265/sgx/t4/zzsgxSol_1.pdb
Heavy atoms:                    /Merge_R12_265/sgx/t4/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4.
#Alpha-helix: 1 helices with 6 amino acids long.
#Beta-sheets: 1 beta sheet(s) 6 amino acids long.
CC_ALL/CC_WEAK    28.54/10.35
PATFOM      31.35
```

Table 4.2.9: R1-R2 Tracing results from SGXPro Novel Structure Solution: R1–R2 Merged processing yielded acceptable CC-ALL/Weak values but substandard PATFOM, phases and tracing statistics.

4.3 MOSFLM

This data reduction program shares many similarities with other programs in this study. The data processing engine MOSFLM (37, 83)(version 7.0.5) includes a new GUI interface accompanying other changes to the program. The output file from MOSFLM is a .mtz containing integrated intensities which can then be used in scaling and further processing. Originally an X-ray integration program used for film data dating back to the 1970's. The first suggestion for including a GUI interface was proposed in 1992 and further advancements were made allowing for processing image plate data. The newest version of MOSFLM implements an advanced GUI interface containing a user-friendly flowchart for data reduction. Initiating the program is achieved by using the command line interface, typing:

>MOSFLM !This will initiate the processing GUI and display the start screen;

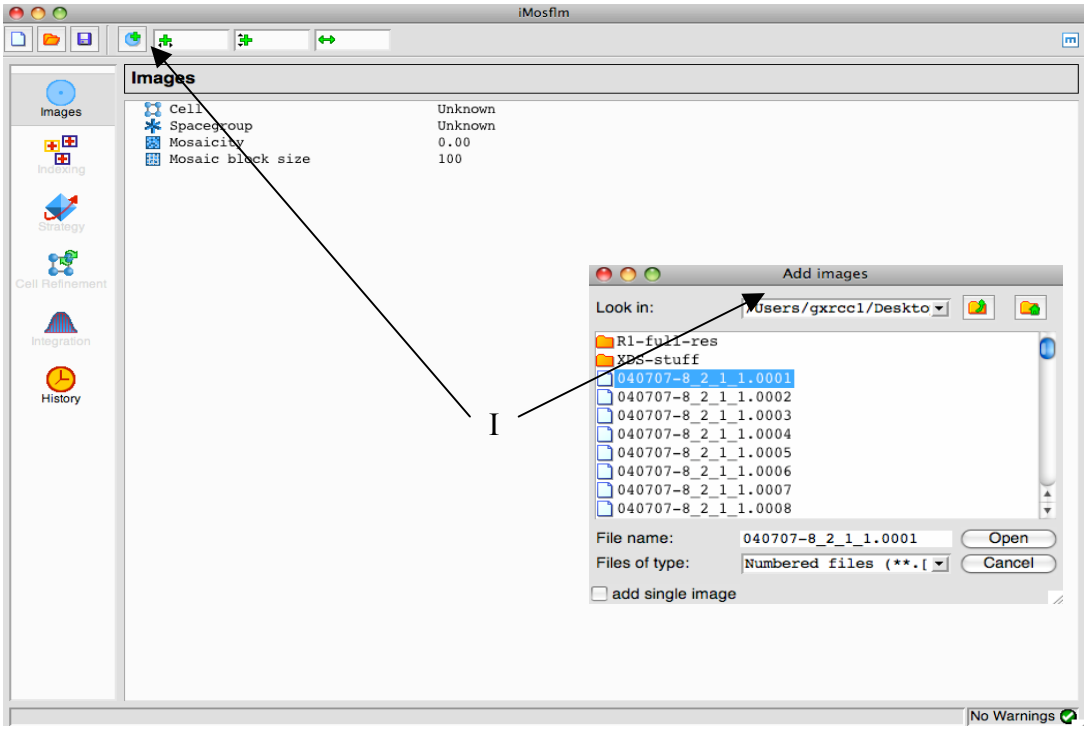

Figure 4.3.1: MOSFLM processing GUI: This window serves as the primary data reduction GUI. (I) Image selection proceeds by selecting the Add Images icon and opening the appropriate diffraction images.

The individual sub-routines for Indexing, Strategy, Cell Refinement and Integration are displayed in a flow chart to the right of the GUI. These cannot be selected until the proper steps are taken such as indexing before integration. Selecting the Add images (Figure 4.3.1; I) icon will allow searching for image files to be used in processing. For this study the Numbered files designation was used to identify the diffraction images. After images are added the first image is displayed within a image GUI;

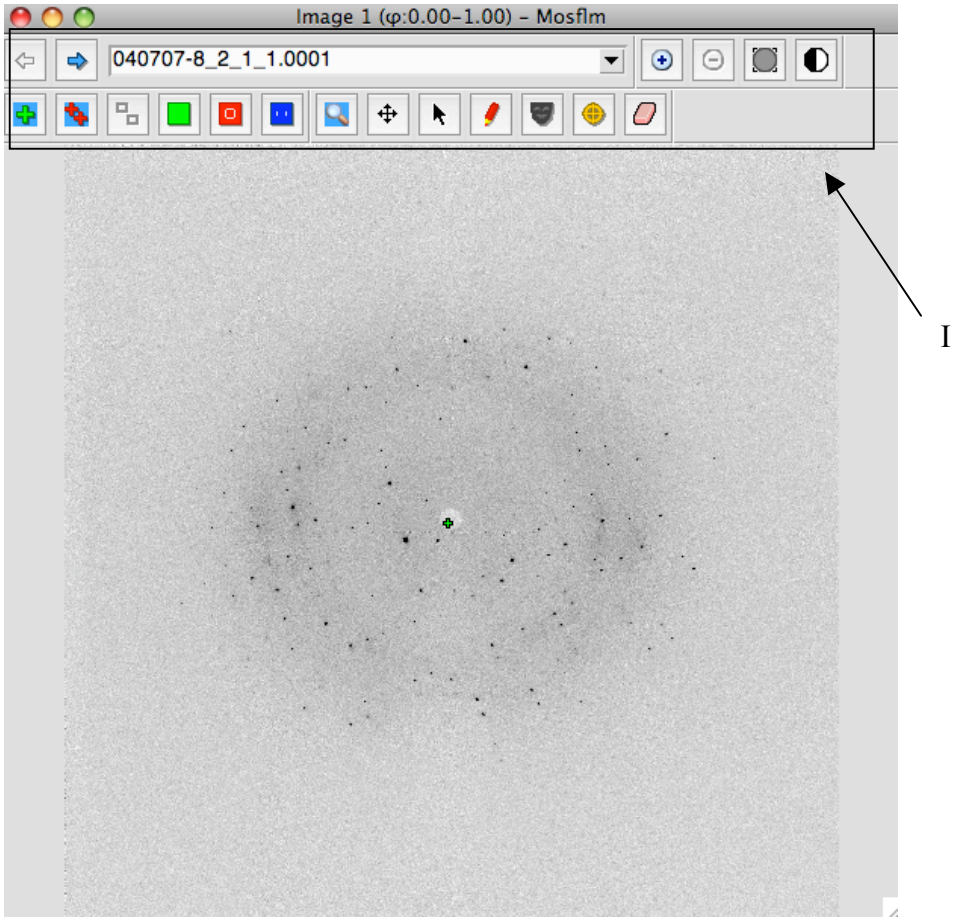

Figure 4.3.2: Image Display Window: Image viewing tool which allows the user several options contained in the toolbar menu (I).

If the program correctly read the header information a small green cross will identify the beam center. The most important options in this GUI are the beam center adjustment, beam stop masking, and the zoom command – allowing for individual reflections can be visual analyzed.

Figure 4.3.3: Top most Image Display toolbar

Starting from the left the arrow keys on the toolbar will scroll through the images to be processed. Each image is identified by its title. The remaining options on the right side of the toolbar are for the purposes of zoom, resetting and altering the contrast of the image. These tools are primarily used to perform a visual account of image and spot quality.

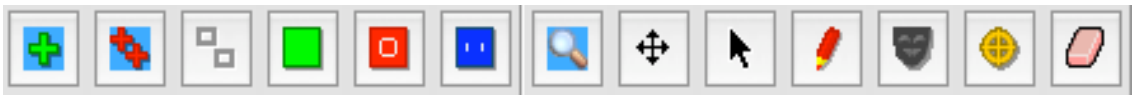

Figure 4.3.4: Image Display toolbar

The six icons to the left of the magnifying glass (for selectively magnifying portions of the image) represent the incident X-ray beam position, spots found during indexing, predicted spot locations, masked areas and search area used during spot finding respectively. To the right of the magnifying glass are icons for panning, selection, spot addition, mask editing, circular fitting, and an eraser for removing errant masks or spots.

Finding spots

After adding images, the indexing option becomes active from the processing flowchart. Default parameters used are images 1° and 90°. If the data set contains less than 90 degrees the choices will be the first degree (1°) and largest degree possible. The spot finding parameters are located via parameters under “view” and “processing option” in the program toolbar;

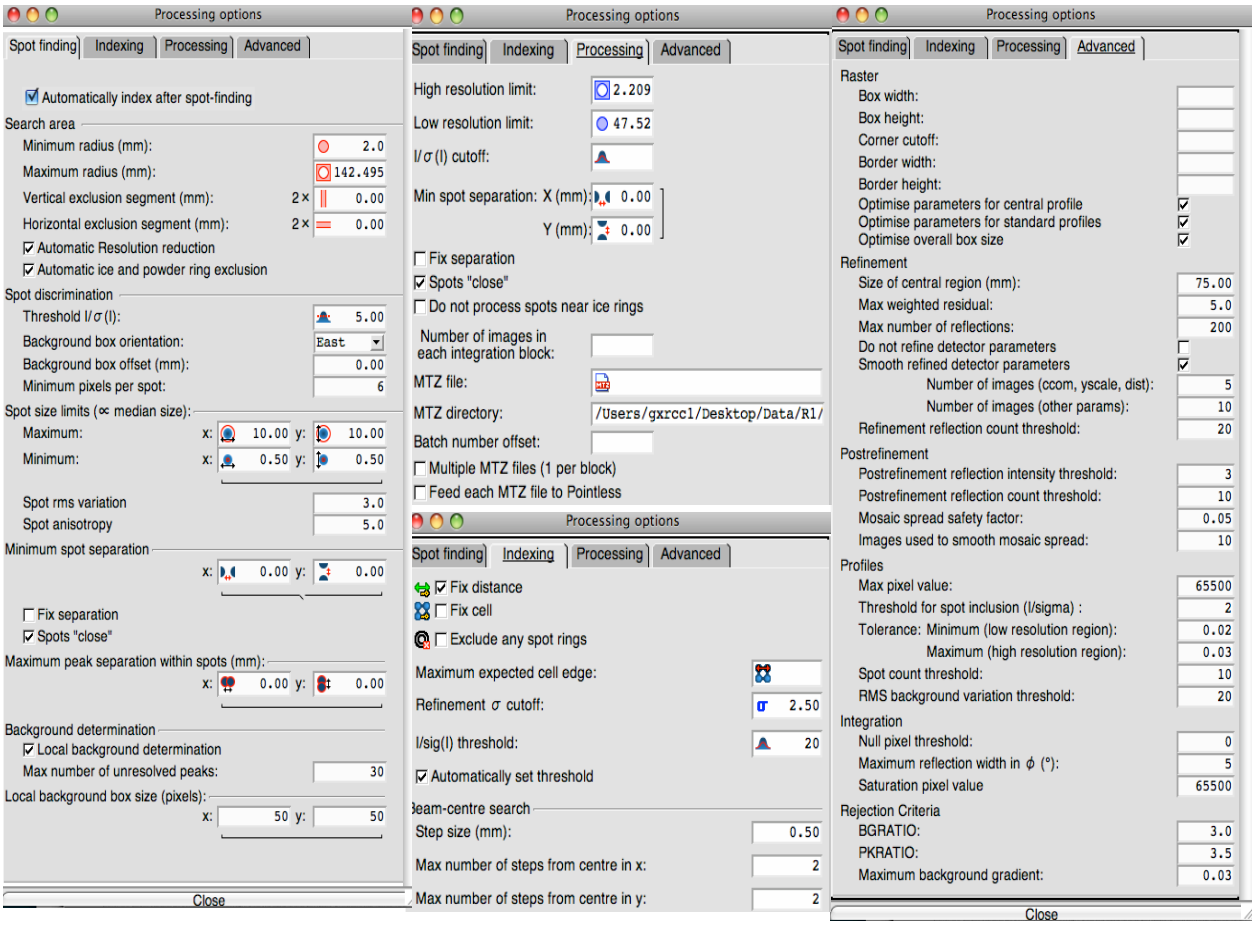

Figure 4.3.5: Processing options: MOSFLM contains a small set of options for each function of the program, which can be altered to improve data processing. Although these options require more work from the user to discern their function there is adequate documentation available to do so.

The resultant window contains four tabs for adjusting the parameters pertaining to spot finding, indexing, processing (integration) and advanced (portions of refinement and integration). The options exist for manual addition or removal of spots not included with the initial spot finding process. The user also has the option to change the area used to search for and the I/σ_I threshold for spot versus background determination, allowable spot size, spot separation limits, and in the case of split spots, the maximum allowable peak separation within spots. The background determination function offers the user a choice between a local and radial method. If no changes are necessary for satisfying the MOSFLM's requirements for number and location of

spots, indexing will automatically continue. However if parameters need to be changed the option to Automatically index after spot finding must be disabled. This is contained under "view" and "processing option" in the program toolbar in the spot finding tab.

Indexing

MOSFLM, like HKL2000, can use as little as one image or a span of images to locate diffraction spots and determine the unit cell and orientation of the crystal, as long as the spot requirements are met. Once the user has selected the index function, a list of solutions are generated and listed in a display box.

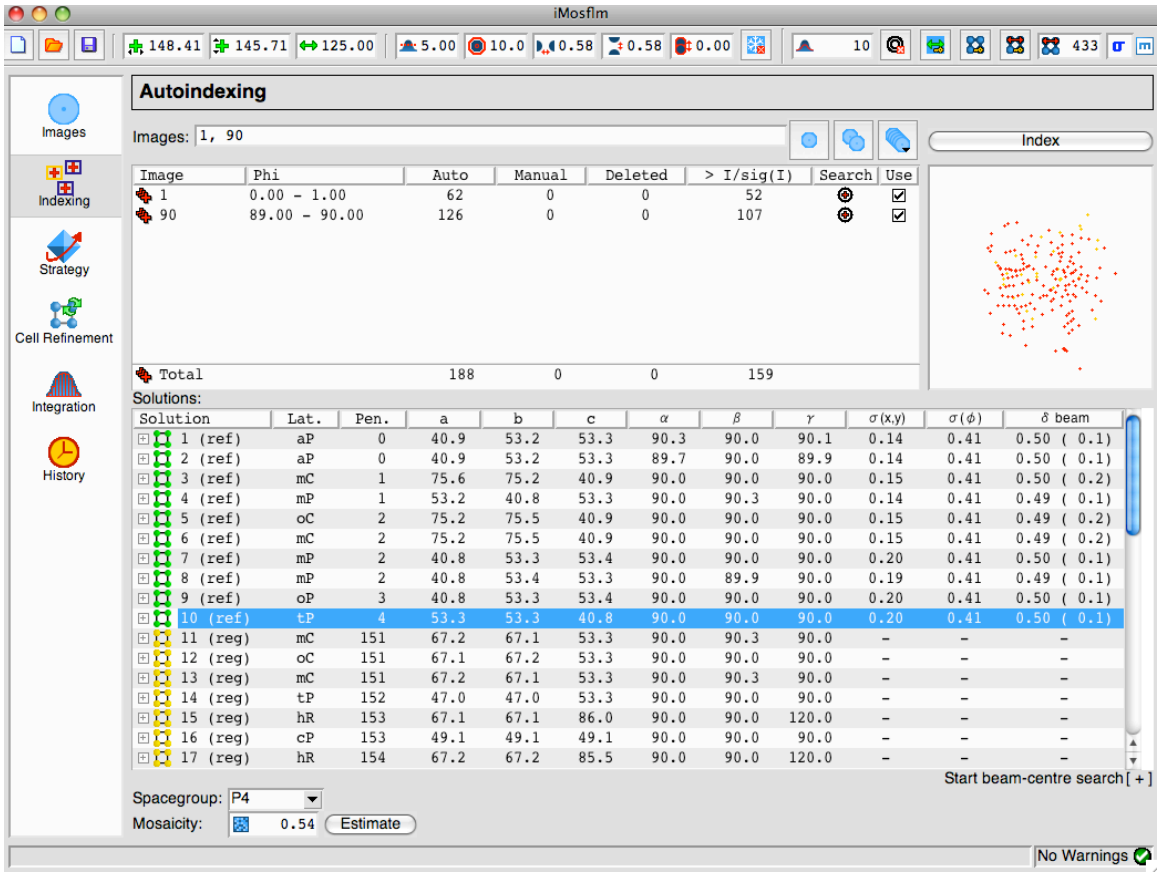

Figure 4.3.6: MOSFLM Indexing GUI: After selecting the Indexing icon from the Flow Chart the program displays information concerning the images used as well as identifying the best solution based on penalty score.

The possible solutions appear sorted by penalty score and preferred solutions are highlighted in blue. The acceptable solutions should have a penalty between 0-20. Just as in HKL2000, the penalty is similar to the distortion index (Chapter 3 section 1; Indexing). The penalty scores are also accompanied by error analysis in both x, y, and φ (σ(x,y); σ(φ)). If there are no errors in the supplied or calculated direct beam coordinates, detector distance, or wavelength, the user should chose a solution with a penalty score below 20, but possessing the highest symmetry. MOSFLM will also display the indexed images highlighting the predicted pattern for the solution chosen using colors to differentiate the type of spots selected:

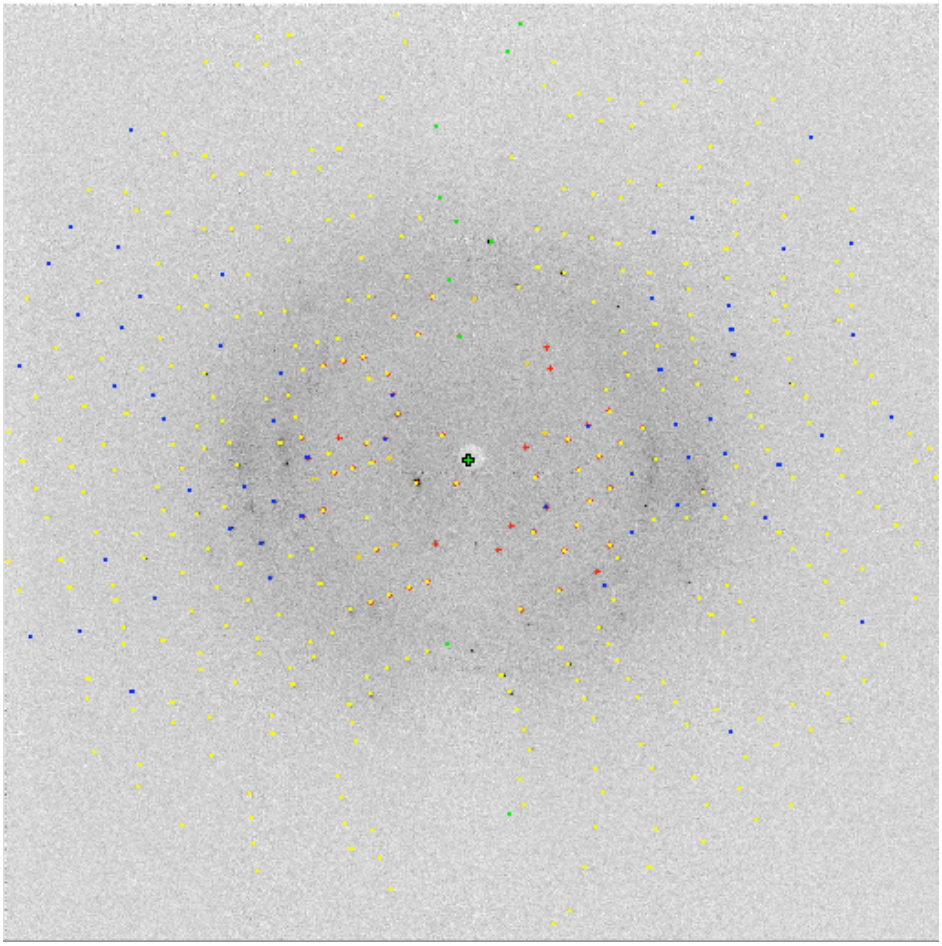

Figure 4.3.7: Spot prediction patterns: The colors highlighted on the image represent MOSFLM's classification of the index corresponding to space group identification.

Green for reflections spanning more than 5 images, Yellow for partially recorded reflections, Red for overlapped reflections which will not be integrated, and Blue for fully recorded reflections. If a alternative solution other than that which was recommended by the program is selected the predicted patters will change. This is the only opportunity the user has to examine the quality of observed versus predicted spot positions. MOSFLM does offer the user the option to edit the default values for $I/\sigma_{(I)}$ cutoffs for spot selection, ice ring exclusion algorithms, and the option to chose the parameters considered during indexing. The Indexing parameters governing these options are contained under "view" and "processing option" within the program toolbar. The most pertinent result of initial indexing and cell refinement is the RMS residual value, the positional error between predicted and observed spot.

Refinement

The importance of determining the cell parameters accurately during the refinement process has already been discussed in this work. Initial refinement occurs during auto-indexing, and additional accuracy concerning the crystal parameters (cell dimensions, orientation matrix, and mosaicity), Beam parameters (orientation, divergence) and Detector parameters (detector position) is achieved during post refinement. This procedure requires batch integration of images at widely differing φ values. The intensity distributions of partial reflections throughout these images are used to further refine the crystal orientation, unit cell, and mosaic spread. The user must assure predicted and observed spots are well aligned, or the post-refinement will not function properly. The post refinement technique used in MOSFLM offers the user a choice of cell refinement solutions. The values corresponding to the initial refinement conducted during indexing or the parameters generated in post refinement can be chosen to represent the best refinement cycle. Its best to allow MOSFLM to do this unless prior information is known

concerning the crystal. A comparison between the output RMS residual values of the refinement routines should be compared to ensure the user chooses the best refinement state for integration. I chose to use images 1-20 for post refinement as the default values of the first four images did not generate a differences in cell parameter values

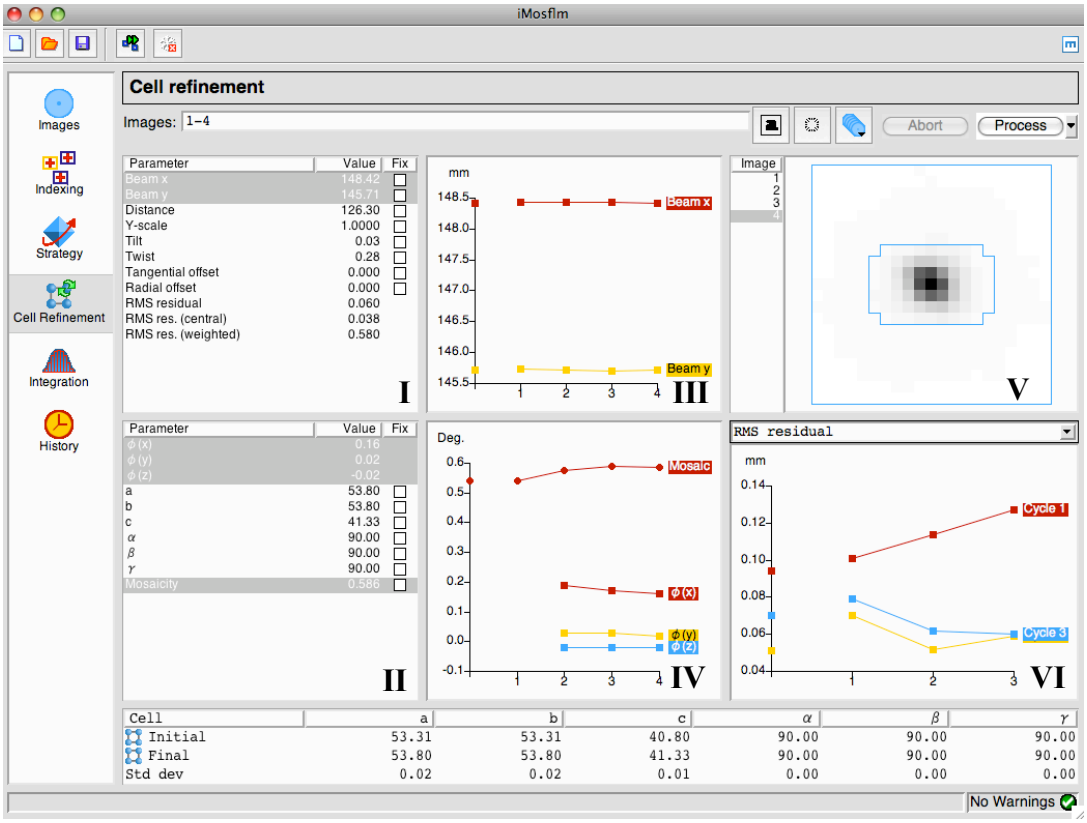

Figure 4.3.8: MOSFLM Cell Refinement: The Refinement GUI within MOSFLM highlighting the detector and crystal parameters (I, II) accompanied by a visual representation of offsets from initial indexing measured values in mm and degrees respectively (III, IV). The central spot profile window (V) represents the average profile for spots in the central region of the detector and (VI) finally the RMS residual illustrates the correlation of different refinement trials.

A well matched box and spot represents efficient integration. Last is the summation window (VI) reflects the RMS values, traditionally the lower this value the better the refinement. Images 1-4 were used for each cycle of refinement. The final refinement solution achieved slightly lower a lower RMS values than 1st cycle, but as I have discussed precision is everything when determining the proper index for data processing.

The user can select parameters to be fixed or allowed to vary due to parameter correlations during post refinement; provided the default settings for MOSFLM are not changed (Figure 4.3.5). These parameters include: Beam (x,y), Detector distance, Y-scale, Detector Tilt, Detector Twist, Tangential offset, Radial Offset, RMS residual, RMS res. central and RMS res. weighted. Additional parameters can, but rarely should, be fixed such as the φ(x,y,z), a, b, c, α, β, γ, and mosaicity. Of these values the RMS residual, central and weighted depend directly on the positional error in predicting reflections. I noticed that the central RMS res. value of 0.038 is at the high end of acceptable ranges as defined by MOSFLM. This is usually linked to error in the cell parameters. In addition to this the weighted RMS res. value should be close to unity which is obviously not the case at 0.58 (Figure 4.3.8). The reason for this difference in expected value is apparently linked to the GAIN of the detector. However I have not found a means to edit the GAIN value associated with the Marr 300 CCD detector used in this work. Lastly the RMS residual value was within acceptable ranges for the program. Attempts were made during multiple rounds of refinement to fix parameters which consistently appeared errant such as the Beam y, mosaicity, and ϕ(x,y,z) values. These efforts generated disastrous results at the conclusion of integration and, as expected, output from scaling. Next I proceed to integration accepting the default values and fix/un-fixed parameters within MOSFLM.

Integration

It is recommended that integration be conducted with a sub-set of the entire data set such that the user can discern if the cell parameters are acceptable. I used 10 images and repeatedly encountered errors concerning the detector GAIN values as well as null pixel values at the center of the detector.

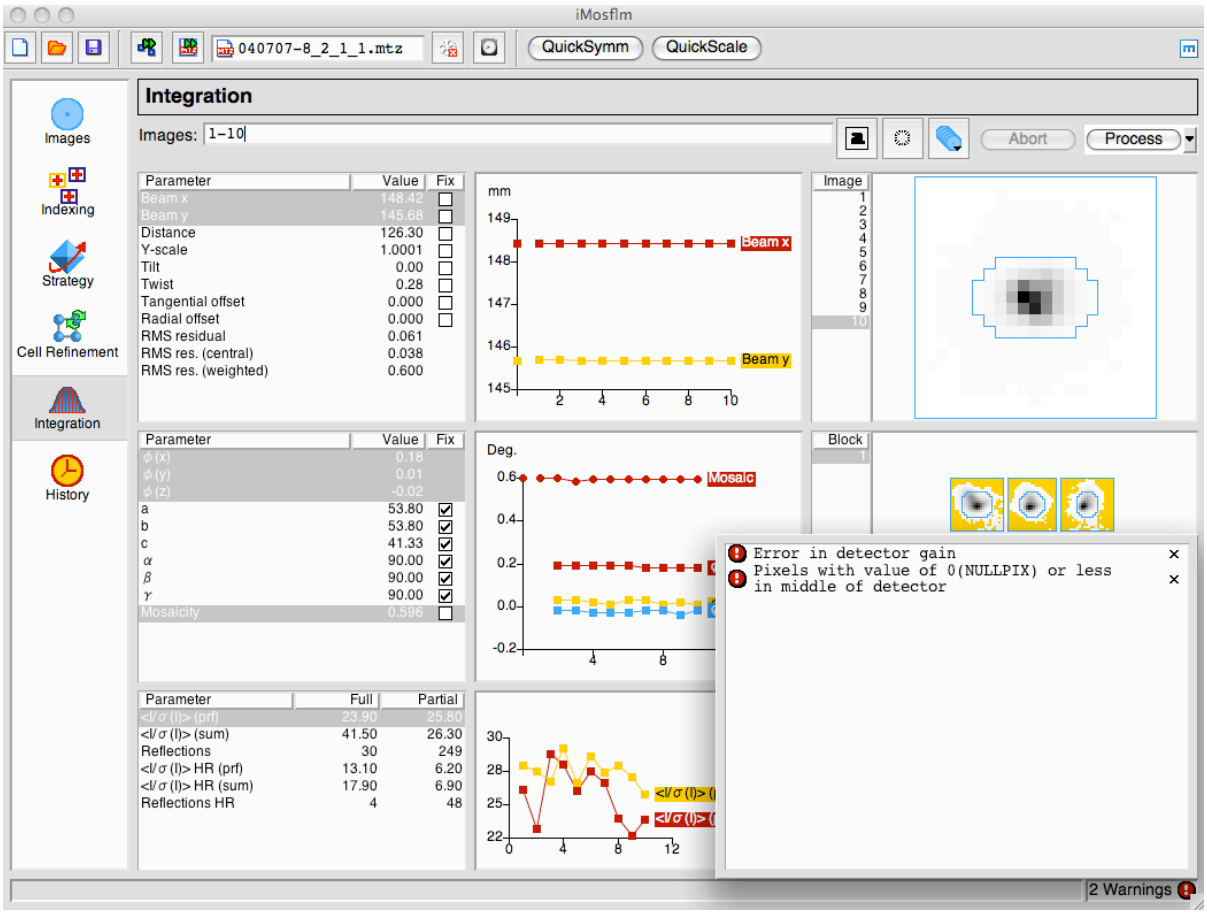

Figure 4.3.9: Test Integration: The first 10 frames were used to "test" the refinement parameters generated in the previous step.

Searching for a solution to these errors I consulted the MOSFLM tutorial (84). The two error flags correspond to the following;

Error in detector gain is based on the weighting for the residuals is calculated with the gain - for standard detectors the gain is given a default value of 0.03, I did change this value as suggested in the error warning to 0.39. This removed this error but did not improve the finial processing results.

Pixels with value of 0 (NULLPIX) or less in the middle of the detector is triggered by bad pixels (which have a count of 0) or the gap between tiles has been marked incorrectly by

MOSFLM, or you have a very low background values which rarely happens with proteins. I attempted to mask the beam stop in an effort to combat this issue. This action produced no change in the generation of this error either.

MOSFLM and XDS are the only programs I have found to both give the user a clear indication of errors during data processing accompanied by suggestions to address these errors. For a Novice user this is in all likelihood the only way to know how to approach and error without painstaking reading or hunting the internet for assistance. Unfortunately the errors could not be remedied by my level of understanding of the program despite following the suggestions the authors were gracious enough to offer. Just as programs such as HKL2000 and d*TREK may indicate less than ideal processing parameters by colored warnings or within log files, there is no steadfast rule that any errors noted by the program leave structure determination impossible. I next proceeded with standard integration of the full data set.

Integration within MOSFLM is a two-pass process. First, batches of images are used for group refinement. Second, standard profiles are generated from the spots selected from each image for integration and output intensities to a MTZ file. During the first pass of refinement, only the crystal orientation and mosaicity are refined (unit cell dimensions are fixed). During integration MOSFLM displays several tables and selected parameters including the profile display (Figure 4.3.10; VIII). The user should consider the standard profile for difference regions of the detector. If the spot profile dimensions and box centering are in doubt this could be a indication that the prediction is poor.

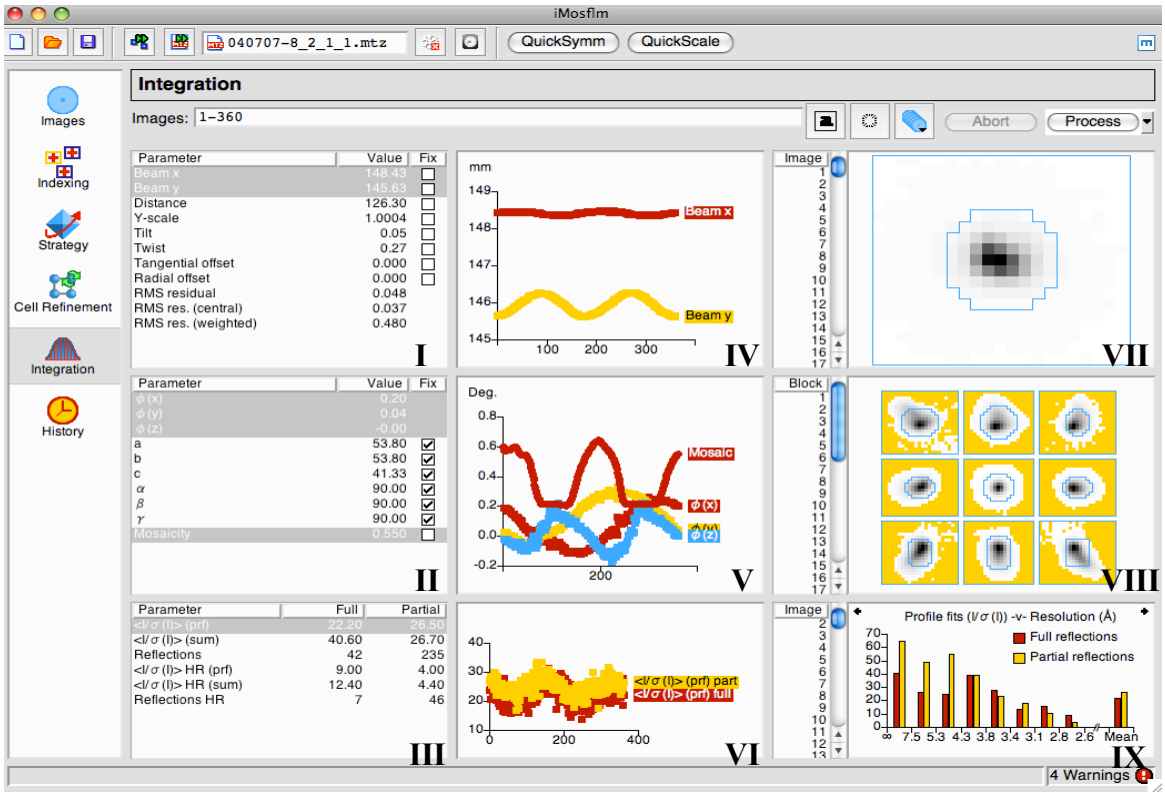

Figure 4.3.10: MOSFLM Integration: The Integration window displays 9 charts containing information collected in real time during the integration process. (I, II, VII) Are explained in Figure 4.3.9 as the refined crystal and detector parameters and the average profile for spots in the central region of the detector. (III) Relays information concerning the I/σ_I values throughout the data set. (IV, V, VI) Are graphical extensions of the information located in the window(s) to their immediate left. (VIII, IX) Represent the standard profiles from different regions of the detector and the I/σ_I values as a function of resolution.

The warnings generated in the bottom right-hand portion of Figure 4.3.11 are alerts the integration engine notes during integration, and are as follows;

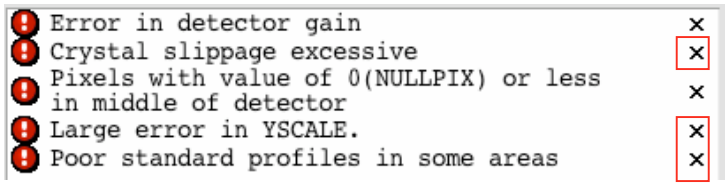

Figure 4.3.11: Error(s) generated during MOSFLM Integration: In an addition to the errors found during Test Integration (Figure 1.7) three errors were generated (boxed in red)

At the conclusion of integration the program recorded additional errors. The first of these was *Crystal slippage excessive,* this implies the ϕ(x,y,z) missetting angels are changing more than allowed by MOSFLM. This is usually due to the rotation axis not being perpendicular to the X-ray beam, not the crystal slipping. This was a more alarming error as the missetting angles originate from the orientation matrix [A] (Chapter 2, section Indexing). I was told by the author not to be concerned unless the graphical representations of these values "jump about erratically". Upon review of these values (Figure 4.3.10; V) movement of the calculated missettings were smooth throughout integration. *Large error in YSCALE,* is an indicator of a possible problem in the generation of the orientation matrix [A], however MOSFLM triggers this error if the value deviates from unity by more than 0.0002. The integration of the data from this study resulted in a deviation of 0.0003, this is generally considered bearable because of the resolution range of 2.65 angstroms. *Poor standard profiles in some areas* is a ambiguous error flag which I attempted to remedy by reducing the integration box sizes using an increase in profile tolerance also I attempted an increasing the block size to include more images per batch during integration. I experimented with a wide variety of measurements as recommended within MOSFLM documentation. An additional option is limiting the diffraction of the data; I attempted this back to 3.0 angstroms by 0.25 angstrom increments with no success.

The output mtz file can be then scaled using QuickSymm and QuickScale from the top of the Integration window.

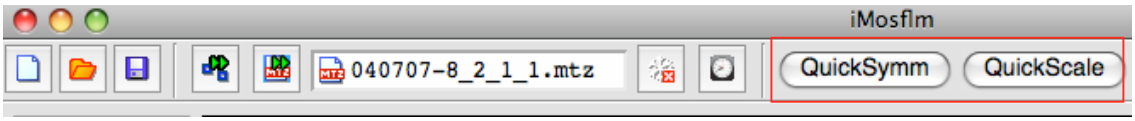

Figure 4.3.12: Scaling portion of the MOSFLM GUI: Selecting 1st QuickSymm executes the CCP4 program pointless, which determines the accurate space group while testing alternative indexing schemes, followed by selecting QuickScale which executes SCALA another CCP4 program, for scaling in.

The choice of Pointless (38) and SCALA (85) for scaling was based solely on their inclusion in the MOSFLM GUI. The methodology used in the SCALA algorithm is discussed in Chapter 3, and covered in grater detail by Evans (86). The mtz output from integration was not properly formatted for scaling using 3DSCALE. Efforts were made in conjunction with Dr. Zin Quing Fu concerning this conversion but were unsuccessful. The results of scaling are listed below;

R-factors -

$$R_{sym} = \sum_h \sum_i \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

R_{sym} is the R-factor chosen 3DSCALE to relate differences in symmetry related reflections. This is a measure of the accuracy of the data. The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection.

$$R_{meas} = \sum_h \sqrt{n_h /(n_h - 1)} \sum_i \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

A alternate indicator of data quality proposed by Diederichs and Karplus (87) to remove the redundancy dependence of R_{sym}. This value, R_{meas}, includes a term $\sqrt{[n/(n-1)]}$ which appropriately weights individual reflections (h) according to their multiplicity (n_h).

R1 - Scaling Statistics using SCALA

R1- MOSFLM

Summary data for Project: JTS: New Dataset: R1

	Overall	InnerShell	OuterShell
Low resolution limit	41.33	41.33	2.79
High resolution limit	2.65	8.38	2.65
Rmerge	**0.083**	0.064	**0.217**
Rmerge in top intensity bin	0.066	-	-
Rmeas (within I+/I-)	0.090	0.069	0.234
Rmeas (all I+ & I-)	**0.089**	0.069	**0.231**
Total number of observations	50091	1559	7101
Total number unique	3542	124	516
Mean((I)/sd(I))	**22.6**	32.3	**8.7**
Completeness	**100.0**	99.2	100.4
Multiplicity	**14.1**	12.6	**13.8**
Anomalous completeness	100.0	98.3	100.4
Anomalous multiplicity	7.4	7.3	7.1

Average unit cell: 53.80 53.80 41.33 90.00 90.00 90.00
Space group: P 42
Average mosaicity: 0.37

Table 4.3.1: R1 Scaling results from MOSFLM processing using SCALA: R1 processing yields high but still acceptable R_{merge} results for continued processing.

R1 Automated tracing results using SGXPro

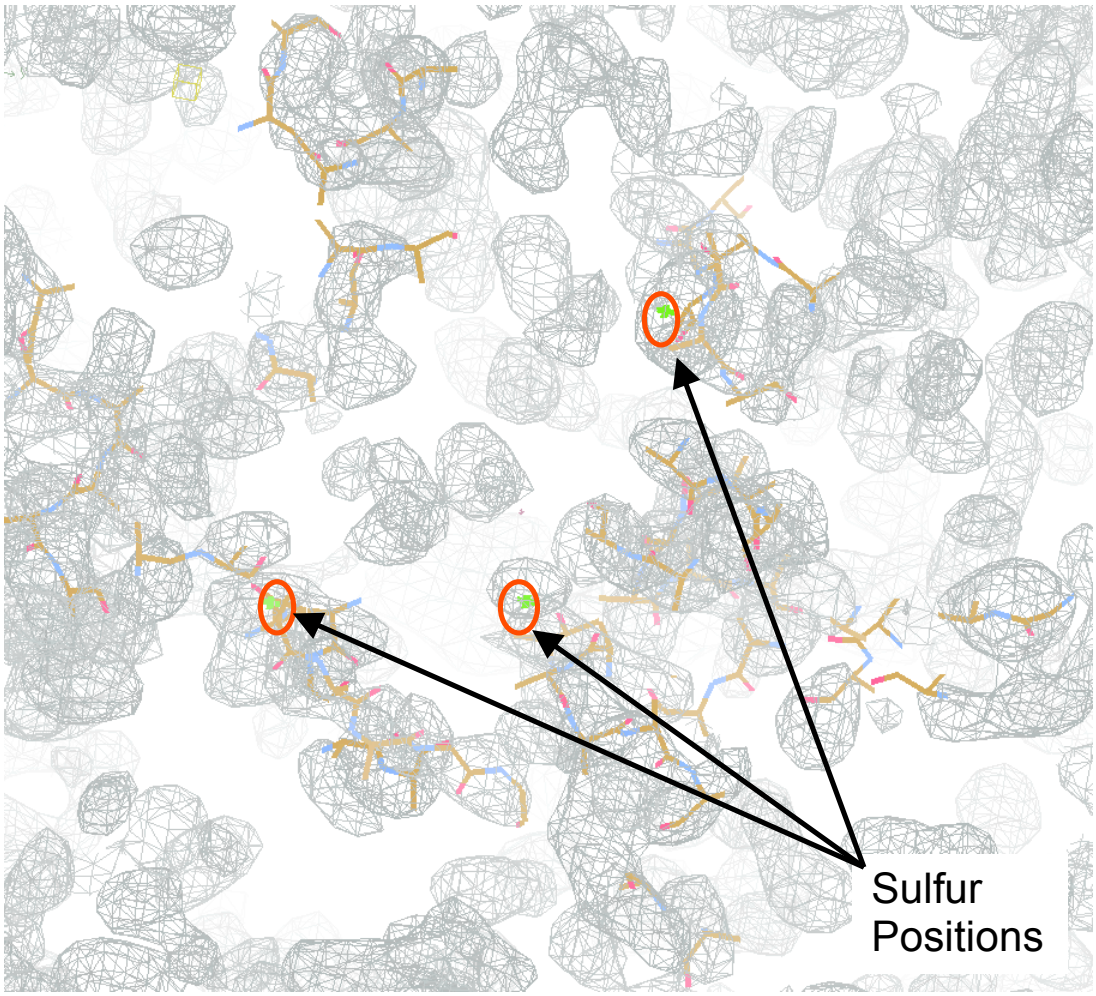

Figure 4.3.13: Tracing result from MOSFLM processing of R1 data set: Three Sulfur positions were correctly identified in Red, with 69% of the total amino acids traced. The trace from RESOLVE did not agree with the refined model.

R1 - Heavy Atom/Tracing statistics using SGXPro

No#	NumBuilt	NumSegs	Top3Segs	Model Built
1	66	13	15 7 5	/MOSFLM/R1/t1/zzsgxSol_1.pdb

```
Heavy atoms:                          /MOSFLM/R1/t1/zzsgxSol_1_ha.xyz
Sulfur atoms found:     3
#Alpha-helix: 3 helices 8, 8, 8 amino acids long
#Beta-sheets: no discernible beta sheets
CC_ALL/CC_WEAK    23.09/11.59
PATFOM      64.33
```

Table 4.3.2: R1 Tracing results from SGXPro Novel Structure Solution: R1 processing yielded poor CC-ALL/Weak and PATFOM values. The phases and tracing statistics were substandard as well.

R2 - Scaling Statistics using SCALA

R2- MOSFLM results

Summary data for Project: JTS: New Dataset: R2

	Overall	InnerShell	OuterShell
Low resolution limit	41.31	41.31	2.79
High resolution limit	2.65	8.38	2.65
Rmerge	**0.083**	0.074	**0.147**
Rmerge in top intensity bin	0.068	-	-
Rmeas (within I+/I-)	0.090	0.083	0.158
Rmeas (all I+ & I-)	**0.089**	0.081	**0.156**
Total number of observations	50134	1216	7196
Total number unique	3529	121	510
Mean((I)/sd(I))	**23.6**	25.6	**12.1**
Completeness	**99.9**	97.1	100.0
Multiplicity	**14.2**	10.0	**14.1**
Anomalous completeness	99.9	94.9	100.0
Anomalous multiplicity	7.4	5.7	7.3

Average unit cell: 53.74 53.74 41.31 90.00 90.00 90.00
Space group: P 42
Average mosaicity: 0.35

Table 4.3.3: R2 Scaling results from MOSFLM processing using SCALA: R2 processing yields high but still acceptable R_{merge} results for continued processing.

R2 Automated tracing results using SGXPro

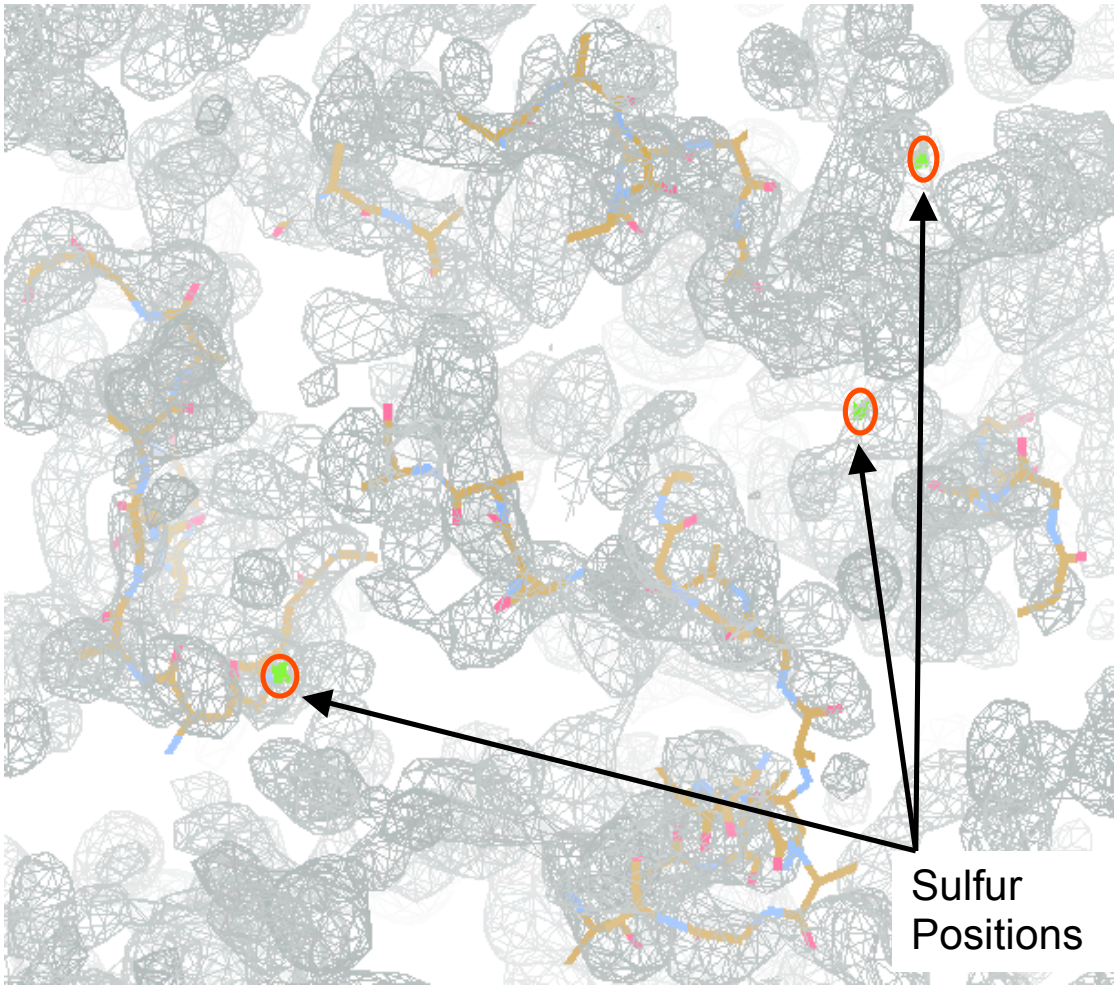

Figure 4.3.14: Tracing result from MOSFLM processing of R2 data set: Three Sulfur positions were correctly identified in Red, with 65% of the total Amino Acids traced. The trace from RESOLVE did not agree with the refined model.

R2 - Heavy Atom/Tracing statistics using SGXPro

No#	NumBuilt	NumSegs	Top3Segs	Model Built
3	**62**	**10**	**10 7 7**	/MOSFLM/R2/t1/zzsgxSol_3.pdb

Heavy atoms: /MOSFLM/R2/t1/zzsgxSol_3_ha.xyz
Sulfur atoms found: 3
#Alpha-helix: 2 helices both 7, 5 amino acids long
#Beta-sheets: 1 beta sheet 10 amino acids long
CC_ALL/CC_WEAK 17.3/5.3
PATFOM 92.42

Table 4.3.4: R2 Tracing results from SGXPro Novel Structure Solution: R2 processing yielded poor CC-ALL/Weak and PATFOM values. The phases and tracing statistics were substandard as well.

R1-R2 merged - Scaling Statistics using SCALA

R1-R2_SCALA results

Summary data for Project: JTS: New Dataset: R1-R2

	Overall	InnerShell	OuterShell
Low resolution limit	41.32	41.32	2.79
High resolution limit	2.65	8.38	2.65
Rmerge	**0.089**	0.076	**0.192**
Rmerge in top intensity bin	0.073	-	-
Rmeas (within I+/I-)	0.092	0.080	0.199
Rmeas (all I+ & I-)	**0.092**	0.080	**0.198**
Total number of observations	100218	2777	14249
Total number unique	3542	124	516
Mean((I)/sd(I))	**32.6**	40.9	**14.3**
Completeness	**100.0**	99.2	100.4
Multiplicity	**28.3**	22.4	**27.6**
Anomalous completeness	100.0	98.3	100.4
Anomalous multiplicity	14.8	12.9	14.2

Average unit cell: 53.77 53.77 41.32 90.00 90.00 90.00
Space group: P 42
Average mosaicity: 0.36

Table 4.3.5: R2 Scaling results from MOSFLM processing using SCALA: R1-R2 Merged processing yields high but still acceptable R_{merge} results for continued processing.

R1-R2 Merged Automated tracing results using SGXPro

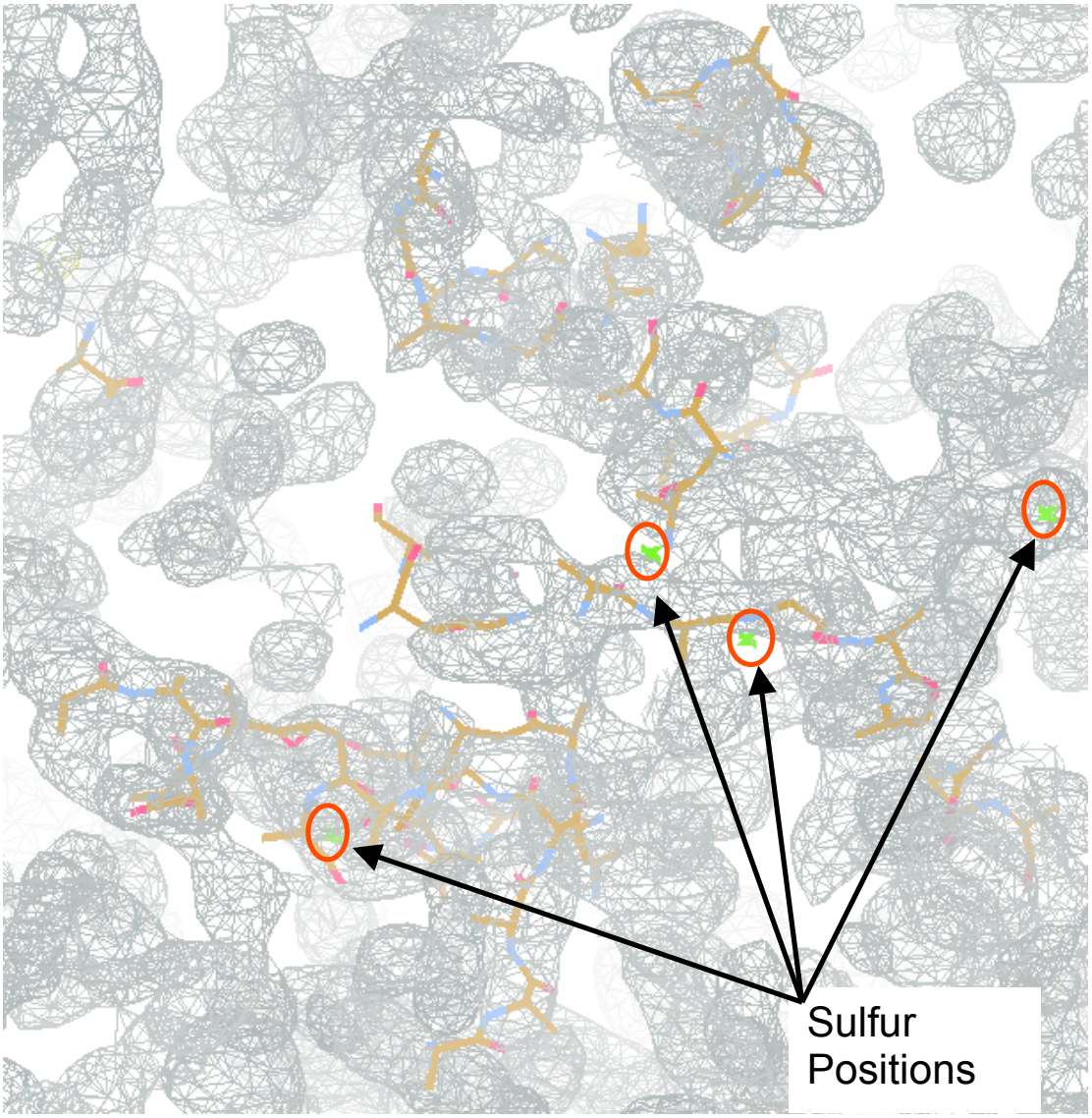

Figure 4.3.15: Tracing result from MOSFLM processing of R1-R2 merged data sets: Four Sulfur positions were correctly identified in Red, with 72% of Amino acids traced. The trace from RESOLVE did not agree with the refined model.

R1-R2 merged - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------- -----------
 2      68       13     8 8 8   /MOSFLM/R1-R2/t2/zzsgxSol_2.pdb
Heavy atoms:                    /MOSFLM/R1-R2/t2/zzsgxSol_2_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 3 helices 8, 8, 8 amino acids long
#Beta-sheets: no discernible beta sheets
CC_ALL/CC_WEAK    23.02/10.66
PATFOM      74.72
```

Table 4.3.6: R1-R2 Tracing results from SGXPro Novel Structure Solution: R1-R2 processing yielded poor CC-ALL/Weak and PATFOM values. The phases and tracing statistics were substandard as well.

4.4 XDS

The XDS (extended Development System) software suite is the only purely text based data reduction program within this study. XDS was conceived in 1991 by Wolfgang Kabsch and the first version of the system was implemented 1992 under the name OM2. The XDS algorithm was developed for the first automatic interpretation of reciprocal lattice points in 1993 (88). There exist no GUI interface and all commands are contained within individual scripts. This is undoubtedly a daunting program for users which are accustom to GUI based versions of classical data reduction programs such as HKL2000, d*TREK, MOSFLM and PROTEUM2. It is necessary to mention that nearly all data reduction programs were originally script/text based. The newer GUI interfaces offer present day users a level of comfort involved with visual processing as well as graphical statistical quality checks. The XDS scripts can be written on a basic or advanced level depending on the parameters the user wishes to address. The XDS script uses sub menus (*written in italics*), keyword commands (**written in bold**), and various subroutines (underlined) to determine the parameters of processing. In the most basic of scripts, the user need only identify the detector used, x-y origins of the beam center, detector distance, wavelength, location of the images and the number of images to be used during background calculation during spot selection (Figure 4.4.1). XDS does not read any header information, thus, all parameters must be input into the XDS.INP file by the user. The sub menu entitled JOB CONTROL PARAMETERS controls the tasks conducted by the XDS.INP. This section contains the commands for the following subroutines: XYCORR creates spatial correction tables for every pixel on the detector; INIT determines the initial background and gain of the detector; COLSPOT determines the locations of strong reflections; IDXREF conducts initial indexing from strong reflections for identification of unit cell information; DEFPIX used in isolation of

masked portions of the detector; INTEGRATE calculates the intensities of the predicted reflections in three dimensions; and CORRECT which scales integrated intensities while accounting for sensitivity variations of the detector face, refining diffraction parameters and corrects intensities due to decay. An important factor of XDS functioning purely from a single script is the necessity of acquiring the script form a individual with knowledge of functions contained therein. At minimum the custodian of the script should have prior knowledge of the scripts use similar experiments to those desired by the experimenter. A novice user would find understanding which of the approximately 424 possible values and combinations should be altered to produce the best results a truly daunting task. The program authors claim there are only 30 relevant parameters within the script and of these only 15 that are commonly changed. This may be true in the case of high quality or even average quality data. Therefore, I found it best to contact the author of the program who advised me where I should download template for XDS data reduction;

```
!*****************************************************************************
! File XDS.INP  containing named arguments for running XDS (arbitrary order).
! Characters in a line to the right of an exclamation mark are comment.
!*****************************************************************************
!********** Example for MAR CCD-detector at ESRF beamline ID14-1 *************
!********** and the 1024 X 1024 CCD-detector at CHESS            *************
!************************DETECTOR PARAMETERS*********************************
DETECTOR=CCDCHESS        MINIMUM_VALID_PIXEL_VALUE=1     OVERLOAD=65000
DIRECTION_OF_DETECTOR_X-AXIS= 1.0 0.0 0.0
DIRECTION_OF_DETECTOR_Y-AXIS= 0.0 1.0 0.0
TRUSTED_REGION=0.0 0.99 !Relative radii limiting trusted detector region
!File name, access, format of dark-current (non-Xray background) image
!DARK_CURRENT_IMAGE=../images/blank.tif    !hardly ever used

!MAXIMUM_NUMBER_OF_JOBS=4  !Speeds-up COLSPOT & INTEGRATE on a Linux-cluster
MAXIMUM_NUMBER_OF_PROCESSORS=6!<25;ignored by single cpu version of xds
!MINUTE=0    !Maximum number of minutes to wait until data image must appears
!TEST=1      !Test flag. 1,2 additional diagnostics and images

!NX=number of fast pixels (along X); QX=length of an X-pixel (mm)
!NY=number of slow pixels (along Y); QY=length of a  Y-pixel (mm)
!Select the correct detector parameters by uncommenting the appropriate line
NX=4096 NY=4096 QX=0.073242 QY=0.073242!MARCCD 300mm at APS 22ID
!NX=3072 NY=3072 QX=0.07345  QY=0.07345 !MARCCD 225mm at APS 22BM
!NX=2048 NY=2048 QX=0.079    QY=0.079   !MARCCD 165mm version
!NX=2048 NY=2048 QX=0.064    QY=0.064   !MARCCD 133mm version
!NX=1024 NY=1024 QX=0.0508   QY=0.0508  !CCD at CHESS

! Do not forget to define ORGX and ORGY, which are app. ORGX=NX/2, ORGY=NY/2
ORGX=2028 ORGY=1996 !Detector origin
```

```
!====================== JOB CONTROL PARAMETERS ================================
JOB= ALL !XYCORR INIT COLSPOT IDXREF DEFPIX XPLAN INTEGRATE CORRECT

!====================== GEOMETRICAL PARAMETERS ================================
DETECTOR_DISTANCE= 125.0    !(mm)

ROTATION_AXIS=1.0 0.0 0.0
OSCILLATION_RANGE=1.0             !degrees (>0)

X-RAY_WAVELENGTH=1.9          !Angstrom
INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0
!FRACTION_OF_POLARIZATION=0.99 !default=0.5 for unpolarized beam;0.90 at DESY;
POLARIZATION_PLANE_NORMAL= 0.0 1.0 0.0
!AIR=0.001     !Air absorption coefficient of x-rays as computed by XDS

!====================== CRYSTAL PARAMETERS ================================
SPACE_GROUP_NUMBER=0   !0 for unknown crystals; cell constants are ignored.
UNIT_CELL_CONSTANTS= 0 0 0 0 0 0

! You may specify here the x,y,z components for the unit cell vectors if
! known from a previous run using the same crystal in the same orientation
!UNIT_CELL_A-AXIS=
!UNIT_CELL_B-AXIS=
!UNIT_CELL_C-AXIS=

!Optional reindexing transformation to apply on reflection indices
!REIDX=    0   0 -1   0   0 -1   0   0 -1   0   0   0

FRIEDEL'S_LAW=FALSE !Default is TRUE.

!==================== SELECTION OF DATA IMAGES ================================
!Generic file name, access, and format of data images
NAME_TEMPLATE_OF_DATA_FRAMES=../../../Robin1/040707-8_2_1_1.???? DIRECT  TIFF

DATA_RANGE=1  360        !Numbers of first and last data image collected

 BACKGROUND_RANGE=1 5  !Numbers of first and last data image for background

 SPOT_RANGE=1  180         !First and last data image number for finding spots

!==================== DATA COLLECTION STRATEGY (XPLAN) ======================
!                        !!! Warning !!!
! If you processed your data for a crystal with unknown cell constants and
! space group symmetry, XPLAN will report the results for space group P1.

!STARTING_ANGLE=  0.0      STARTING_FRAME=1
!used to define the angular origin about the rotation axis.
!Default:  STARTING_ANGLE=  0 at STARTING_FRAME=first data image

!RESOLUTION_SHELLS=10 6 5 4 3 2 1.5 1.3 1.2

!STARTING_ANGLES_OF_SPINDLE_ROTATION= 0 180 10

!TOTAL_SPINDLE_ROTATION_RANGES=30.0 120 15

!REFERENCE_DATA_SET= CK.HKL   !Name of a reference data set (optional)

!====================== INDEXING PARAMETERS ================================
!Never forget to check this, since the default 0 0 0 is almost always correct!
!INDEX_ORIGIN= 0 0 0          ! used by "IDXREF" to add an index offset

!Additional parameters for fine tuning that rarely need to be changed
!INDEX_ERROR=0.05      ! Maximum allowed deviation from 'integerness'
INDEX_MAGNITUDE=8      ! Maximum magnitude of index differences between
                              reflections
INDEX_QUALITY=0.8      ! Minimum quality of indices required for a reflection to
                             be included in the shortest tree
!SEPMIN=6.0  ! Minimum distance (pixels) between diffraction spots considered when
                    !looking for difference vector clusters
CLUSTER_RADIUS=3       ! Maximum radius of a difference vector cluster
!MAXIMUM_ERROR_OF_SPOT_POSITION=3.0   ! Maximum acceptable deviation (pixel units)
```

```
                    !between observed and calculated location of a diffraction peak

!============== DECISION CONSTANTS FOR FINDING CRYSTAL SYMMETRY =============
!Decision constants for detection of lattice symmetry (IDXREF, CORRECT)
MAX_CELL_AXIS_ERROR=0.03 ! Maximum relative error in cell axes tolerated
MAX_CELL_ANGLE_ERROR=2.0 ! Maximum cell angle error tolerated

!Decision constants for detection of space group symmetry (CORRECT).
!Resolution range for accepting reflections for space group determination in
!the CORRECT step. It should cover a sufficient number of strong reflections.
TEST_RESOLUTION_RANGE=8.0 4.5
MIN_RFL_Rmeas= 50 ! Minimum #reflections needed for calculation of Rmeas
MAX_FAC_Rmeas=2.0 ! Sets an upper limit for acceptable Rmeas

!================= PARAMETERS CONTROLLING REFINEMENTS ======================
!REFINE(IDXREF)=BEAM AXIS ORIENTATION CELL !DISTANCE
!REFINE(INTEGRATE)=!DISTANCE BEAM ORIENTATION CELL !AXIS
!REFINE(CORRECT)=DISTANCE BEAM ORIENTATION CELL AXIS

!================== CRITERIA FOR ACCEPTING REFLECTIONS ======================
VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= 6000 30000 !Used by DEFPIX
                                        !for excluding shaded parts of the detector.

INCLUDE_RESOLUTION_RANGE=20.0 2.65 !Angstroem; used by DEFPIX,INTEGRATE,CORRECT

!used by CORRECT to exclude ice-reflections
!EXCLUDE_RESOLUTION_RANGE= 3.93 3.87 !ice-ring at 3.897 Angstrom
!EXCLUDE_RESOLUTION_RANGE= 3.70 3.64 !ice-ring at 3.669 Angstrom
!EXCLUDE_RESOLUTION_RANGE= 3.47 3.41 !ice-ring at 3.441 Angstrom
!EXCLUDE_RESOLUTION_RANGE= 2.70 2.64 !ice-ring at 2.671 Angstrom
!EXCLUDE_RESOLUTION_RANGE= 2.28 2.22 !ice-ring at 2.249 Angstrom
!EXCLUDE_RESOLUTION_RANGE= 2.102 2.042 !ice-ring at 2.072 Angstrom - strong
!EXCLUDE_RESOLUTION_RANGE= 1.978 1.918 !ice-ring at 1.948 Angstrom - weak
!EXCLUDE_RESOLUTION_RANGE= 1.948 1.888 !ice-ring at 1.918 Angstrom - strong
!EXCLUDE_RESOLUTION_RANGE= 1.913 1.853 !ice-ring at 1.883 Angstrom - weak
!EXCLUDE_RESOLUTION_RANGE= 1.751 1.691 !ice-ring at 1.721 Angstrom - weak

!MINIMUM_ZETA=0.01 !Defines width of 'blind region' (XPLAN,INTEGRATE,CORRECT)

!WFAC1=0.75  !This controls the number of rejected MISFITS in CORRECT;
                  !a larger value leads to fewer rejections.

!============== INTEGRATION AND PEAK PROFILE PARAMETERS ====================
!Specification of the peak profile parameters below overrides the automatic
!determination from the images
!Suggested values are listed near the end of INTEGRATE.LP
!BEAM_DIVERGENCE=   0.80         !arctan(spot diameter/DETECTOR_DISTANCE)
!BEAM_DIVERGENCE_E.S.D.=   0.080 !half-width (Sigma) of BEAM_DIVERGENCE
!REFLECTING_RANGE=  0.780 !for crossing the Ewald sphere on shortest route
!REFLECTING_RANGE_E.S.D.=  0.113 !half-width (mosaicity) of REFLECTING_RANGE

!NUMBER_OF_PROFILE_GRID_POINTS_ALONG_ALPHA/BETA=9 !used by: INTEGRATE
!NUMBER_OF_PROFILE_GRID_POINTS_ALONG_GAMMA= 9     !used by: INTEGRATE

!CUT=2.0    !defines the integration region for profile fitting
!MINPK=75.0 !minimum required percentage of observed reflection intensity
!DELPHI= 5.0!controls the number of reference profiles and scaling factors

!======= PARAMETERS CONTROLLING CORRECTION FACTORS (used by: CORRECT) =======
!MINIMUM_I/SIGMA=3.0 !minimum intensity/sigma required for scaling reflections
!NBATCH=-1  !controls the number of correction factors along image numbers
!REFLECTIONS/CORRECTION_FACTOR=50   !minimum #reflections/correction needed
!PATCH_SHUTTER_PROBLEM=TRUE         !FALSE is default
!STRICT_ABSORPTION_CORRECTION=TRUE  !FALSE is default
!CORRECTIONS= DECAY MODULATION ABSORPTION

!=========== PARAMETERS DEFINING BACKGROUND AND PEAK PIXELS ================
!STRONG_PIXEL=3.0                          !used by: COLSPOT
!A 'strong' pixel to be included in a spot must exceed the background
!by more than the given multiple of standard deviations.
```

```
!MAXIMUM_NUMBER_OF_STRONG_PIXELS=1500000       !used by: COLSPOT

!SPOT_MAXIMUM-CENTROID=3.0                     !used by: COLSPOT

!MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=6          !used by: COLSPOT
!This allows to suppress spurious isolated pixels from entering the
!spot list generated by "COLSPOT".

!NBX=3  NBY=3  !Define a rectangle of size (2*NBX+1)*(2*NBY+1)
!The variation of counts within the rectangle centered at each image pixel
!is used for distinguishing between background and spot pixels.

!BACKGROUND_PIXEL=6.0                          !used by: COLSPOT,INTEGRATE
!An image pixel does not belong to the background region if the local
!pixel variation exceeds the expected variation by the given number of
!standard deviations.

!SIGNAL_PIXEL=3.0                              !used by: INTEGRATE
!A pixel above the threshold contributes to the spot centroid
```

Figure 4.4.1: XDS.INP script used for data reduction: The various input parameters are commented to better explain their purpose, those highlighted items signify the eleven minimum input parameters needed to execute the program. There are other detector specific terms which required no editing in this study.

To properly use this script the user must define several parameters contained in the script. XDS reads no header information from images to be processed. I was told to consult an XDSwiki page authored by Dr. Kay Diederichs which advised users on the user of the program ADXV (89). This program reads the header information to determine the ORGX/ORGY beam center, detector distance, oscillation range, wavelength. The detector choice was listed within the script for the SERCAT 22ID line, no information concerning the space group was initially entered, image locations identified, and the Background and Spot range values were default values within the original script offered.

The forthcoming explanations of what values should be used within the various sections of the XDS.INP file were adapted from the release notes of Wolfgang Kabsch, January 20, 2009. It is important to mention however that the XDS.INP file contains many variables which are interrelated such that unknowingly altering may values within a specific sub menus may lead to adverse effects in others.

Finding Spots

The section of the XDS script entitled *SELECTION OF THE DATA IMAGES* uses the keywords **Name_Template_of_Data_Frames, Data_**, **Background_**, and **Spot_Range** for locating of the images to be processed and defining the range of the data set, number of images to be used for background, and spot approximation, respectively.

```
!==================== SELECTION OF DATA IMAGES ==============================
!Generic file name, access, and format of data images
NAME_TEMPLATE_OF_DATA_FRAMES=040707-8_2_1_1.???? DIRECT  TIFF
DATA_RANGE=1  360       !Numbers of first and last data image collected
BACKGROUND_RANGE=1 180  !Numbers of first and last data image for background
SPOT_RANGE=1  180        !First and last data image number for finding spots
```

Figure 4.4.2: Find Spot parameters: A portion of the XDS.INP file dealing with the Finding Spot routine

Further parameter definitions are determined by the portion of the script entitled *PARAMETERS DEFINING BACKGROUND AND PEAK PIXELS.* The first keyword determines the requirement for a **STRONG_PIXEL**. If the intensity of a pixel is above the mean pixel values plus and additional 3σ from the surrounding background pixels, it is considered strong. Instead of the traditional x, y, φ values defining the spot position in 3D, XDS uses x, y, z such that the progression of images are considered part of the z-direction. Any two "strong pixels" found to be adjacent to each other are considered part of the same spot. Adjacent pixels are considered part of the same spot from the original x,y,z of the first pixel in either direction; x+1,y,z; x,y+1,z; x,y,z+1. The coordinates of the spots are defined by z-centroids. XDS does not use x, y, φ to classify the orientation of spots, instead assigns a z-value representing Φ. Therefore, weak reflections will not be used in prediction of spot coordinates during indexing. The aforementioned actions are performed by a subroutine entitled COLSPOT which locates the strong diffraction spots and saves the calculated centroids in the SPOT.XDS file.

The *PARAMETERS DEFINING BACKGROUND AND PEAK PIXELS* menu also has the option to alter the **MAXMIUM_NUMBER_OF_STRONG_PIXELS** value, such that the

weakest of the selected spots are discarded. Also a cut off value can be determined for the number of pixels contained within the selected spots be setting the **MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT** value. Also, the **SPOT_MAXIMUM_CENTROID** keyword is useful for discarding spots if the centroid location exceeds a pre-defined range from the predicted of the strongest pixel within the spot. The **NBX**, **NBY** values are used to define a rectangular array centered on each pixel on the face of the detector in order to detect spot shape from background. The **BACKGROUND** and **SPOT_RANGE** pixel values establish a default threshold to distinguish a pixel to be counted as either spot or background from the NBX, NBY analysis.

```
!=========== PARAMETERS DEFINING BACKGROUND AND PEAK PIXELS =================
!STRONG_PIXEL=3.0                          !used by: COLSPOT
!A 'strong' pixel to be included in a spot must exceed the background
!by more than the given multiple of standard deviations.
!MAXIMUM_NUMBER_OF_STRONG_PIXELS=1500000       !used by: COLSPOT
!SPOT_MAXIMUM-CENTROID=3.0                     !used by: COLSPOT
!MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=6          !used by: COLSPOT
!This allows to suppress spurious isolated pixels from entering the
!spot list generated by "COLSPOT".
!NBX=3  NBY=3  !Define a rectangle of size (2*NBX+1)*(2*NBY+1)
!The variation of counts within the rectangle centered at each image pixel
!is used for distinguishing between background and spot pixels.
!BACKGROUND_PIXEL=6.0                          !used by: COLSPOT,INTEGRATE
!An image pixel does not belong to the background region if the local
!pixel variation exceeds the expected variation by the given number of
!standard deviations.
!SIGNAL_PIXEL=3.0                              !used by: INTEGRATE
!A pixel above the threshold contributes to the spot centroid
```

Figure 4.4.3: Spot definition parameters: A portion of the XDS.INP file dealing with the Spot selection parameters

Indexing

Before auto-indexing can be performed, three subroutines are required: XYCORR, INIT, and COLSPOT all of which are governed by keywords within the *INDEXING PARAMETERS* menu. The XYCORR creates spatial correction tables wherein correction values in x and y can be accessed when observed coordinates of a pixel array, in relation to the laboratory coordinates, are offset and in need of adjustment. INIT determines three additional tables for classifying

background from spots or strong pixels. These are the tables: BLANK.CBF, GAIN.CBF, and BKGINIT.CBF. The Blank.CBF table relies on a non-X-ray background image (dark current). If a dark current image is not available, the table is generated from the OFFSET parameter which specifies the **INDEX_ORIGIN** as default = 0 0 0, or a constant value estimation for detector noise using the mean values at the four corners of several images. The **INDEX_ERROR**, **_MAGNITUDE**, and **_QUALITY** keywords account for the maximum allowable deviation from intergerness for the **h**, **k**, **l** values, differences between reflection magnitudes, and the quality of the **h**, **k**, **l** values to be included in the indexing algorithm, respectively.

```
!====================== INDEXING PARAMETERS ==================================
!Never forget to check this, since the default 0 0 0 is almost always correct!
!INDEX_ORIGIN= 0 0 0          ! used by "IDXREF" to add an index offset

!Additional parameters for fine tuning that rarely need to be changed
!INDEX_ERROR=0.05      ! Maximum allowed deviation from 'integerness'
INDEX_MAGNITUDE=8      ! Maximum magnitude of index differences between
                         reflections
INDEX_QUALITY=0.8      ! Minimum quality of indices required for a reflection to
                       !be included in the shortest tree
!SEPMIN=6.0  ! Minimum distance (pixels) between diffraction spots considered when
             !looking for difference vector clusters
CLUSTER_RADIUS=3       ! Maximum radius of a difference vector cluster
!MAXIMUM_ERROR_OF_SPOT_POSITION=3.0   ! Maximum acceptable deviation (pixel units)
                    !between observed and calculated location of a diffraction peak
```

Figure 4.4.4: Indexing Parameters: Highlights several of the parameters which are used for the creation of cluster vectors

The Gain.CBF attempts to account for the variation of pixel contents within the background region of the data. The file assist in distinguishing “strong pixels” from background pixels via a box constructed from (2*NBX+1) by (2*NBY+1) which is oriented at the center of each pixel of the images searched for background determination and the variation of pixel values therein. In the absence of any spot the values in this table are used to estimate the pixel variation to distinguish “strong pixels” from background pixels. The default vales of NBX and NBY are 3 and 3 respectively. Finally, the Bkginit.CBF tables utilize the user input for images included within the **BACKGROUND_RANGE** to estimate the global background for data processing.

The X-ray background from each image is added to best account for variations through the data set, this includes regions of the detector such as beam stops or other user defined areas of the detector. The function of COLSPOT highlighted in the earlier section on finding spots is used to locate both "strong pixels" and diffraction spots specified by the input parameters. The output is saved in a file titled SPOT.XDS

A portion of the spots within the SPOT.XDS file are used to discern the orientation matrix, cell constants, and symmetry of the crystal lattice. The diffracted beam wave vector responsible for diffraction spots allows for the formulation of the laboratory coordinates using the input values. The keywords used in this calculation belong to the *DETECTOR* and *GEOMETRICAL PARAMETERS* menu are **QX** and **QY**, which identify the detector parameters. **X-ray Wavelength** identifies the wavelength used for data collection. **DIRECTION_OF_DETECTOR_X-AXIS**, is a matrix comprised of orthonormal vectors which denote the orientation of the detector in respect to the laboratory coordinate system. **DETECTOR_DISTANCE** simply states the distance from sample to detector. **ORGX** and **ORGY** denote the location of the minimum distance between the detector and the crystal or the origin (0,0,0) at which the direct beam impacts the detector face.

The difference between the unit vectors along the incident and reflected beam result in a reciprocal lattice vector. This vector can also be found for the stationary crystal from the centroid information within the SPOT.XDS file in conjunction with the following: **ROTATION_AXIS** (vector) describes the directional cosines of the rotation axis versus the laboratory system. A default value is provided by the XDS.inp file 0.0 1.0 0.0 which is interpreted to mean the crystal would rotate clockwise as data collection proceeds from a detector orientation/view point. **STARTING_ANGLE/FRAME** combined with **OSCILLATION_RANGE** define the φ angle

of the crystal rotation. The *INDEXING PARAMETERS* subroutine uses only those reciprocal lattice vectors which satisfies the user defined minimal length differences, **SEPMIN**. These vectors are represented in a 3 dimensionalhistogram, which will result in clusters of the vectors since many pairs are nearly identical in difference. From the vector clusters a maximum is located by either default or user cutoffs termed **CLUSTER_RADIUS**. A basis set of 3 independent linear cluster vectors are chosen which are used to express the remaining clusters as integral multiples in respect to the original choice. The basis vectors of the top 60 clusters are listed in the IDXREF.LP file. If the user has input, the known space group and cell constants, the clustered vectors are interpreted with respect to the provided parameters. If the user provides no space group infromation, XDS uses a reduced triclinic cell and all parameters are recorded in the IDXREF.LP file. With the new parameters of the unit cell and space group identified, by default, up to 3,000 of the strongest spots are used in local indexing. This process groups spots into nodes based on the best fit between location of spots and predetermined unit cell characteristics. Reflections from the most qualified node, those exhibiting the greatest number of integer indices, are used in refining the basis vectors. The DEFPIX command removes detector regions outside the detector range. Once initial refinement of the cell parameters is complete, more spots are added to the "refinement queue" only rejecting those spots (corresponding to lattice vectors) which do not fit the accepted unit cell parameters. Another means of removing unwanted peaks uses the **MAXIMUM_ERROR_OF_SPOT_POSITION**, which affects the allowable deviation between predicted and observed spot peaks.

Next XDS uses the *DECISION CONSTANTS FOR FINDING CRYSTAL SYMMETRY* menu to search the 44 available lattice types for the best fit using two key words, **MAX_CELL_AXIS_ERROR** and the **MAX_CELL_ANGLE_ERROR.** These values are

used to weed out those symmetries which offer the best fit within triclinic symmetry. The remaining keywords determine the acceptable ranges for determining the Rmeas (R-measure) statistics, **MINIMUM_RFL_Rmeas** and **MAXIMUM_FAC_Rmeas**. R_{meas} is a corrected R-factor which XDS uses as an indicator of diffraction quality. Traditionally R_{sym}, which is commonly used interchangeably with the term R_{merge}, is widely accepted as the preferred statistic for judging merit of scaled data (90, 91). R_{sym} contains an implicit redundancy dependence, which affects its magnitude such that less data can make this value appear better. Mathematical and empirical arguments have been presented against the use of R_{sym} in lieu of R_{meas} (87), for the purpose of this work the terms used are not important as the final quality check for each program will depend on traceability of each result and the resultant quality of phases generated.

```
!============== DECISION CONSTANTS FOR FINDING CRYSTAL SYMMETRY =============
!Decision constants for detection of lattice symmetry (IDXREF, CORRECT)
MAX_CELL_AXIS_ERROR=0.03 ! Maximum relative error in cell axes tolerated
MAX_CELL_ANGLE_ERROR=2.0 ! Maximum cell angle error tolerated

!Decision constants for detection of space group symmetry (CORRECT).
!Resolution range for accepting reflections for space group determination in
!the CORRECT step. It should cover a sufficient number of strong reflections.
TEST_RESOLUTION_RANGE=8.0 4.5
MIN_RFL_Rmeas= 50 ! Minimum #reflections needed for calculation of Rmeas
MAX_FAC_Rmeas=2.0 ! Sets an upper limit for acceptable Rmeas
```

Figure 4.4.5: The portion of the XDS.INP script which sets indexing constants

There are, however, several errors that can cause incomplete or poor processing pertaining to crystal symmetry. XDS does an excellent job of posting error messages and possible resolutions at the completion of each subroutine. Various parameters can be refined during indexing in an effort to improve the output such as the beam center (**BEAM**), rotation axis (**AXIS**), unit cell orientation (**CELL**), orientation matrix (**ORIENTATION**), and the detector distance (**DISTANCE**). When first processing the data in XDS, including distance in the refinement is not encouraged. The data should be processed with this refinement parameter included only to determine which method produces more favorable results.

```
!================= PARAMETERS CONTROLLING REFINEMENTS ======================
REFINE(IDXREF)=BEAM AXIS ORIENTATION CELL !DISTANCE
REFINE(INTEGRATE)=!DISTANCE BEAM ORIENTATION CELL !AXIS
REFINE(CORRECT)=DISTANCE BEAM ORIENTATION CELL AXIS
```

Figure 4.4.6: Refinement parameters used during Indexing, Integration and Scaling

Integration

XDS records and saves the observed intensities, the corresponding standard deviations, and location of each predicted spot peak throughout the rotational data images in an INTEGRATE.HKL file. The parameters needed for predicting spot locations are provided in the XPARM.XDS file. These parameters can be refined by referencing strong spots as integration commences. Just as in indexing, the user can also determine which factors will be refined during integration such as the detector distance (**DISTANCE**), beam center (**BEAM**), rotation axis (**AXIS**), orientation matrix (**ORIENTATION**), and unit cell orientation (**CELL**). This is accomplished with in the **Refine (Integrate)=** keywords, including the use of the keyword **ALL** to encompass all possible refinement parameters (Figure 4.4.7).

The integration process is commonly known as Kabsch profile fitting. This is a 3 dimensional integration process nearly identical to the integration engine used in PROTEUM2. As stated earlier the process used during integration by d*TREK are also considered 3 dimensional, but this method resembles Kabsch profile fitting only in the fact it to processes full reflections not partials located on a varying number of images. The Process involves isolating each reflection projected onto the Ewald sphere and further enclosing each spot by a 9x9x9 pixel array serving as a integration box (55).

To accomplish this, the Kabsch algorithm first assigns h, k, l indices to every pixel on the surface of the detector based on the nearest reflection it may belong to. The indice locations are cross-referenced with known intensity locations based on the pre-determined space group during

indexing. Any pixels not obeying the limiting conditions pertaining to possible reflection location relegated to background values. Those pixels, which are sufficiently close to known spot locations, are recoded and their distances from the Ewald sphere are calculated. These remaining pixels are subject to finer inspection of their locations versus the known distances between the detector and the Ewald sphere. Pixels that fail this test are, just as before, relegated to background. Those, which perform to this test, are considered "spot". This process in discussed in greater detail as described by (55).

Those pixels contained in the area considered "spot" are analyzed according to the parameters in the subroutine, *INDEXING PARAMETERS*. The pixels identified as strong are linked to other strong pixels by analyzing all surrounding pixels according to σ cut offs, which will outline the spot shape in 2 dimensions per image. Since every pixel on the detector face is considered a part of the closest reflection there exist no possibility of spots being considered only once. Once spot positions have been properly identified the intensities are calculated for each reflection via background estimation and 3 dimensional profile fitting. The average profiles created are collected from strong reflections forming a grid such that a threshold is established for profile determination noted as CUT within the XDS.INP. The detector is divided into 9 equal parts and the standard profiles calculated from these areas are refined every 5 images. Profile fitting ensures the habit of diffraction spots within the same area of the detector are taken into account when fitting spot shape to integration peaks. The intensity estimation calculated within XDS is estimated as;

$$I = \frac{\sum_{i \in D} (c_i - b_i) \frac{p_i}{v_i}}{\sum_{i \in D} \frac{p_i^2}{v_i}}$$

Which is further minimized to

$$\Psi(I) = \sum_{i \in D} \left(c_i - I \bullet p_i - b_i\right)^2 \Big/ v_i \; ; \text{ such that } \sum_{i \in D} p_i = 1.$$

Where c_i is measured contents, I the intensity to be determined, p_i expected fraction in pixel I from profile fitting, b_i the background of each pixel, and v_i variance of pixel. The summation factor $i \epsilon D$ represents an expected intensity distribution of an observed profile. The variance of the pixels are determined iteratively stating at $v_i = b_i + \mathrm{I}\, p_i$ (92). The *INTEGRATION AND PEAK PROFILE PARAMETERS* portion of the XDS script describes the various parameters dealing with the profile fitting portion of integration.

```
!============== INTEGRATION AND PEAK PROFILE PARAMETERS ======================
!Specification of the peak profile parameters below overrides the automatic
!determination from the images
!Suggested values are listed near the end of INTEGRATE.LP
!BEAM_DIVERGENCE=   0.80         !arctan(spot diameter/DETECTOR_DISTANCE)
!BEAM_DIVERGENCE_E.S.D.=   0.080 !half-width (Sigma) of BEAM_DIVERGENCE
!REFLECTING_RANGE=  0.780 !for crossing the Ewald sphere on shortest route
!REFLECTING_RANGE_E.S.D.=  0.113 !half-width (mosaicity) of REFLECTING_RANGE

!NUMBER_OF_PROFILE_GRID_POINTS_ALONG_ALPHA/BETA=9 !used by: INTEGRATE
!NUMBER_OF_PROFILE_GRID_POINTS_ALONG_GAMMA= 9     !used by: INTEGRATE

!CUT=2.0    !defines the integration region for profile fitting
!MINPK=75.0 !minimum required percentage of observed reflection intensity
!DELPHI= 5.0!controls the number of reference profiles and scaling factors
```

Figure 4.4.7: Integration parameters used by XDS

For each reflection the Kabsch algorithm estimates the background and assembles the 3 dimensional profile from frames contributing to a single spot. Those pixels which register a higher intensity than established background without surrounding pixel intensity, which would be assessed as a possible spot, are considered overlap and rejected. As a method of crosschecking for those pixels considered either background or spot, an additional refinement pass is conducted for each reflection checking from a theoretical profile against a predetermined threshold **MINPK** within the *INTEGRATION AND PEAK PROFILE PARAMETERS* menu. If the integrated intensity is below this threshold, the spot is disregarded; otherwise, the spot is

considered satisfactory data (93). XDS does not attempt to center integration boxes based on spot position. Instead, attempts are made to minimize errors in box placement compared to spot centroid location through refinement. This differs from d*TREK, HKL and MOSFLM methodologies which use the spot habit to choose the proposed centroid location as a basis to begin peak pixel searching within the spot.

Scaling

The integration engine within XDS does not produce an output which is currently compatible with 3DSCALE. Therefore, the CORRECT or scaling portion of the XDS package is used to accomplish this task. The *PARAMETERS CONTROLLING CORRECTION FACTORS* are primarily contained within the following:

```
!======= PARAMETERS CONTROLLING CORRECTION FACTORS (used by: CORRECT) =======
!MINIMUM_I/SIGMA=3.0 !minimum intensity/sigma required for scaling reflections
!NBATCH=-1  !controls the number of correction factors along image numbers
!REFLECTIONS/CORRECTION_FACTOR=50   !minimum #reflections/correction needed
!PATCH_SHUTTER_PROBLEM=TRUE         !FALSE is default
!STRICT_ABSORPTION_CORRECTION=TRUE  !FALSE is default
!CORRECTIONS= DECAY MODULATION ABSORPTION


!================== CRITERIA FOR ACCEPTING REFLECTIONS ======================
!MINIMUM_ZETA=0.01 !Defines width of 'blind region' (XPLAN,INTEGRATE,CORRECT)
!WFAC1=0.75  !This controls the number of rejected MISFITS in CORRECT;
                  !a larger value leads to fewer rejections.
 INCLUDE_RESOLUTION_RANGE=20.0 2.65 !Angstrom; used by DEFPIX,INTEGRATE,CORRECT
```

Figure 4.4.8: Scaling parameters used by XDS

The final two key words within in the *DECISION CONSTANTS FOR FINDING CRYSTAL SYMMETRY* assist in the calculation of the R_{meas} value. The **MIN_RFL_Rmeas** keyword determines the minimum number of reflections required for R_{meas} calculation and **MAX_FAC_Rmeas** sets the upper range for allowable R_{meas} values. The CORRECT step is the finial process conducted by XDS.INP file. The output file is a XDS_ACSII.HKL file containing

the scaled intensities. The statistics generated from XDS scaling are stored in the CORRECT.LP file.

Additional Refinement

There is an additional refinement step which can "polish-off" the data reduction outcome. Initial integration is performed in the default triclinic space group. This can be easily changed to the correct space group by substituting the unit cell information generated by the CORRECT step stored in a file entitled GXPARM.XDS. The following script is used to overwrite the contents of XPARAM.XDS with the values contained in GXPARAM.XDS and reprocesses the data beginning at the INTEGRATE step. This refinement is executed using command line format as follows:

```
>cp GXPARM.XDS XPARM.XDS !copies the contents of GXPARM to XPARM
>mv CORRECT.LP CORRECT.LP.old        !renames the CORRECT.LP file
>egrep –v 'JOB|REIDX' XDS.INP > XDS.INP.new                ! inserts the REIDX
                                                    ! information into the new XDS.INP file
>echo "JOB=XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE >CORRECT" > XDS.INP

>echo"JOB=INTEGRATE CORRECT">>XDS.INP ! resets the JOB parameters
                                                        ! for the next round of processing

>cat XDS.INP.new >>XDS>INP                ! Renames the XDS.INP.new file
```

Figure 4.4.9: Polishing refinement reprocessing: Re-executing the XDS.INP file using refined cell parameters from CORRECT step.

This step will integrate and scale the data using the space group and unit cell parameters identified by the Correct step of XDS. This step also retains the output files from the original processing in case the results from the "polishing step" are not desirable.

File Conversion for Structural Studies

The Novel Structure Solution routine within the SGXPro programming pallet does not recognize the output file type from the XDS scaling routine. A series scripts and conversions are necessary to convert scaled intensities from single data sets as well as multiple merged data sets into a file format recognized by SGXPro. These routines are titled XDSCONV and XSCALE.

1)XDSCONV

Given the output from a single or multiple scaled data sets, re-formatting the file begins by using the XDSCONV.INP script. This script initiates conversion of the scaled XDS_ASCII.HKL output file to various user specified formats such as CNS, SHELX, CCP4_F and CCP4_I for structure determination. The file is outlined below and executed using the line command:

>xdsconv

```
!=================== xdsconv ============================
INPUT_FILE=R1_XDS_ASCII 20 2.65  ! specifies the reflection data that XDSCONV should convert
OUTPUT_FILE=temp.hkl CCP4        ! specifies the converted output file for subsequent use by various crystal
                                 ! structure analysis packages
FRIEDEL'S_LAW=FALSE              ! used if h,k,l and –h,-k,-l are expected to have different intensities
MERGE=FALSE                      ! prevents merging of data
WILSON_STATISTICS=FALSE          ! truncated normal distribution instead of a Wilson disribution is used as a prior guess for
                                 ! estimating structure factor amplitudes
```

Figure 4.4.10: XDSCONV initial conversion script: The R1 and R2 XDS_ASCII formatted files from scaling are merged for further conversion. – adapted from XDS input parameters; MPI for Medical Research, Wolfgang Kabsch, 2010

This will create a "temp.mtz" file and a "F2MTZ.INP" script. The F2MTZ file is written as text within the output of XDSCONV. This file can be made into a executable script or executed using command line format as follows;

```
>f2mtz HKLOUT temp.mtz < F2MTZ.INP
>cad HKLIN1 temp.mtz HKLOUT new1.mtz << EOF
>LABIN FILE 1 ALL
>END
>EOF
```

Figure 4.4.11: Polishing refinement reprocessing: Re-executing the XDS.INP file using refined cell parameters from CORRECT step.

This will result in a "new1.mtz" containing HKL, F+, F-, and the corresponding ±σ columns. Next at the command line type:

>uniqueify new1.mtz new2.mtz

This will isolate 5% of the data for use as R_{free}. Now the new2.mtz file can be used in the finial conversion necessary for SGXPro's Novel Structure Solution engine. A program entitled mtz2sca ver0.3 (94)(Grune, 2008) was used to convert properly labeled mtz files , in particular from XDS, into a format known as .sca which can be processed using SGXPro Novel Structure Solution routine.

2) XSCALE - for merging and scaling

When merging the multiple data sets, R1 and R2 in this study, a script entitled XSCALE.INP is used to combine the corresponding XDS_ASCII files. The file is outlined below and executed using the line command:

>xscale

```
!==================== xdsconv ==============================
MAXIMUM_NUMBER_OF_PROCESSORS=8        !if avalible XDS will use multiple cpu's for
                                      !for processors
RESOLUTION_SHELLS=8.0 5.38 4.27 3.73 3.39 3.15 2.96 2.82 2.69 2.65 !reported res. shells
SPACE_GROUP_NUMBER=75                 !Space group an cell constants to be used in scaling
UNIT_CELL_CONSTANTS=52.70 52.70 40.50 90.000 90.000 90.000

MINIMUM_I/SIGMA=3.0                   ! defines strong reflections to be used for scaling
REFLECTIONS/CORRECTION_FACTOR=50 !minimum #reflections /
                                                        !correction_factor
!0-DOSE_SIGNIFICANCE_LEVEL=0.0125

OUTPUT_FILE=R1-2.ahkl !at minimum of f'
FRIEDEL'S_LAW=FALSE    !Default is True
MERGE=FALSE                    !Default is True
STRICT_ABSORPT6ION_CORRECTION=TRUE !FALSE is default
INPUT_FILE=./R1_XDS_ASCII.HKL
INCLUDE_RESOLUTION_RANGE=20 2.65
!CORRECTIONS=DECAY MODULATION ABSORPTION
!STARTING_DOSE=0.0  DOSE_RATE=1.0
CRYSTAL_NAME=a
INPUT_FILE=./R2_XDS_ASCII.HKL
INCLUDE_RESOLUTION_RANGE=20 2.65
!CORRECTIONS=DECAY MODULATION ABSORPTION
!STARTING_DOSE=0.0  DOSE_RATE=1.0
CRYSTAL_NAME=b
```

Figure 4.4.12: XSCALE merging file for combining individual results: The R1 and R2 XDS_ASCII formatted files from scaling are merged for further conversion. – adapted from XDS input parameters; MPI for Medical Research, Wolfgang Kabsch, 2010

At the conclusion of executing the XSCALE script, the output file is converted to a sca file in the same manner as each individual processing conversion mentioned earlier in the **File Conversion** section.

The values from the respective R1, R2, and R1-R2 data set are as follows:

R-factors -

$$R_{sym} = \sum_{h}\sum_{i}\left|I_{hi} - I_{mean}\right| \Big/ \sum_{hi} I_{hi}$$

R_{sym} is the R-factor chosen 3DSCALE to relate differences in symmetry related reflections. This is a measure of the accuracy of the data. The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection.

$$R_{meas} = \sum_{h}\sqrt{n_h/(n_h - 1)}\sum_{i}\left|I_{hi} - I_{mean}\right| \Big/ \sum_{hi} I_{hi}$$

A alternate indicator of data quality proposed by Diederichs and Karplus (95) to remove the redundancy dependence of R_{sym}. This value, R_{meas}, includes a term $\sqrt{[n/(n-1)]}$ which appropriately weights individual reflections (h) according to their multiplicity (n_h).

R1- XDS Scaling Statistics

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected
7.82	1878	251	256	98.0%	2.9%	3.4%
5.58	3360	435	435	100.0%	3.0%	3.8%
4.57	4228	556	556	100.0%	3.3%	3.8%
3.97	4906	656	657	99.8%	4.0%	4.0%
3.55	5235	728	728	100.0%	5.7%	4.7%
3.25	5915	837	837	100.0%	9.0%	7.6%
3.01	6385	906	906	100.0%	16.5%	14.8%
2.81	6727	947	948	99.9%	31.2%	31.4%
2.65	6903	985	992	99.3%	**59.0%**	69.0%
total	45537	6301	6315	**99.8%**	**5.8%**	5.9%

RES LIMIT	COMPARED	I/SIGMA	R-meas	Rmrgd-F	Anomal Corr	SigAno	Nano
7.82	1877	53.03	3.2%	1.4%	74%	1.470	107
5.58	3360	45.57	3.2%	1.5%	61%	1.265	200
4.57	4228	45.66	3.5%	1.4%	49%	1.037	261
3.97	4906	41.46	4.3%	2.1%	17%	1.089	311
3.55	5235	35.33	6.1%	2.8%	38%	1.423	348
3.25	5915	24.35	9.8%	4.2%	23%	1.158	400
3.01	6385	14.28	17.8%	8.9%	35%	1.241	435
2.81	6727	7.36	33.8%	18.2%	60%	1.410	457
2.65	6892	**3.28**	**63.7%**	43.7%	64%	1.274	467
total	45525	**24.59**	**6.3%**	5.8%	53%	1.258	2986

Table 4.4.1: R1 Scaling results from XDS processing: R1 processing yields acceptable R-factor and R_{meas} results for continued processing.

R1 Data Automated tracing results using SGXPro

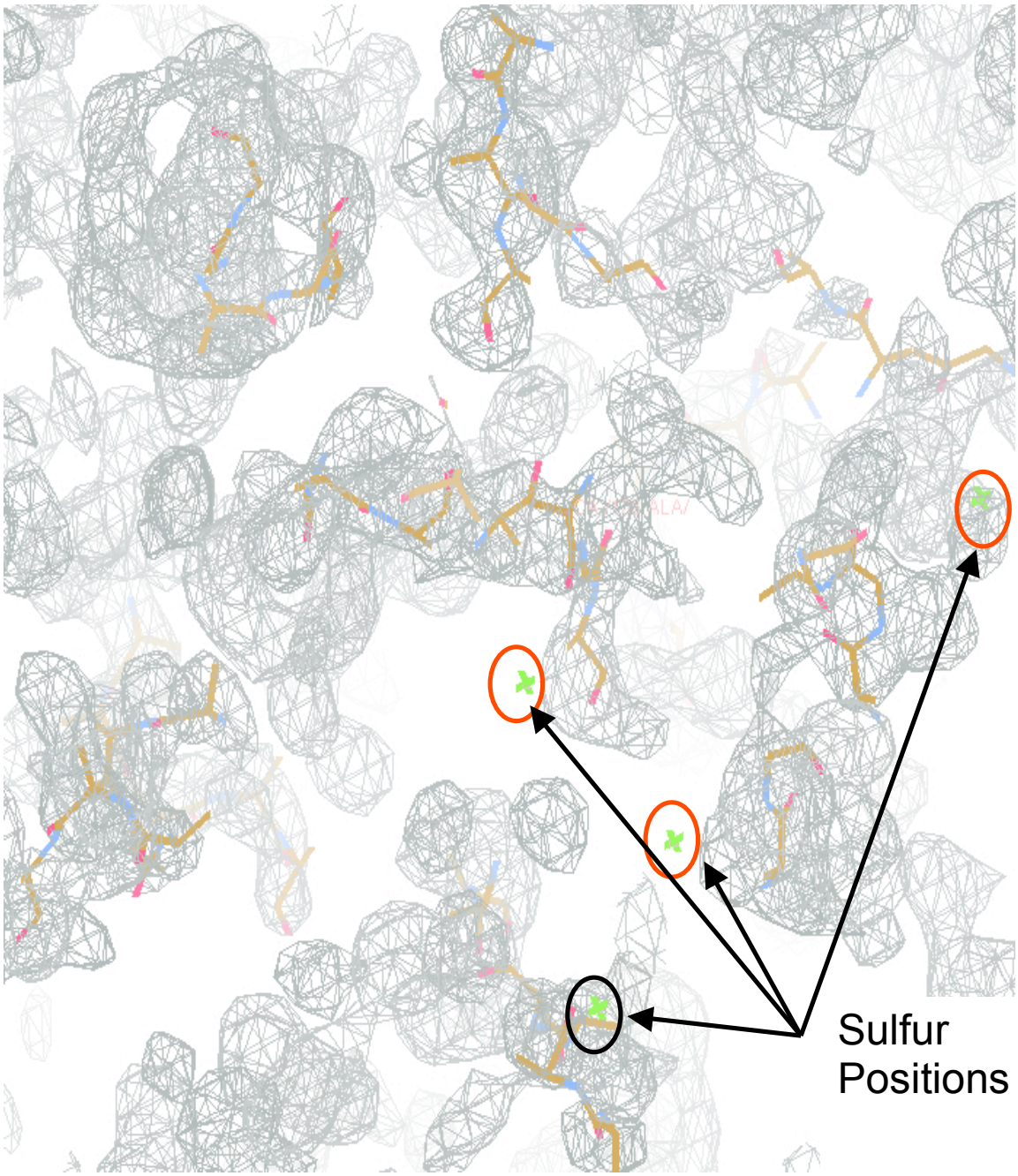

Figure 4.4.13: Tracing result from XDS processing of R1 data set: Three of the four Sulfur positions were correctly identified while an errant Sulfur position identified by SGXPro is circled in black. A total of 62% of amino acids were traced. The trace from RESOLVE did not agree with the refined model.

R1 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
---  -------- -------  --------- -----------
 4      59      12       7 7 6    /XDS-BEST-Period/R1/t2/zzsgxSol_4.pdb
Heavy atoms:                     /XDS-BEST-PERiuod/R1/t2/zzsgxSol_4_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 1 helices 4 amino acids long
#Beta-sheets: no discernible beta sheets
CC_ALL/CC_WEAK    22.32/7.19
PATFOM      31.35
```

Table 4.4.2: R1 Tracing results from SGXPro Novel Structure Solution: R1 data set processing yielded poor CC-ALL/Weak and PATFOM but poor phases and tracing statistics.

R2- XDS Scaling Statistics

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected
7.76	1670	243	256	94.9%	4.3%	5.2%
5.54	3450	445	445	100.0%	3.8%	5.6%
4.54	4421	572	572	100.0%	3.8%	5.6%
3.94	4985	666	667	99.9%	5.2%	5.7%
3.52	5396	764	764	100.0%	8.2%	6.5%
3.22	5684	842	843	99.9%	12.4%	9.4%
2.98	6416	912	912	100.0%	17.7%	14.8%
2.79	6779	979	980	99.9%	29.3%	27.0%
2.63	6281	987	1031	95.7%	**43.5%**	47.0%
total	45082	6410	6470	**99.1%**	**7.1%**	7.4%

RES LIMIT	COMPARED	I/SIGMA	R-meas	Rmrgd-F	Anomal Corr	SigAno	Nano
7.76	1667	33.56	4.7%	2.2%	59%	1.182	102
5.54	3450	32.61	4.1%	1.7%	55%	1.040	204
4.54	4421	33.08	4.1%	1.7%	27%	0.770	270
3.94	4985	30.54	5.6%	2.4%	28%	0.954	315
3.52	5396	26.41	8.9%	3.4%	36%	1.460	365
3.22	5684	19.69	13.5%	4.6%	9%	0.937	402
2.98	6416	13.94	19.2%	6.1%	29%	0.983	440
2.79	6779	8.95	31.8%	9.2%	29%	0.881	472
2.63	6257	**5.18**	**47.3%**	15.8%	9%	0.730	453
total	45055	**19.54**	**7.7%**	4.2%	30%	0.969	3023

Table 4.4.3: R2 Scaling results from XDS processing: R2 processing yields acceptable R-meas results for continued processing.

R2 Data Automated tracing results using SGXPro

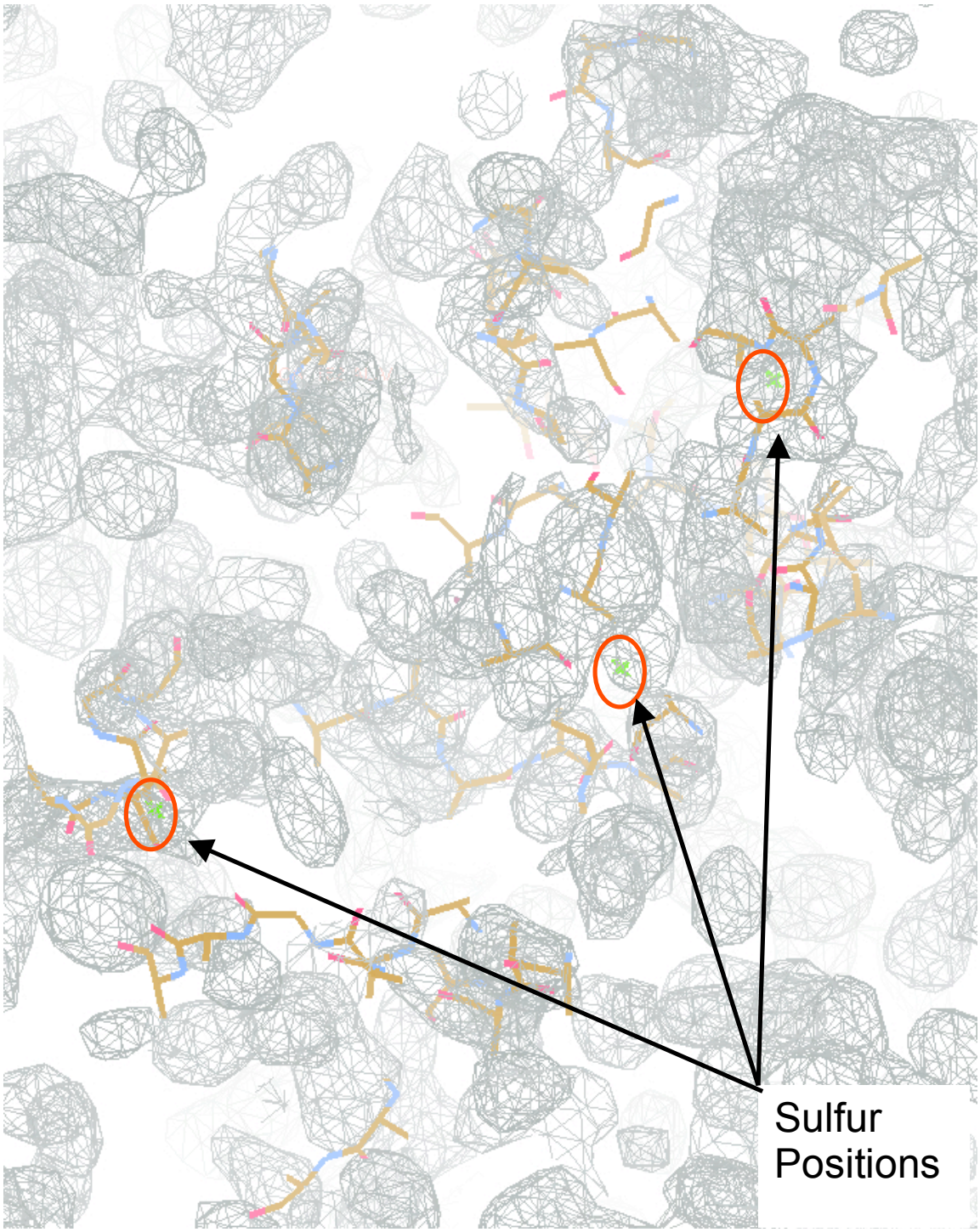

Figure 4.4.14: Tracing result from XDS processing of R2 data set: Three Sulfur positions were identified. A total of 91% of the total amino acids were traced. The trace from RESOLVE did not agree with the refined model.

R2 - Heavy Atom/Tracing statistics using SGXPro

No#	NumBuilt	NumSegs	Top3Segs	Model Built
1	**87**	**17**	**9 7 6**	/XDS/Project_1a/t1/zzsgxSol_1.pdb

Heavy atoms: /XDS/Project_1a/t1/zzsgxSol_1_ha.xyz
Sulfur atoms found: 4
#Alpha-helix: 1 helices 4 amino acids long
#Beta-sheets: no discernible beta sheets
CC_ALL/CC_WEAK 36.22/17.71
PATFOM 58.38

Table 4.4.4: R2 Tracing results from SGXPro Novel Structure Solution: R2 data set processing yielded poor CC-ALL/Weak and PATFOM but poor phases and tracing statistics.

R1-R2 XDS Scaling Statistics

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected
11.79	980	76	79	96.2%	4.4%	4.1%
8.33	1738	119	121	98.3%	3.9%	4.2%
5.89	3066	198	198	100.0%	3.9%	5.0%
4.81	4052	265	265	100.0%	3.8%	4.9%
4.17	4637	305	305	100.0%	4.2%	5.0%
3.73	4666	323	323	100.0%	6.5%	5.5%
3.40	5245	379	379	100.0%	9.3%	8.0%
3.15	5526	396	396	100.0%	15.0%	13.0%
2.95	5814	418	418	100.0%	22.7%	21.3%
2.78	6322	454	454	100.0%	40.0%	43.4%
2.64	4612	445	471	94.5%	**58.8%**	71.6%
total	90555	6418	6451	**99.5%**	**6.9%**	7.1%

RES LIMIT	COMPARED	I/SIGMA	R-meas	Rmrgd-F	Anomal Corr	SigAno	Nano
11.79	979	58.90	4.6%	1.6%	74%	2.056	30
8.33	1738	66.68	4.0%	1.2%	73%	1.514	52
5.89	3066	53.61	4.0%	1.2%	63%	1.525	92
4.81	4052	55.26	4.0%	1.2%	58%	1.208	124
4.17	4637	52.30	4.3%	1.3%	39%	1.047	145
3.73	4666	46.57	6.7%	1.8%	8%	1.055	154
3.40	5245	33.45	9.7%	2.6%	21%	1.095	181
3.15	5526	23.60	15.6%	3.8%	11%	0.871	189
2.95	5814	15.22	23.6%	6.2%	1%	0.829	203
2.78	6322	8.85	41.6%	10.8%	38%	0.792	221
2.64	4589	**4.57**	**61.5%**	19.8%	29%	0.728	194
total	90531	**30.77**	**7.1%**	3.4%	28%	1.004	3029

Table 4.4.5: R1-R2 Scaling results from XDS: R1-R2 data set processing yield acceptable results for continued processing.

R1-R2 Data Automated tracing results using SGXPro

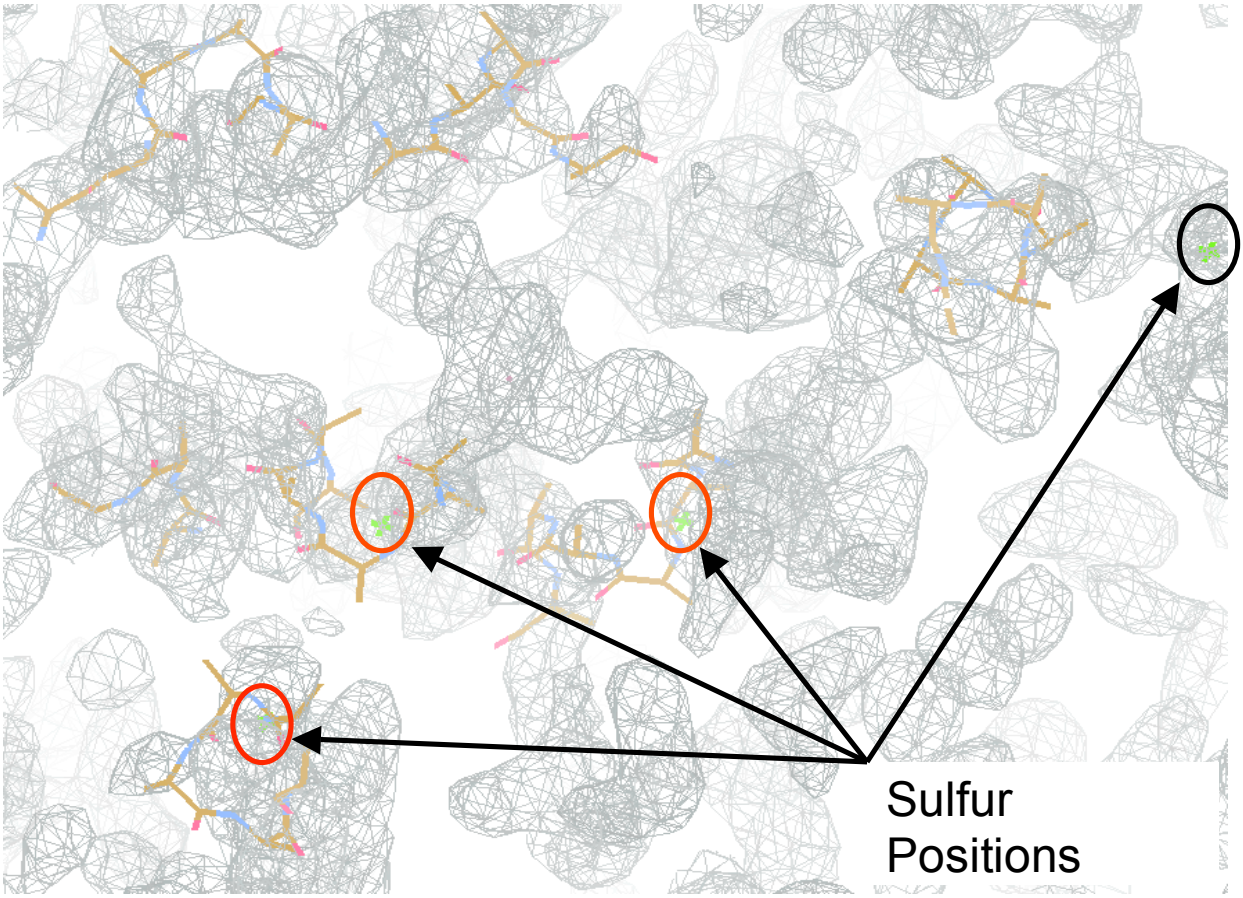

Figure 4.4.15: Tracing result from XDS processing of R1-R2 data set: Three of the four Sulfur positions were correctly identified in Red, while an errant Sulfur position identified by SGXPro is circled in black. A total of 42% of the total amino acids traced. The trace from RESOLVE did not agree with the refined model.

R1-R2 Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------- -----------
 3      40       7       9 6 6        /XDS/Project_1a/t3/zzsgxSol_3.pdb
Heavy atoms:                        /XDS/Project_1a/t3/zzsgxSol_3_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 3 helices 9, 6, 6 amino acids long
#Beta-sheets: no discernible beta sheets
CC_ALL/CC_WEAK    26.84/7.70
PATFOM      65.28
```

Table 4.4.6: R1-R2 Tracing results from SGXPro Novel Structure Solution: R1-R2 data set processing yielded poor CC-ALL/Weak and PATFOM but poor phases and tracing statistics.

From a novice prospective, the XDS data reduction package performs poorly. This is in large part due to the lack of visual aids and the comfort of a GUI interface with which the current generation of crystallographers are accustomed. The program generates error messages, but in the absence of a critical error, from which the program cannot continue, the only means to discerning the effectiveness of your efforts require picking apart log files for relevant statistics. The difficulties inherent to the XDS data reduction program primarily deal with the simplicity of the XDS.INP file. There are approximately 424 possible values and combinations of input parameters which can also be altered in an attempt to optimize processing. Perhaps many of these values should be treated as default. Considering this as a possibility attempts were made to use minimal scripts as instructed by the program authors which were unsuccessful. Although XDS script is documented, the program executes as a black box operation. The user inputs what they believe to be correct and hopes for the best. The purpose of many lines in the XDS.INP script and the various correlations, which exits between many of these values, is not clearly listed in any location I could find. An experimenter who wishes to use XDS will have to rely on experienced users and sparse official documentation, with no official walk-through, in an attempt to understand the eccentricities of the XDS.INP script and its function. These reasons do not

separate XDS from the other processing programs covered in this work as novice friendly. As mentioned earlier XDS, and PROTEUM2 utilize the Kabsch integration technique or a variation of the same. I believe that concise indexing and the use of 3 dimensional integration will be better suited for S-SAD phasing due to the low percent contribution, 1-2%, of Sulfur from the intensity of each diffraction spot. The methods used by Kabsch involving indexing, refinement and integration are intriguing yet my novice approach is in all likelihood unable to fully utilize the depth of the program. I conclude my efforts with XDS with the hope that GUI development and perhaps a more elementary user guide will eventually allow a novice to better utilize XDS for medium resolution S-SAD data.

4.5 PROTEUM2

PROTEUM2 is the only data reduction program covered in this study initially designed for the data collection and processing using a specific detector (34). Continued development and redesign has enabled PROTEUM2 to process data collected on detectors other than the Bruker SMART 6000 detector. This program offers users an excellent GUI interface which is as easy to understand as the MOSFLM GUI with conveniently adjustable parameters within a flowchart as seen in d*TREK. The individual buttons within the program such as Harvest Spots, Index, Bravais and Refine only become active after the appropriate information has been added to the system. The user also has the option to choose between algorithms pertaining to indexing such as FFT, Difference Vectors, or least squares methods. During integration the user can impose various minimum and maximum values on diffraction intensities determining weather the summation method or profile fitting would be used. Another differentiating factor between PROTEUM2 and other data reduction programs hinge on the ease of adjusting parameters, harvesting, indexing, and integration functions. PROTEUM2 also offers the user more intuitive visual aids during data reduction. These refinement tools are easily mastered and used in evaluating spot profiles and quality of indexing. An additional point of interest is program stability. It is commonly known that if errors occur during data reduction the user often needs to restart the program and reprocess the data to effectively. This lack of stability was found to be most prevalent in HKL and MOSFLM, but does occur sporadically in d*TREK while attempting to re-initiate spot finding, indexing or refinement processes in an effort to improve results. PROTEUM2 proved quite stable throughout the data reduction process which was convenient for retracing steps to confirm the parameters used.

Finding Spots

The data used in this study were collected at APS sector 22ID. As stated earlier, PROTEUM2 was initially developed for use with home source detectors therefore image conversion is required. The necessary modifications to convert images are conducted by a sub-routine within PROTEUM2 entitled Unwarp, located within the Instrument tab. The values necessary for this conversion essentially adjust the images collected on a Mar 300 CCD detector at APS into more manageable frames by altering various angles and intensity measures.

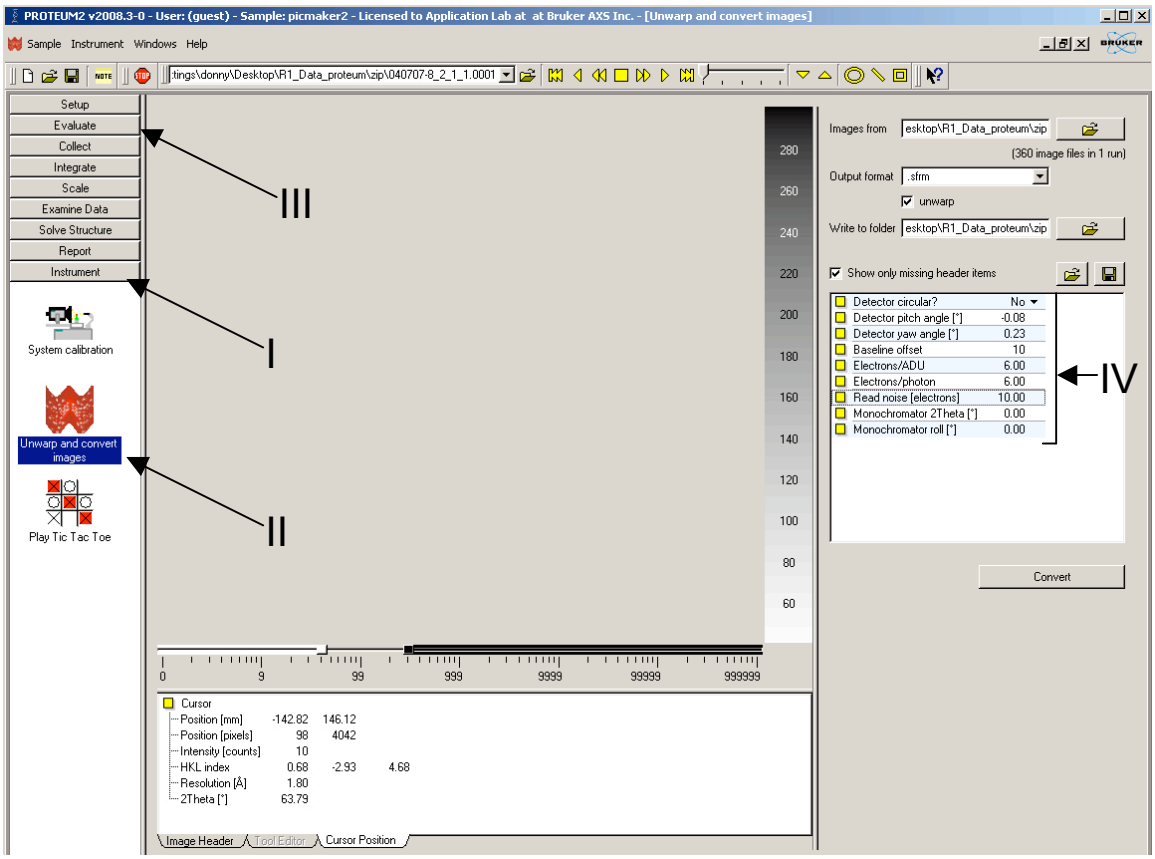

Figure 4.5.1: Unwarp Image Conversion: Unwarp menu for file conversion (I) Instrument Tab, (II) Unwarp and convert images subroutine, (III) Evaluate Tab, and (IV) conversion parameters for 22ID line at Argonne National Laboratory, APS.

Following conversion, the Evaluate tab is selected from the flow chart on the left of the PROTEUM2 GUI and Determine Unit Cell subroutine was chosen. From this interface the newly Unwarpped images are loaded and the newly active Harvest Spots button is selected.

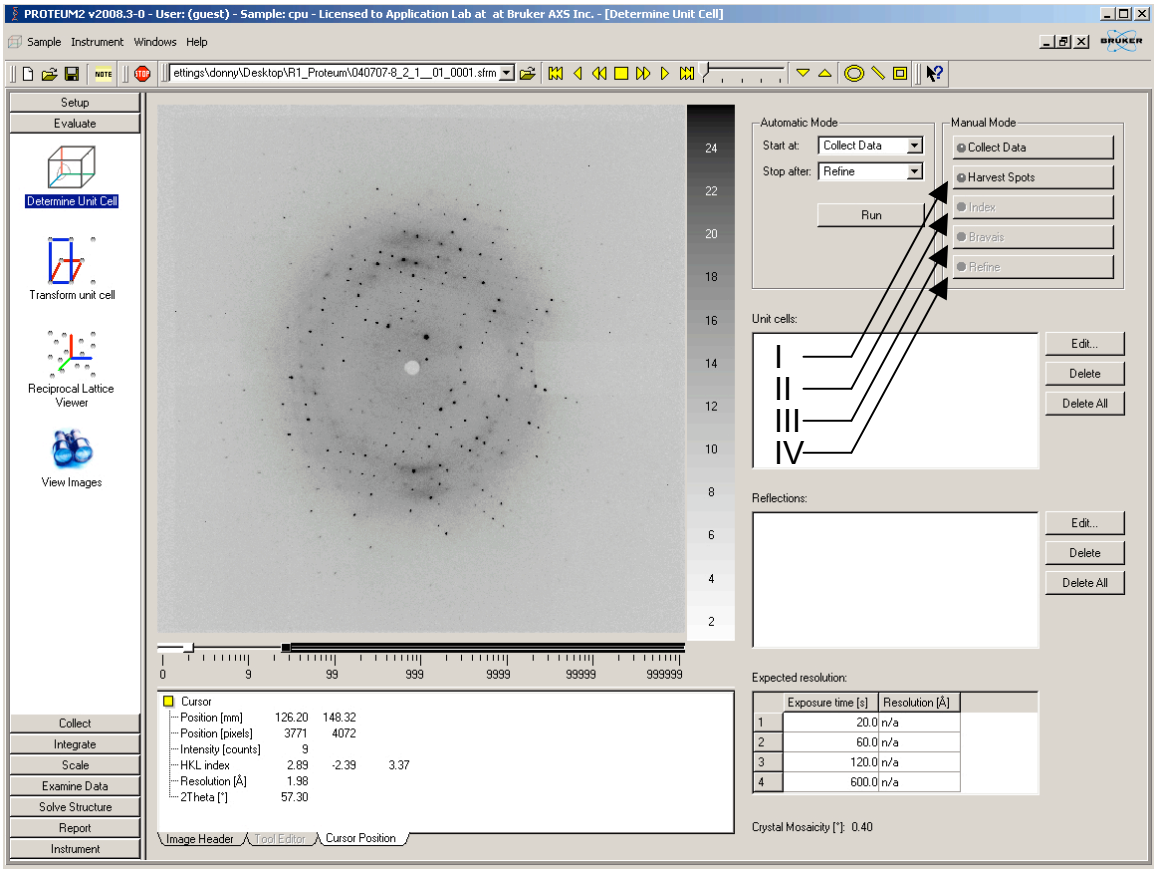

Figure 4.5.2: Initial PROTEUM2 Data Reduction GUI: Each portion of the data reduction routine is highlighted, (I) Harvest Spots, (II) Index, (III) Bravais, and (IV) Refine.

The Harvesting portion of PROTEUM2 attempts to identify reflections within chosen images. The interface has a simple layout in which the user can easily increase or decrease the minimum I/σ_I cutoff for spot selection. This is an adjustable criteria based on pixel intensity. The spots selected for Harvesting are displayed by green circles.

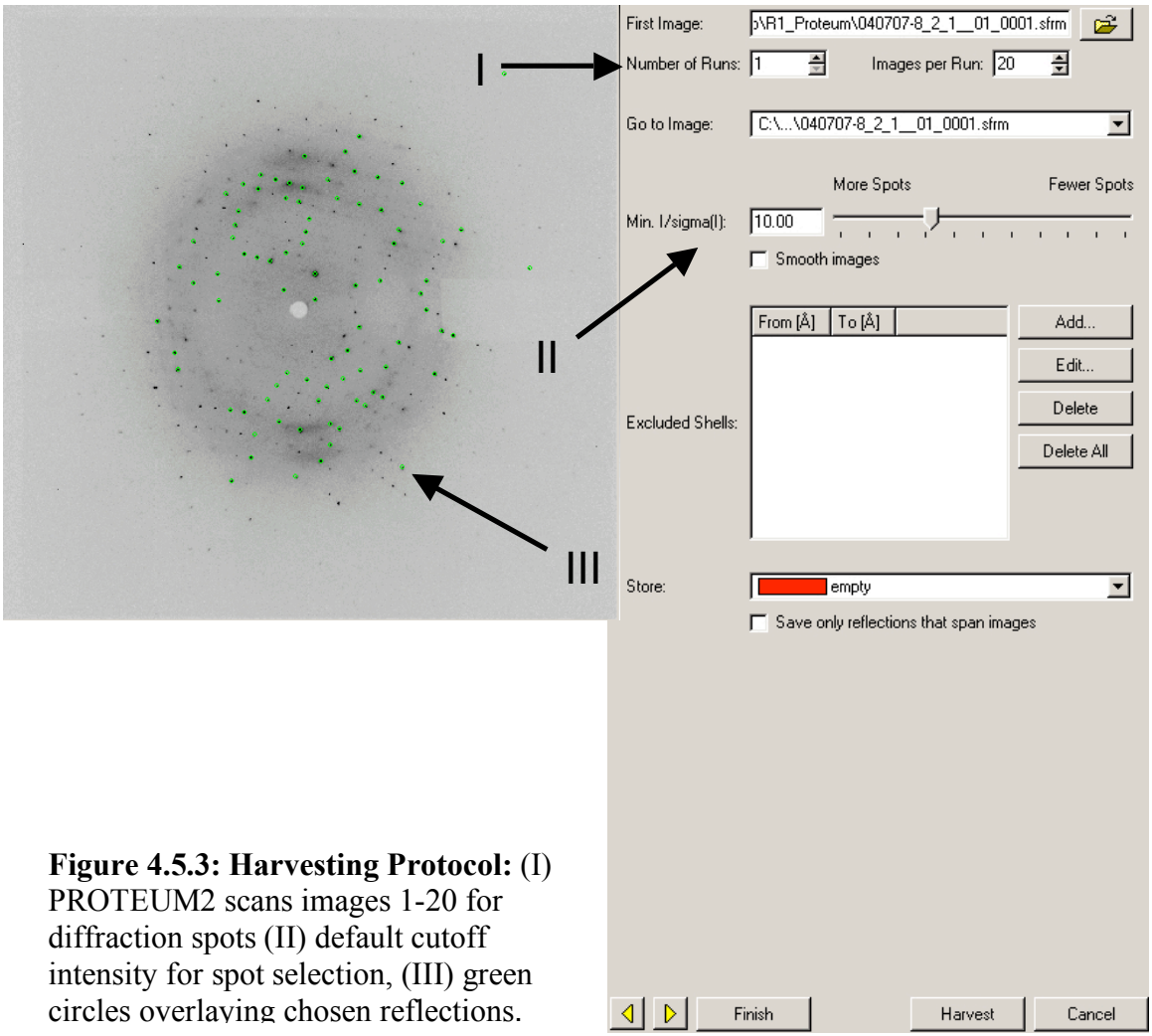

Figure 4.5.3: Harvesting Protocol: (I) PROTEUM2 scans images 1-20 for diffraction spots (II) default cutoff intensity for spot selection, (III) green circles overlaying chosen reflections.

All reflections initially chosen by the spot harvesting program are to be encompassed by green circles. It is not prescribed to use the Smooth Images function when dealing with proteins and although rarely used in initial indexing, the user has the option to exclude resolution shells from the harvesting procedure. Unless the spot harvesting process produced unsuccessful mapping, such as no spots identified, it is advisable to accept the default results and proceed to indexing by selecting Harvest at the bottom of the GUI.

Indexing

After harvesting spots the indexing option becomes active (Figure 4.5.2). There are several methods of spot filtering available to the user to determine a subset of the total spot harvesting which should be used for indexing. These include ensuring the spots are either isolated or span multiple images, and must be whole.

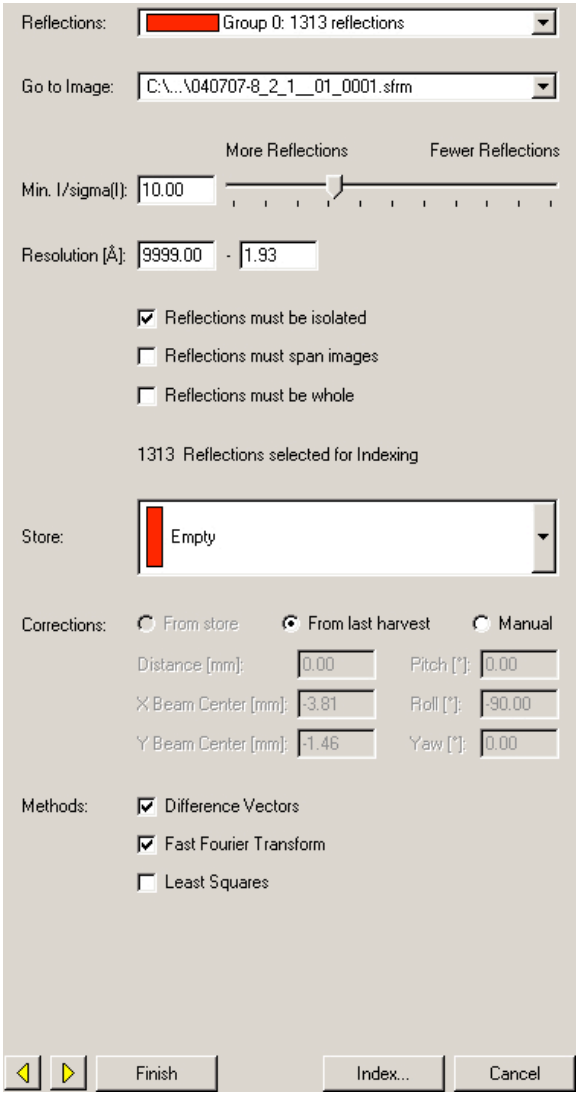

Figure 4.5.4: PROTEUM2 Indexing Menu: A simplified indexing GUI which allows the user to select the methods of indexing to be used during data reduction.

The latter two options, Reflections must – be whole / span images, are not advised by the PROTEUM2 manual for initial indexing unless the user has previous information pertaining to

mosacity. The user does have the option to determine the I/σ_I cutoff for choosing spots to be indexed. By default the PROTEUM2 package will index the data using both Difference vectors, SMART algorithm(96), and FFT, DENZO/DPS algorithims (78, 97) reviewed in Chapter 2 section 2.1. If desired, the user can also select Least Squares method for indexing.

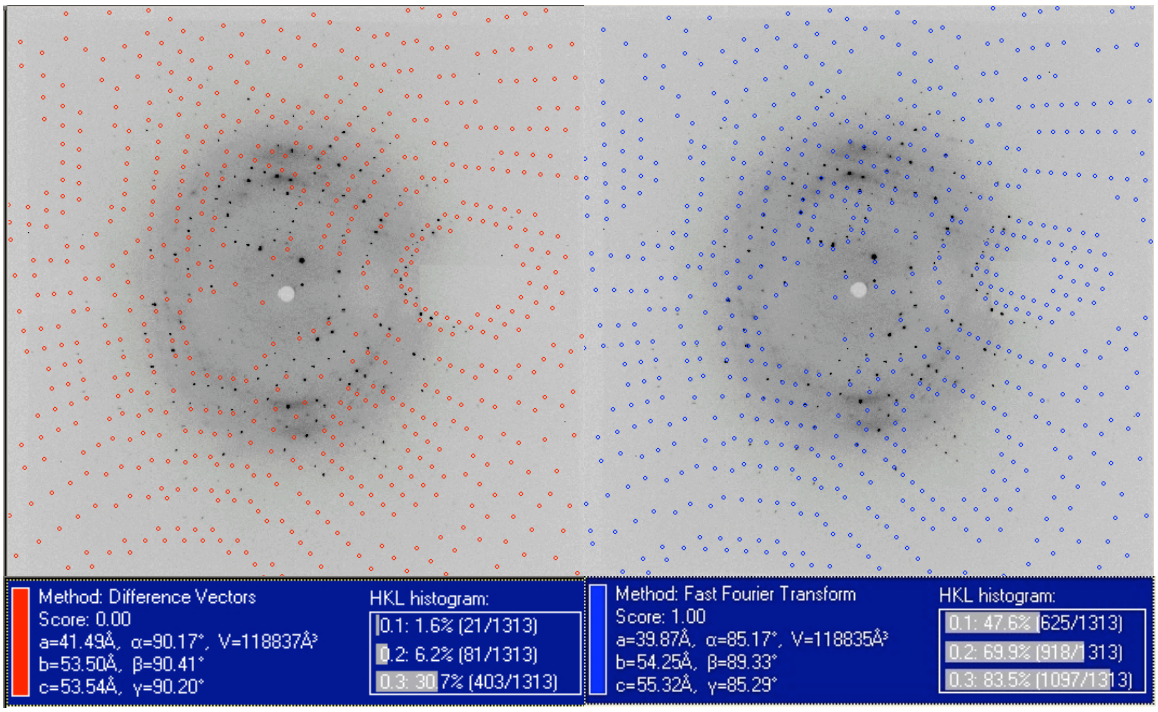

Figure 4.5.5: Multiple method Indexing results: Initial Difference Vector and Fast Fourier Transform indexing results from PROTEUM2: The Fast Fourier Transform spot prediction indexing method yields the best result.

The output from the indexing algorithms is displayed in a graphical format which displays the diffraction image with predicted spot locations overlapping experimental spots.. Additionally the h, k l indices assigned during this stage will have a integer distance attributed to each. The percentage of these reflection indices which were in fact assigned integer values is displayed as HKL histograms in Figure 4.5.6, the higher the percentages the better the result. It is readily apparent that for the AF1382 data the Fast Fourier Transform indexing method was

selected. The program now uses a refinement procedure to increase the accuracy of the indexing protocol.

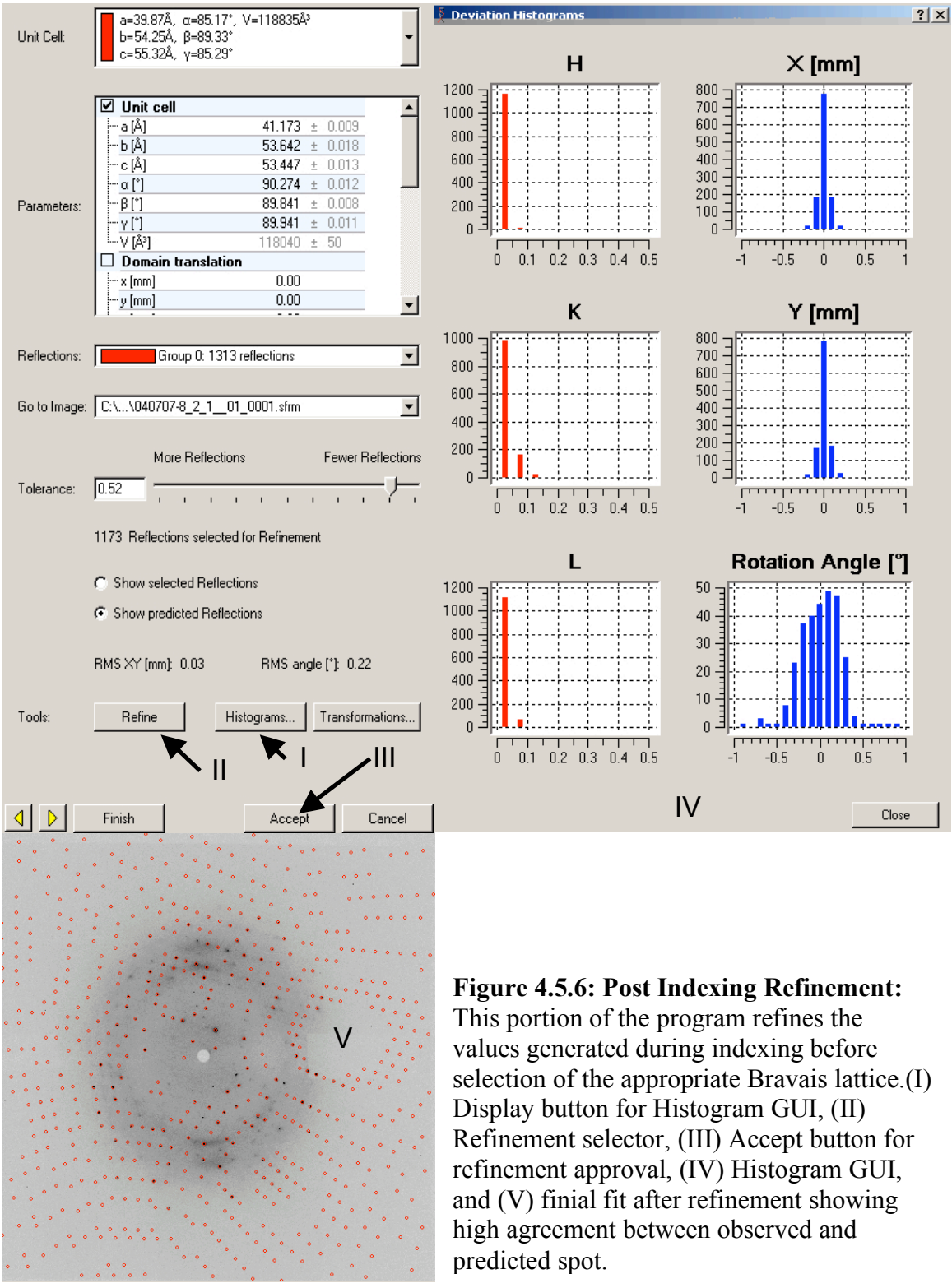

Figure 4.5.6: Post Indexing Refinement: This portion of the program refines the values generated during indexing before selection of the appropriate Bravais lattice.(I) Display button for Histogram GUI, (II) Refinement selector, (III) Accept button for refinement approval, (IV) Histogram GUI, and (V) finial fit after refinement showing high agreement between observed and predicted spot.

After choosing a indexing solution PROTEUM2 conducts a cycle of refinement. To refine the initial spot prediction, the Histogram button is selected which displays a graphical representation of the intergerness of the HKL values and the X, Y, and φ values assigned to each spot (Figure 4.5.6; IV). Each time the Refine button is selected, changes occur in both the RMS values and Histogram shape. Refinement is continued until negligible changes occurred in both indicators. Ideally the histograms for HKL should consist of a single bar nearest to the zero values as possible. This is a reflection of the only integer values for the h, k, l indices assigned to the diffraction spots. The X. Y (mm) and Φ values would also ideally consist of a single bar located at the zero position representing a perfect match between location on the face of the detector and throughout the various frames of data for each diffraction spot. Lower the RMS values indicate a high accuracy involving PROTEUM2's interpretation of the data but the user should still inspect the circular spot selection profiles from the GUI interface to ensure the program has accurately located observed and predicted spots. Of the programs covered in this work only PROTEUM2 offers both easily identifiable and understood text and graphical approximations representing the quality of refinement seen as RMS vales and Histograms.

Next, the user is tasked with selecting the appropriate Bravais Lattice corresponding to the indexing and refinement thus far. The program highlights the most probable lattice. If the indexing and refinement procedures conducted thus far have produced favorable results it is best to accept this selection.

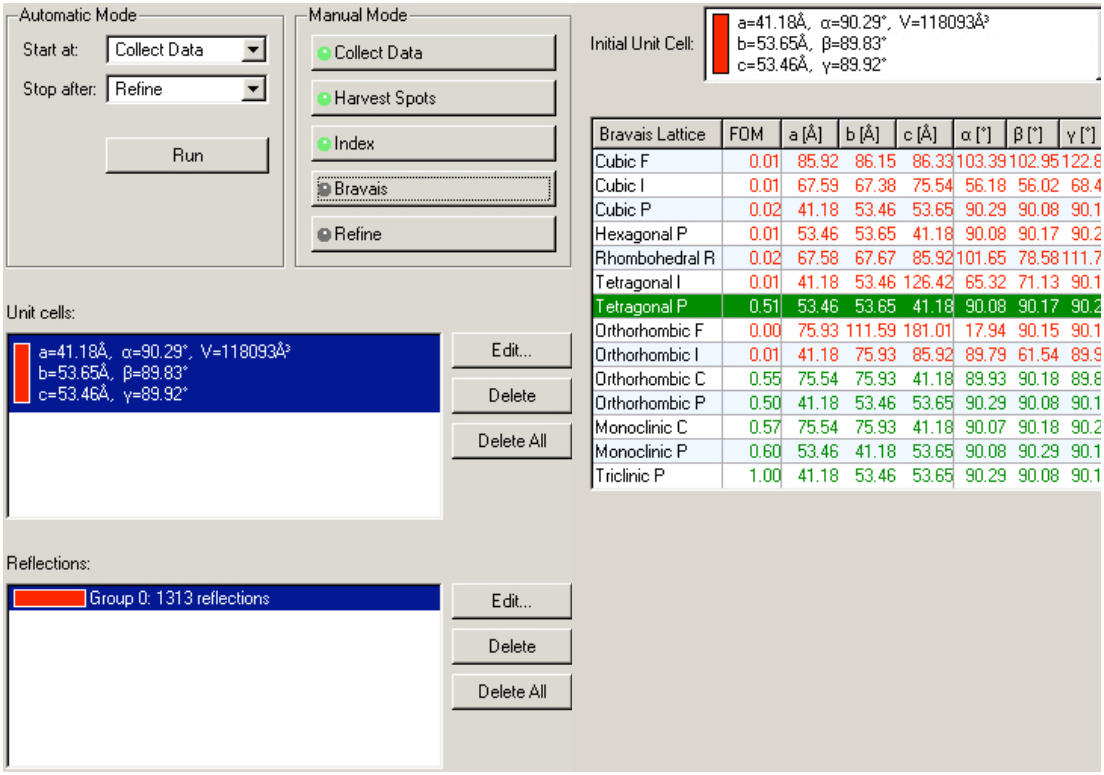

Figure 4.5.7: Bravais Lattice selection: PROTEUM2 highlights the most probable Lattice corresponding to the results from refined Fast Fourier Indexing.

After selecting the proper Bravais Lattice, an additional round of refinement is initiated by PROTEUM2. This is accomplished in the same fashion as refinement shown earlier in Figure 4.5.6.

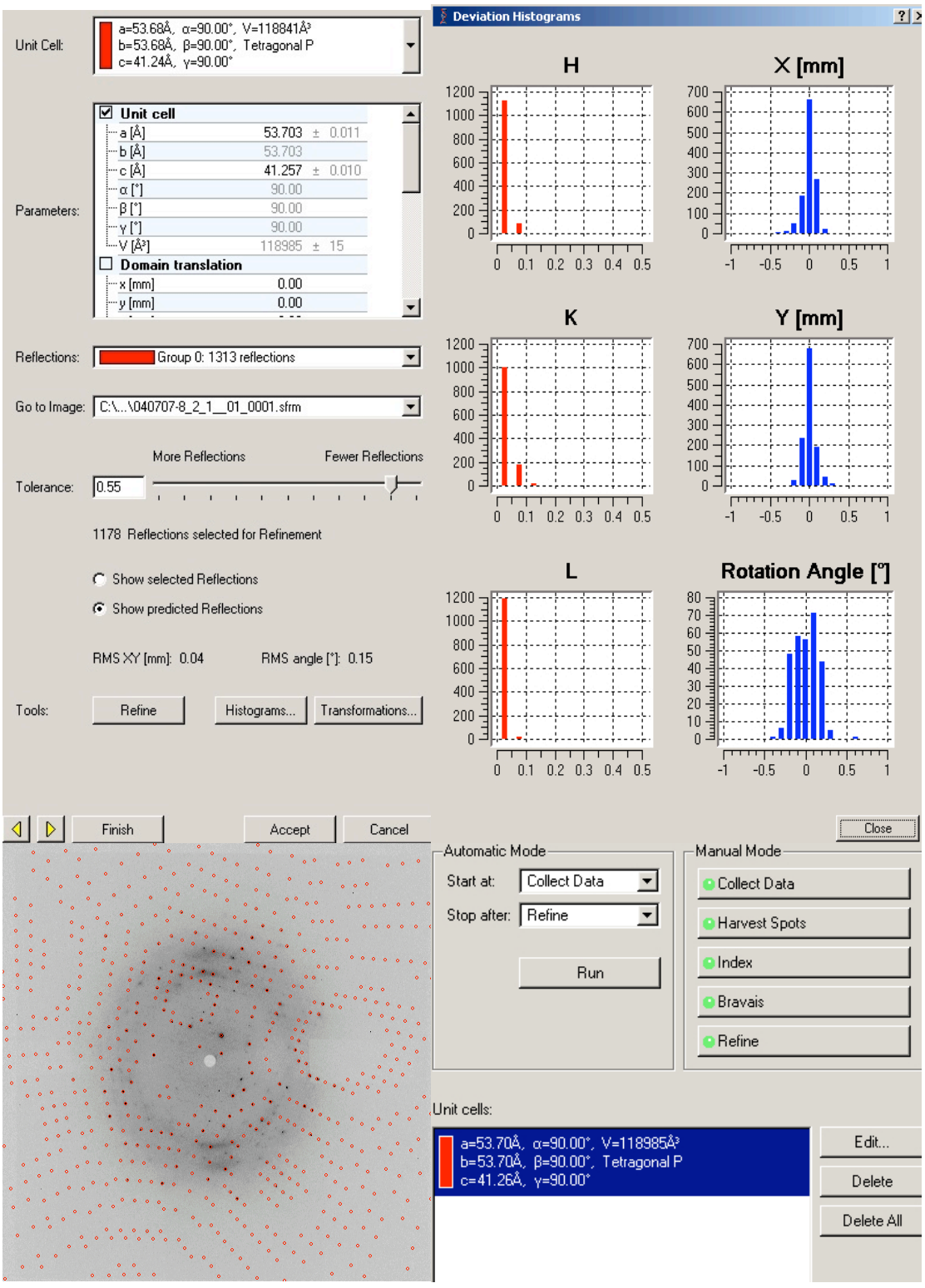

Figure 4.5.8: Final Refinement: The Final Indexing parameters from PROTEUM2.

The finial results from this refinement, which should result in lower RMS values, are saved for further processing via the integration engine within PROTEUM2.

Integration

The PROTEUM2 processing suite utilizes the SAINT integration engine (98), which has been greatly influenced by the integration technique developed by Wolfgang Kabsch (55). To initiate integration from the Integrate tab is chosen from the flowchart on the left of the PROTEUM2 GUI and select the only subroutine, Integrate Images.

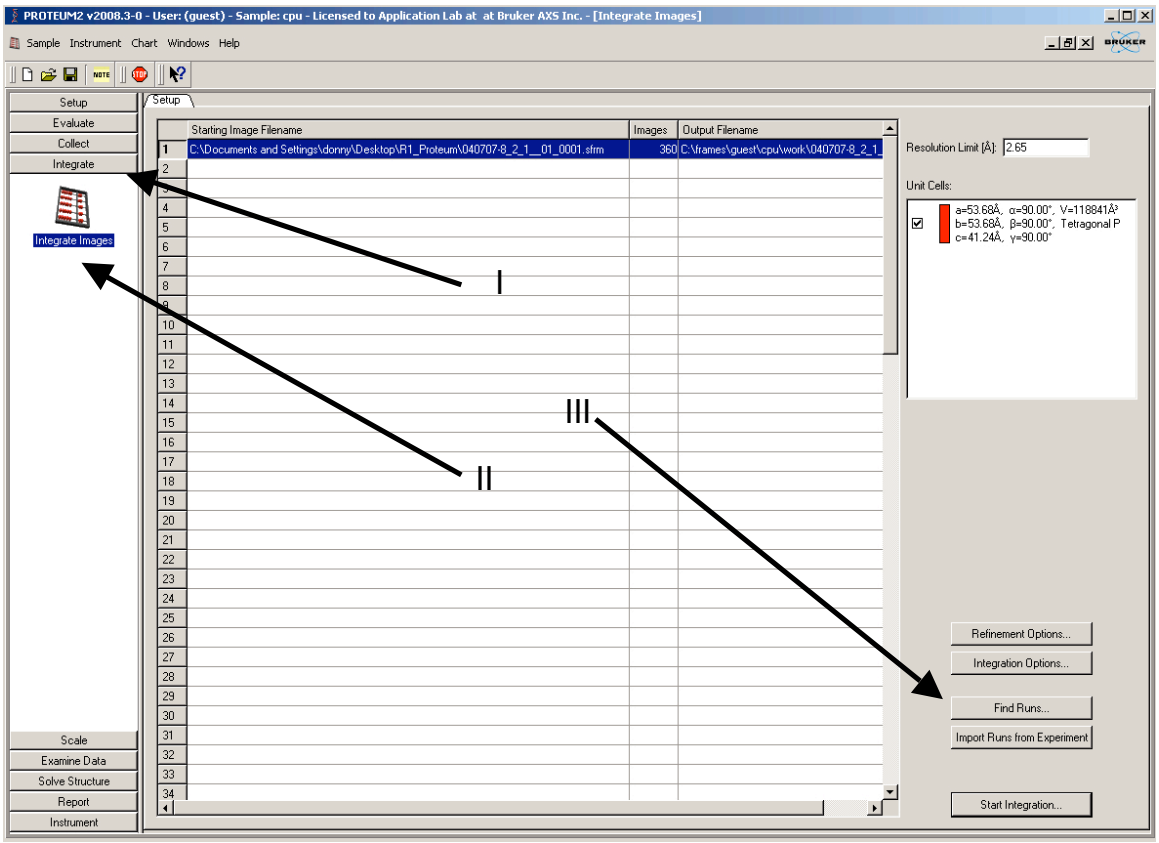

Figure 4.5.9: PROTEUM2 Integration GUI: (I) Integrate tab from PROTEUM2 flowchart, (II) Integrate selection tab, and (III) identifies data set to be integrated.

The user is prompted to load a previously indexed processing attempt, namely, the process just conducted. It is possible, however, to treat the indexing portion of PROTEUM2 separately from integration and continue at a later date if necessary. Both XDS and PROTEUM2 easily allow the

user to halt then resume processing. This may be possible using HKL2000, MOSFLM or d*TREK GUI interface, but may require a more experienced user. The detector is divided into 9 rejoins for the purpose of 3 dimensional profile fitting and the calculation of correlation coefficients used in determining what data should be rejected. Just as in XDS an integration box is defined by a 9x9x9 grid, for background correction, which extends through the angular range of the spot distribution. The user has several options which can be implemented to personalize the PROTEUM2 integration engine. The most pivotal of these are the x, y, and z estimates of the in-plane and angular spot distribution sizes for box determination. These values are generated by the indexing algorithm and relayed to the SAINT integration engine (98). As with all other data reduction programs the importance of box size being too large is far better than too small. For this study we accepted default values generated by PROTEUM2. Among several remaining options such as “box optimization”, “decay correction”, “matrix updating”, and various other constraints I selected only to vary the choice of “narrow” or “wide” frame processing. The use of either wide or narrow integration methods involve the rocking curve and Φ values at which the data was collect. Simply stated if the scan range, Φ, is ¼ to ½ the width of the curve width then “narrow” processing is preferred if the scan width is greater than previously mentioned then the user should select “wide” frame processing (98). Although SAINT is touted to work best in “narrow” frame processing, in this work both methods were tested and I found wide frame processing to be more beneficial.

During the integration process SAINT updates the orientation parameters for both the crystal and detector orientation as each frame is processed. The small changes which are applied to the orientation parameters as frames progress are smoothed via a “dying average” algorithm (99). This parameter updating is governed by the following;

$$P' = P_o + (P - P_o) / (4WX)$$

P' is the end result of orientation updating, P_o is the current parameter value, P is the value of the parameter determined during indexing, and 4WX is the "correlation length" in which W is the estimated spot width and X having a default value of 1.0 but can be user defined (98). SAINT performs a global unit cell least squares refinement at the conclusion of integration. This global refinement being conducted before scaling is the primary reason data from HKL2000 cannot be easily scaled using any program other than Scalepack. During the integration process SAINT also refines the crystal system and experimental parameters.

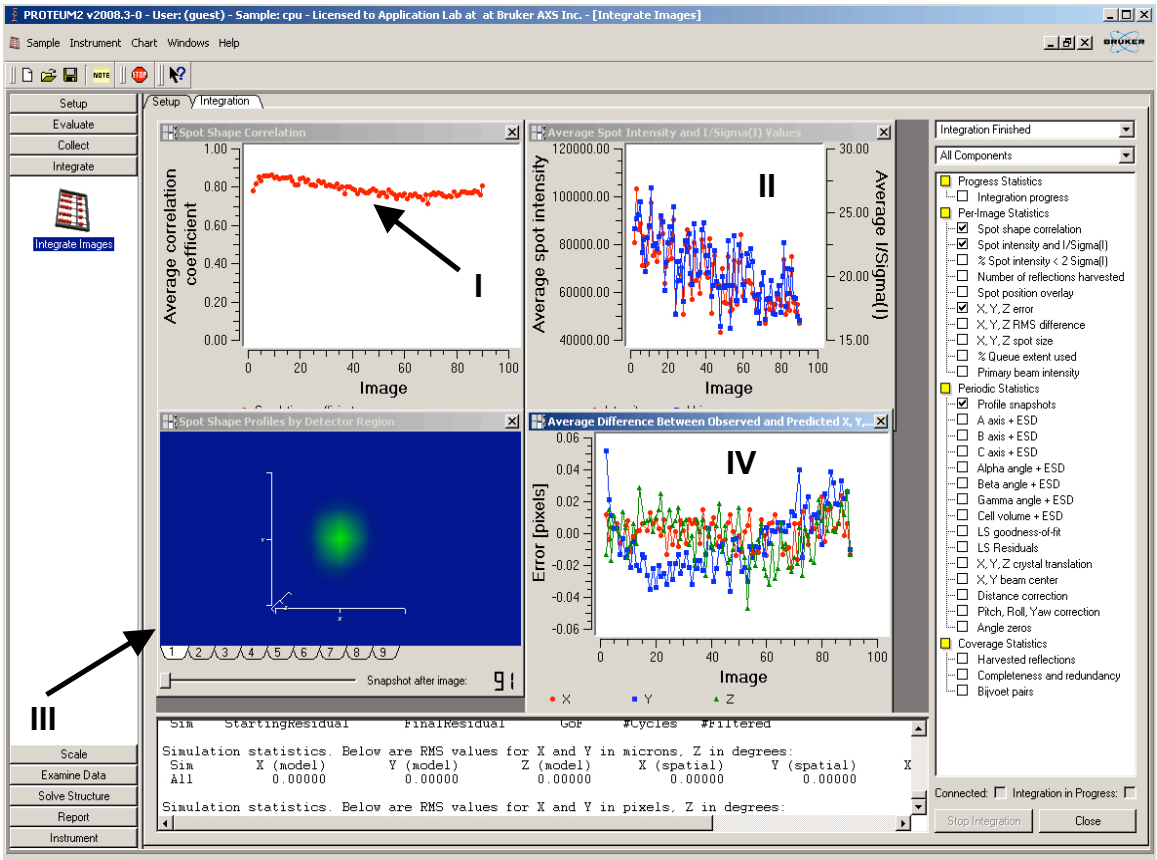

Figure 4.5.10: Integration window: Displaying (I) Spot Shape Correlation – a value of 1.0 being perfect, (II) Average I/σ_I, (III) Spot shape Profile, and (IV) RMS difference in X, Y, and φ.

The Integration window allows the user to view several statistical values associated with the quality of integration. Most importantly the Spot Shape Correlation (I) having a value between 0.8 and 1.0 seems to infer high quality processing. Also the Average I/σ_I graph would be excellent for monitoring the decay rate for a data set.

Scaling

The resultant output of integration consist of several files, and the most important are the .raw and .p4p files and used in scaling. PROTEUM2 scaling engine SADABS was found to produce lower quality results than 3DSCALE and the latter was chosen for scaling.

The results from 3DSCALE are as follows:

R-factors -

$$R_{sym} = \sum_{h} \sum_{i} \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

R_{sym} is the R-factor chosen 3DSCALE to relate differences in symmetry related reflections. This is a measure of the accuracy of the data. The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection.

R1- Scaling Statistics using 3DSCALE

R1-PROTEUM2

Res.Shell	nRefObs	nRefExp	nRefCen	Compl%	Redund
40.98					
to 5.80	351	353	4787	99.43	13.67
to 4.58	702	710	9882	98.87	14.09
to 3.99	1050	1058	14905	99.24	14.22
to 3.61	1401	1420	19993	98.66	14.28
to 3.35	1754	1773	25038	98.93	14.29
to 3.14	2106	2131	30050	98.83	14.29
to 2.99	2451	2476	34961	98.99	14.30
to 2.86	2806	2838	39899	98.87	14.29
to 2.74	3158	3190	44767	99.00	14.27
to 2.65	3510	3542	49596	**99.10**	**14.22**

Res.Shell	Rsym---Shell		Rfree	nRfree	<i/sigi>--Shell		<I/sigI>--Shell		<Chi^2>
40.98									
to 5.80	0.0284	0.0284	0.0451	16	21.72	21.72	79.09	79.09	1.15
to 4.58	0.0297	0.0312	0.0386	39	21.32	20.93	79.18	79.27	1.16
to 3.99	0.0315	0.0357	0.0395	53	20.53	18.98	76.76	71.86	1.19
to 3.61	0.0344	0.0467	0.0415	73	19.73	17.36	74.08	66.03	1.27
to 3.35	0.0368	0.0562	0.0433	94	18.52	13.77	69.73	52.39	1.29
to 3.14	0.0387	0.0656	0.0437	105	17.21	10.66	64.84	40.34	1.26
to 2.99	0.0403	0.0723	0.0455	124	15.94	8.24	60.15	31.39	1.20
to 2.86	0.0420	0.0949	0.0473	143	14.72	6.19	55.62	23.80	1.15
to 2.74	0.0438	0.1233	0.0487	156	13.61	4.69	51.49	17.96	1.09
to 2.65	**0.0457**	**0.1523**	0.0501	173	12.63	3.56	**47.70**	**13.27**	1.04

Res.Shell	<AnoI/SigI>a		<AnoI/SigI>c		---Ras----	
40.98	--shell		--shell		--shell	
to 5.80	3.33	3.33	1.50	1.50	2.22	2.22
to 4.58	2.89	2.49	1.55	1.60	1.86	1.56
to 3.99	2.57	1.98	1.62	1.75	1.59	1.13
to 3.61	2.47	2.17	1.70	1.93	1.45	1.12
to 3.35	2.37	2.01	1.73	1.86	1.37	1.08
to 3.14	2.31	1.99	1.72	1.64	1.34	1.22
to 2.99	2.21	1.64	1.68	1.44	1.32	1.14
to 2.86	2.13	1.59	1.64	1.38	1.30	1.15
to 2.74	2.07	1.60	1.60	1.30	1.29	1.23
to 2.65	2.00	1.41	1.57	1.24	1.28	1.14

70288 reflections read in

Excluded for Scaling		Excluded for Output		
0		0		on the marked-off frames
2		2		low resolution cutoffs 53.763
20154		20154		high resolution cutoffs 2.650
0		0		low I/SigI cutoffs -2.838
0		0		high I/SigI cutoffs 182.823
0		47		rejected as outliers
0	0	0	0	single reflections
2525	173	0	0	randomly selected for Rfree subset
Observed	Unique	Observed	Unique	total reflections used scaling/output
47448	3337	49926	3510	(excluding lattice-center-related extinctions)

Table 4.5.1: R1 Scaling results from PROTEUM2 processing using 3DSCALE: R1 processing yields acceptable R_{sym} results for continued processing.

R1 Data Automated tracing results using SGXPro

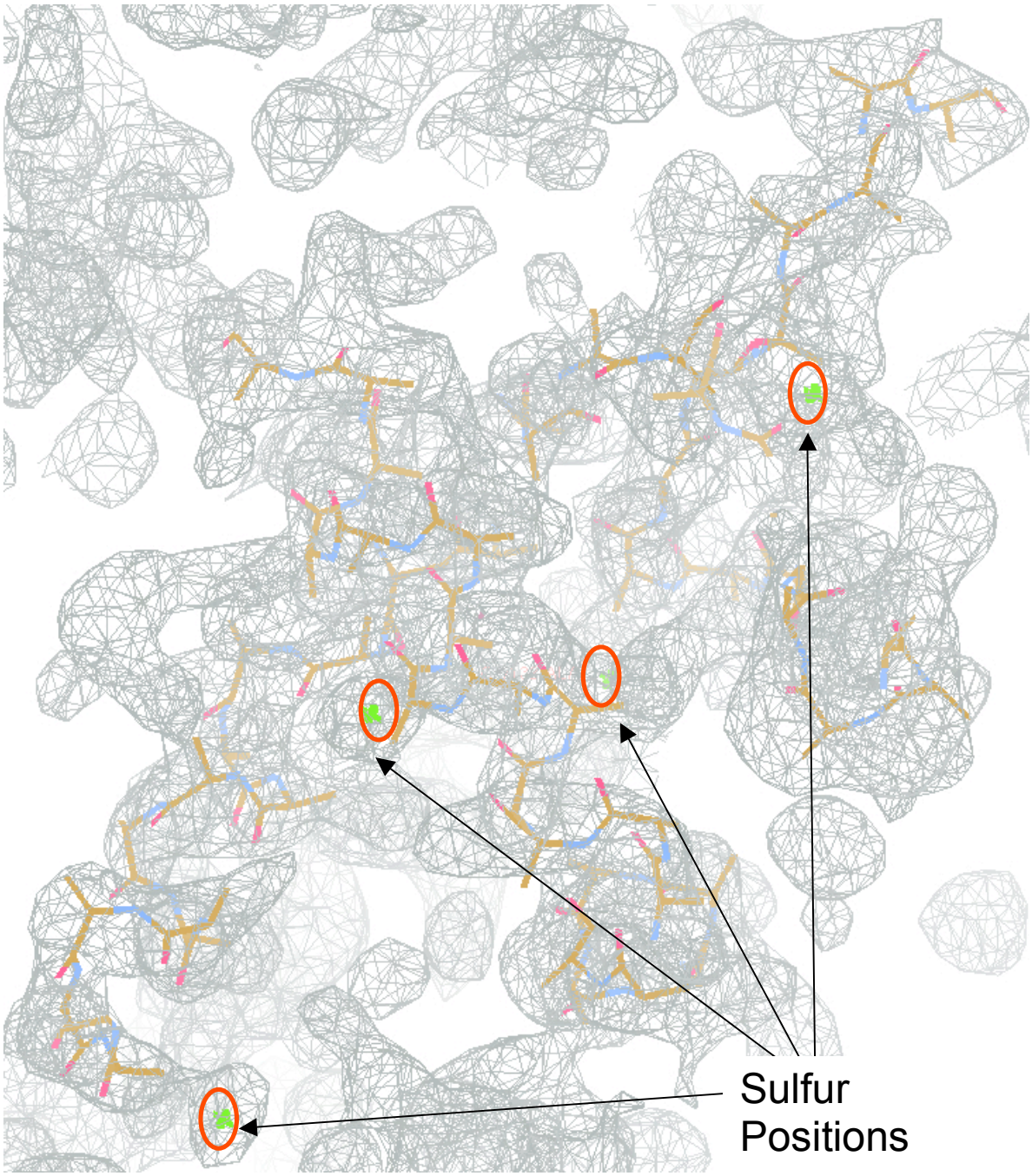

Figure 4.5.11: Tracing result from PROTEUM2 processing of R1 data set: All four Sulfur positions were correctly identified, with 61% of the total Amino Acids traced. The trace from RESOLVE agreed well with the refined model.

R1 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
 --- -------- -------  --------- -----------
 1     58       5       20 15 14       /Desktop/R1-PROTEUM2/t1/zzsgxSol_1.pdb
Heavy atoms:                        /Desktop/ R1-PROTEUM2/t1/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 4 helices with 11,10,4,3 amino acids long
#Beta-sheets: 2 beta sheet(s) 8 and 3 amino acids long
CC_ALL/CC_WEAK    36.3/15.1
PATFOM      134.71
```

Table 4.5.2: R1 Tracing results from SGXPro Novel Structure Solution: R1 processing yielded excellent CC-ALL/Weak and PATFOM values. The phases and tracing statistics were excellent as well.

R2- Scaling Statistics using 3DSCALE

R2-PROTEUM2

Res.Shell		nRefObs	nRefExp	nRefCen	Compl%	Redund
40.98						
to	5.77	348	356	4563	97.75	13.14
to	4.58	697	710	9625	98.17	13.82
to	3.99	1045	1058	14584	98.77	14.00
to	3.61	1396	1411	19645	98.94	14.10
to	3.35	1745	1763	24603	98.98	14.13
to	3.15	2090	2117	29485	98.72	14.14
to	2.99	2440	2467	34367	98.91	14.14
to	2.86	2790	2822	39236	98.87	14.14
to	2.75	3139	3171	44049	98.99	14.13
to	2.65	3496	3528	48983	**99.09**	**14.10**

Res.Shell		Rsym---Shell		Rfree	nRfree	<i/sigi>--Shell		<I/sigI>--Shell		<Chi^2>
40.98										
to	5.77	0.0379	0.0379	0.0446	17	22.13	22.13	78.60	78.60	1.94
to	4.58	0.0365	0.0352	0.0413	34	22.24	22.33	81.50	84.40	1.97
to	3.99	0.0378	0.0405	0.0437	51	21.90	21.24	81.22	80.66	2.09
to	3.61	0.0398	0.0474	0.0455	71	21.57	20.61	80.50	78.33	2.24
to	3.35	0.0416	0.0553	0.0475	91	20.97	18.61	78.47	70.35	2.34
to	3.15	0.0430	0.0617	0.0484	107	20.20	16.34	75.66	61.40	2.38
to	2.99	0.0441	0.0667	0.0494	124	19.32	14.06	72.50	53.44	2.36
to	2.86	0.0452	0.0782	0.0502	137	18.38	11.82	69.10	45.13	2.31
to	2.75	0.0463	0.0906	0.0514	153	17.49	10.30	65.76	38.84	2.26
to	2.65	**0.0474**	**0.1089**	0.0525	171	16.57	8.37	**62.28**	**31.35**	2.22

<<< RAS >>>

Res.Shell		<AnoI/SigI>a		<AnoI/SigI>c		---Ras----	
40.98		--shell		--shell		--shell	
to	5.77	3.86	3.86	1.94	1.94	1.99	1.99
to	4.58	3.52	3.21	1.97	2.00	1.79	1.60
to	3.99	3.22	2.65	2.09	2.33	1.54	1.14
to	3.61	3.08	2.68	2.21	2.56	1.39	1.05
to	3.35	2.98	2.63	2.28	2.57	1.31	1.03
to	3.15	2.92	2.58	2.32	2.55	1.26	1.01
to	2.99	2.81	2.19	2.33	2.40	1.20	0.91
to	2.86	2.75	2.34	2.32	2.24	1.18	1.04
to	2.75	2.70	2.31	2.31	2.19	1.17	1.06
to	2.65	2.65	2.20	2.29	2.13	1.15	1.03

< Number of Reflections Used for Scaling and Output >:

91223 reflections read in

Excluded for Scaling		Excluded for Output		
0		0		on the marked-off frames
2		2		low resolution cutoffs 53.763
41703		41703		high resolution cutoffs 2.650
0		0		low I/SigI cutoffs -2.997
0		0		high I/SigI cutoffs 277.377
0		85		rejected as outliers
2	2	0	0	single reflections
2461	171	0	0	randomly selected for Rfree subset
Observed	Unique	Observed	Unique	total reflections used scaling/output
46931	3323	49309	3496	(excluding lattice-center-related extinctions)

Table 4.5.3: R2 Scaling results from PROTEUM2 processing using 3DSCALE: R2 processing yields acceptable R_{sym} results for continued processing.

R2 Data Automated tracing results using SGXPro

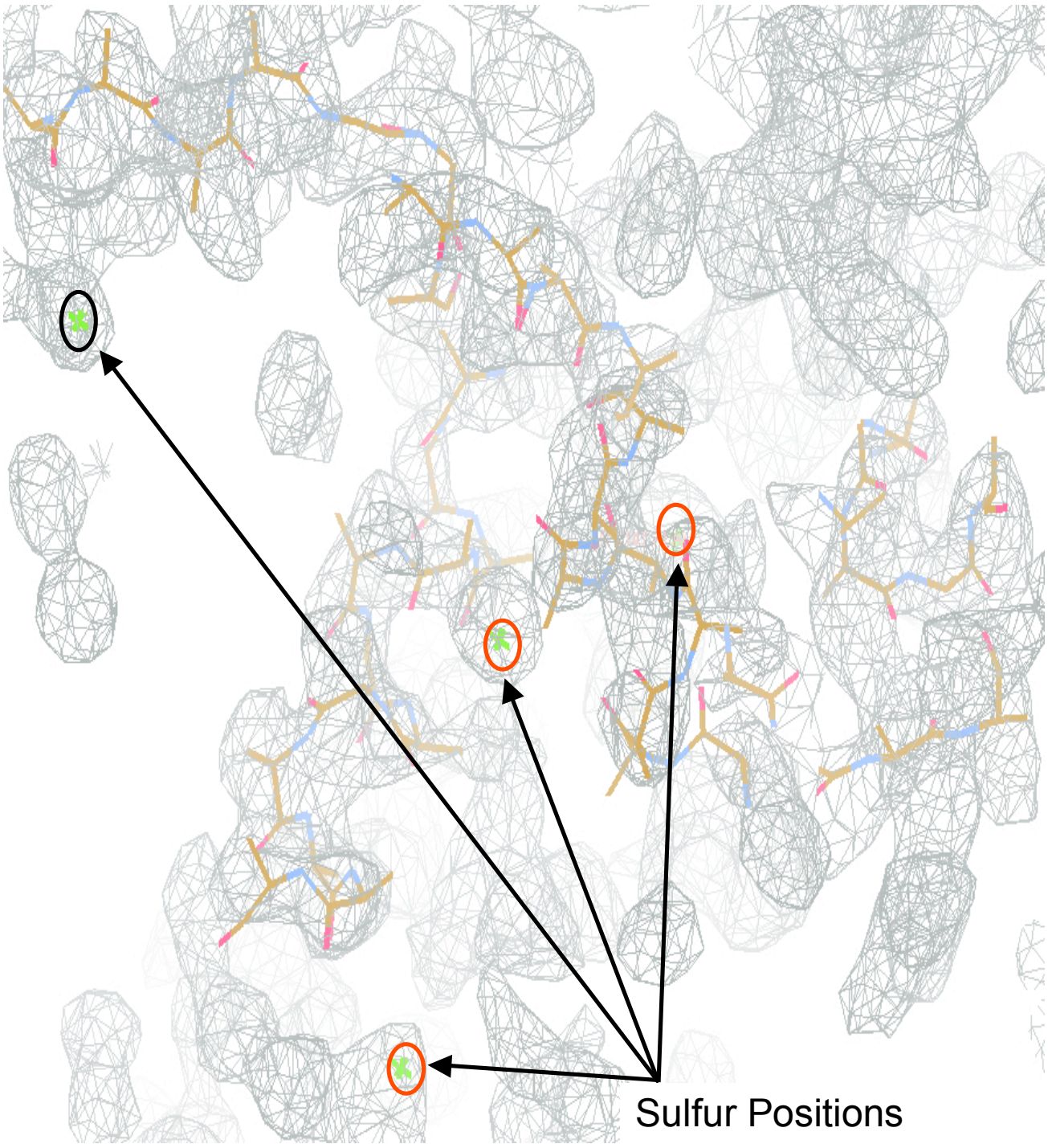

Figure 4.5.12: Tracing result from PROTEUM2 processing of R2 data set: Three of the four Sulfur positions were correctly identified in Red while an errant Sulfur position identified by SGXPro is circled in black. A total of 48% of amino acids were traced. The trace from RESOLVE agreed well with the refined model.

R2 - Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
---  -------- -------  --------- -----------
 1      46       5      19 8 7  /Desktop/R1-PROTEUM2/t3/zzsgxSol_1.pdb
Heavy atoms:                   /Desktop/R1-PROTEUM2/t3/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 4 helices with 11,8,5 amino acids long
#Beta-sheets: No discernible beta sheet(s)
CC_ALL/CC_WEAK    40.14/22.0
PATFOM      66.29
```

Table 4.5.4: R2 Tracing results from SGXPro Novel Structure Solution: R1 processing yielded excellent CC-ALL/Weak but very poor PATFOM values. The phases and tracing statistics were not as high quality as the R1 data set but better than any other data reduction program used in this work.

R1-R2 Scaling Statistics using 3DSCALE

R1-R2 PROTEUM2

Res.Shell	nRefObs	nRefExp	nRefCen	Compl%	Redund
40.98					
to 5.80	351	353	8831	99.43	25.21
to 4.58	702	710	18614	98.87	26.54
to 3.99	1053	1061	28299	99.25	26.97
to 3.61	1398	1417	37969	98.66	27.25
to 3.35	1754	1773	47811	98.93	27.35
to 3.15	2104	2127	57474	98.92	27.40
to 2.99	2457	2480	67213	99.07	27.48
to 2.86	2806	2836	76680	98.94	27.49
to 2.74	3158	3188	86248	99.06	27.50
to 2.65	3510	3540	95573	**99.15**	**27.40**

Res.Shell	Rsym---Shell		Rfree	nRfree	<i/sigi>--Shell		<I/sigI>--Shell		<Chi^2>
40.98									
to 5.80	0.0369	0.0369	0.0435	16	17.35	17.35	86.29	86.29	1.10
to 4.58	0.0366	0.0362	0.0381	39	17.20	17.05	88.22	90.16	1.02
to 3.99	0.0379	0.0405	0.0398	54	16.62	15.51	86.46	82.90	1.00
to 3.61	0.0405	0.0509	0.0429	73	15.97	14.09	83.91	76.08	1.02
to 3.35	0.0429	0.0599	0.0452	94	15.03	11.39	79.33	61.35	1.00
to 3.15	0.0447	0.0685	0.0459	105	13.99	8.84	74.12	47.90	0.95
to 2.99	0.0463	0.0751	0.0481	125	12.96	6.93	69.03	38.46	0.88
to 2.86	0.0479	0.0923	0.0497	143	12.01	5.31	64.15	29.55	0.82
to 2.74	0.0496	0.1119	0.0509	155	11.13	4.22	59.67	23.55	0.77
to 2.65	**0.0513**	**0.1371**	0.0526	173	10.37	3.25	**55.52**	**17.77**	0.72

<<< RAS >>>

Res.Shell	<AnoI/SigI>a		<AnoI/SigI>c		---Ras----	
40.98	--shell		--shell		--shell	
to 5.80	3.62	3.62	1.47	1.47	2.46	2.46
to 4.58	3.22	2.85	1.47	1.46	2.19	1.94
to 3.99	2.87	2.22	1.50	1.55	1.92	1.43
to 3.61	2.76	2.43	1.53	1.64	1.80	1.48
to 3.35	2.63	2.16	1.53	1.54	1.72	1.40
to 3.15	2.52	1.97	1.50	1.34	1.68	1.47
to 2.99	2.37	1.53	1.46	1.20	1.63	1.27
to 2.86	2.26	1.50	1.41	1.08	1.60	1.38
to 2.74	2.17	1.42	1.37	1.03	1.58	1.37
to 2.65	2.07	1.27	1.33	0.97	1.56	1.31

< Number of Reflections Used for Scaling and Output >:

157156 reflections read in

Excluded for Scaling		Excluded for Output			
0		0		on the marked-off frames	
0		0		in the marked-off batches	
3		3		low resolution cutoffs	53.763
60953		60953		high resolution cutoffs	2.650
0		0		low I/SigI cutoffs	-2.997
0		0		high I/SigI cutoffs	274.288
0		28		rejected as outliers	
0	0	0	0	single reflections	
4804	173	0	0	randomly selected for Rfree subset	
Observed	Unique	Observed	Unique	total reflections used scaling/output	
91396	3337	96172	3510	(excluding lattice-center-related extinctions)	

Table 4.5.5: R1-R2 Scaling results from PROTEUM2 processing using 3DSCALE: R1-R2 processing yields acceptable R_{sym} results for continued processing.

R1-R2 Data Automated tracing results using SGXPro

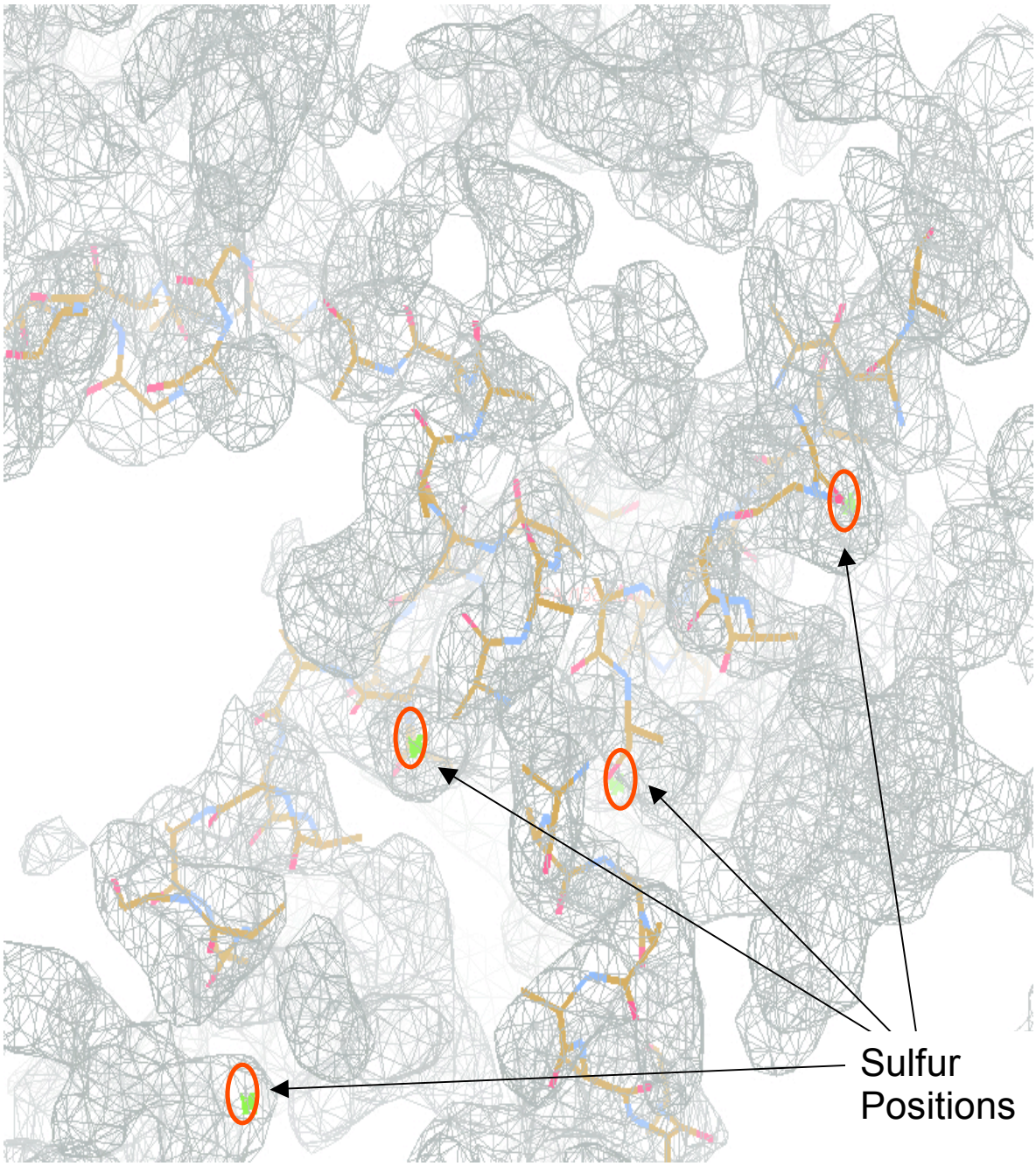

Figure 4.5.13: Tracing result from PROTEUM2 processing of R1-R2 Merged data set: All four Sulfur positions were correctly identified, with 58% of the total Amino Acids traced. The trace from RESOLVE agreed well with the refined model.

R1-R2 merged Heavy Atom/Tracing statistics using SGXPro

```
No#  NumBuilt NumSegs  Top3Segs  Model Built
---  -------- -------  --------- -----------
 1      55       8      12 8 8   /Desktop/PROTEUM2/t2/zzsgxSol_1.pdb
Heavy atoms:                     /Desktop/PROTEUM2/t2/zzsgxSol_1_ha.xyz
Sulfur atoms found:     4
#Alpha-helix: 4 helices with 11,8,6,5,5 amino acids long
#Beta-sheets: 1 beta sheet(s) 7 amino acids long
CC_ALL/CC_WEAK    38.2/19.4
PATFOM      125.9
```

Table 4.5.6: R1-R2 Tracing results from SGXPro Novel Structure Solution: R1-R2 merged data processing yielded excellent CC-ALL/Weak and PATFOM values. The phases and tracing statistics were of excellent quality.

Chapter 5

Phase Comparison

After collecting initial statistical results, attempts were made to quantitatively determine the prerequisites necessary to declare a solution from any one program as superior to another. Statistical analysis of the data produced by each program did not offer a clear quantitative and user independent distinction in determining which data processing program produced the best results. This fact provided a layer of difficulty in qualifying which program would be deemed most effective. The structural representation of the protein generated by SGXPro provided number of segments and length of segments in each solution. Using the traced map a visual comparison of the results of each program versus the 2QVO entry in the PDB was possible; however, a visual comparison of the traced results is not a quantitative measure of solution quality. Although a visual approximation of map quality separates the solutions generated by PROTEUM2 from nearly all others in terms of quality, a effectively means of differentiating between maps deemed with similar parameters required more than a subjective application of visual interpretation.

To quantitatively determine which of the data reduction program yielded the best results or highest quality phases a set of programs were used, entitled PHASEMATCH(100)and PHISTATS(39). These programs were implemented in hopes of providing the quantitative comparison between our known 2QVO phases and those of out target or processing solution phases. For brevity, the solution generated using different data reduction programs will be referred to as target.pdb and 2QVO.pdb will serve as the accepted coordinate model. These phase comparison programs utilize the phase angle and Figure of Merit calculated from each result to be compared. Performing this comparison analysis involves several programs and

procedures pertaining to preparation of the data. It was essential to perform tasks involving map superposition, manual coordinate file editing, reindexing MTZ files, refinement and merging data from two MTZ file before the phase comparison utility could be used.

The first trial faced was to accurately superimpose the experimental target.pdb coordinates from the data reduction programs with the accepted 2QVO.pdb coordinates. I began by using Coot version 0.5-pre-1(75), a model building tool for molecular and Chimera version 1.3(101), a tool for visualization and analysis of molecular structures as my choice for superposition trials.

Superposition

Most of the data reduction programs did not yield a map with secondary structure comparable with those 2QVO.pdb coordinates. Of the solutions which did yield a map including secondary structure, superimposing the 2QVO coordinates and solutions proved difficult. The major issues involved with superimposing any of the target models with the 2QVO coordinates were the numbering systems used for identifying individual amino acids within each coordinate file. The numbering inconsistencies, in concert with inherent gaps within the target trace, yielded a challenge to properly re-number and bridge any amino acids sequences not traced within the target.pbd coordinates. This posed quite a problem when attempting sequence based alignment. Re-formatting the numerical assignments associated with the sequences from each solution target.pdb file generated by SGXPro for each data reduction program (approximately 320 solutions from each data reduction program) was both daunting and unrealistic. This type of re-formatting .pdb files is commonly done in cases where common secondary structure elements exist between two coordinate files, however this is nearly impossible in those cases which produced highly fragmented peptide solutions. Without the location of reference within the

target.pdb coordinates, such as a portion of a well defined helix or sheet, to compare with the 2QVO.pdb coordinates, accurately defining the relative orientation of an experimental solution in reference to the 2QVO.pdb coordinates was unreasonable. Experimental maps which did contain such secondary structure re-formatting attempts were employed to create uniformity between the two coordinate files. Coot and Chimera were tasked to align the target and 2QVO.pdb coordinates using the best aligning pair of chains or using a reference chain matching routines for superposition. Each program continually yielded results, which did not produce accurately superimpose traces and these could often be manually improved upon. This was is primarily attributed to the difficulty in finding a uniform method of structure superposition which did not require user influence or judgment pertaining to the numbering of amino acids or manual amputation of coordinate files. I could not be certain the numbering system chosen during re-formatting was without bias.

A different method had to be used to achieve coordinate file superposition as the primary reason for performing the phase comparison was to avoid any partiality in judging the maps.

Heavy Atom Search

The solution to the problem of superposition was found in the heavy atoms used to phase the experimental maps. Accepting the 2QVO heavy atom coordinates as factual offered a comparative means to judge the quality of data processing for each experimental map. If the heavy atoms locations coincide between the 2QVO.pdb coordinates and a target model, then all other amino acids within both maps should match. This can be expected if the solutions of the heavy atom positions are viable. The program I found most suited to provide a reliable means of superimposeing the heavy atom locations between two models was Chimera. As mentioned during an earlier review of SGXPro, the scaling portion of data processing utilizes software titled

SHELXD to locate heavy atoms or in this case the Sulfur atoms used as anomalous scatterers. The SHELXD output log contains information concerning the inter-atomic distances between the experimentally located heavy atoms, correlation quality of identified heavy atom peak positions and information concerning the Patterson minimum function(102). The information contained in the log files offers a means to differentiating the viable heavy atom locations from poorly generated ones based on Peak values (Figure 1.0; I). Large variations in peak values are indicative of errors pertaining to heavy atom position. By default SHELXD searches for two additional heavy atom positions than the users input incase there are other relevant heavy atom sites. These are separated from the user-defined search but still recorded in the log file. In addition the inter-atomic distances between known heavy atom positions in the 2QVO model could be recorded and compared with the results of the data reduction programs used in this study.

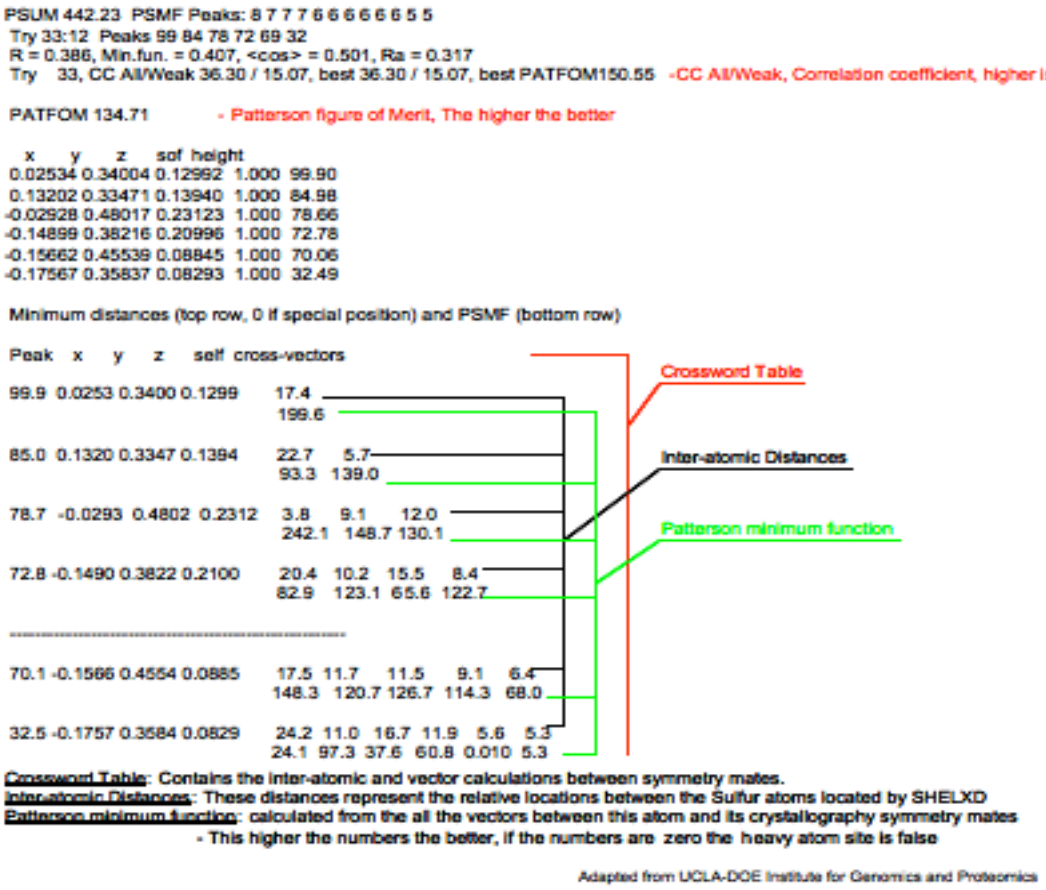

Figure 5.1 SHELXD Analysis of the HA positions: Found within the sheldx.log file output from SGXPro this file contains information pertaining to heavy atom locations within all generated solutions.

The crossword table was based on the use of the Patterson superposition function (103). This offers an easily understood means to recognize which heavy atom sites are correct. The self cross-vector row represents potential heavy atom positions illustrating the inter-atomic distances between heavy atom pairs and the PMF (Patterson Minimum Function) calculated using all vectors between possible pairs. A percentage of confidence is displayed as Peak (Figure 1.0; I) values to indicate the reliability of predicted heavy atom location. By comparing the inter-atomic distances present in the SHELXD output from HKL2000, d*TREK, XDS, MOSFLM and PROTEUM2 with the known values from the 2QVO.pdb coordinates, viable solutions were identified and poor solutions eliminated quickly.

Nomenclature

In this particular case the amino acids in the 2QVO.pdb file were not transformed into alanine. Therefore, it is important to ensure that the sulfur atoms present in the files are labeled the same in both the target.pdb as well as the 2QVO.pdb. By default, Sulfur atoms within a methionine are labeled SD and a cystine SG. Sulfur atoms contained in a target.pdb coordinates file using strictly alanine tracing will be located at the end of the file and labeled as S.

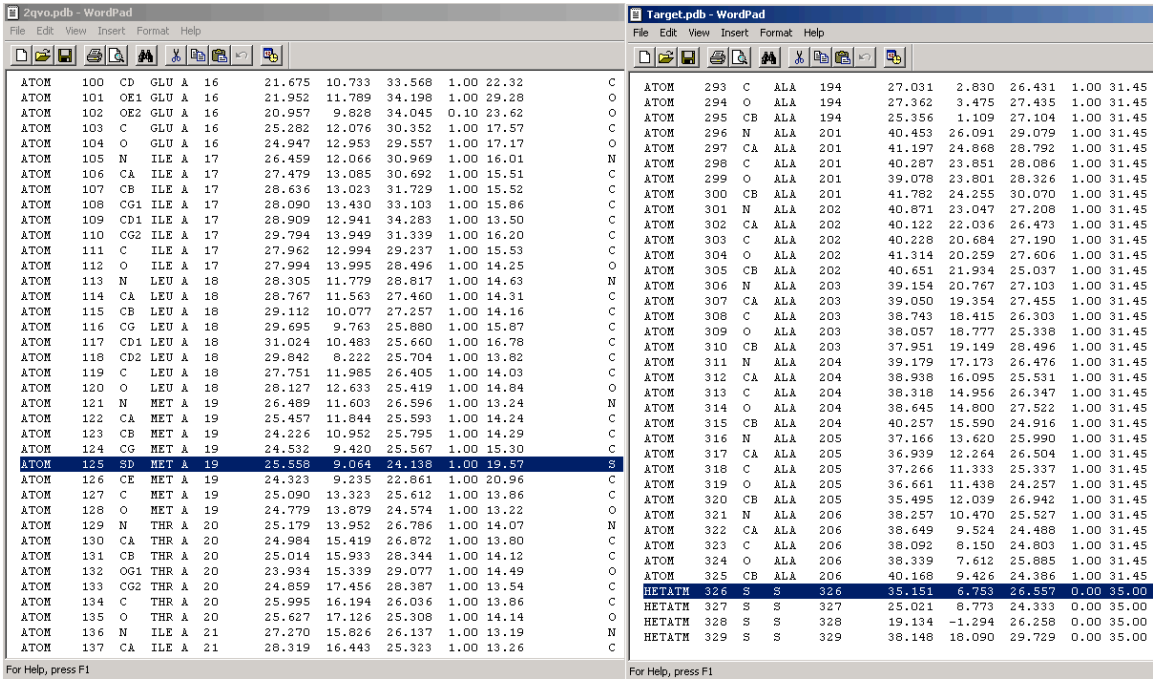

Figure 5.2 Sulfur Atom Nomenclature: The Sulfur atoms are identified differently depending on the method in which the file was generated.

Editing the 2VO.pdb file such that all SG or SD designations are converted to S will solve the nomenclature issues between the two files.

Chimera

After initializing Chimera, a GUI based molecular visualization program, the accepted 2QVO.pdb coordinate is opened followed by the coordinate files from SGXPro. This will designate the 2QVO.pdb file as model #0 and the target.pdb file as model #1. If the heavy atoms located in the files are not readily discernable within the Chimera GUI choose both chains and Select→ Chemistry→ element→ S (or the heavy atom you wish to view choice), and finally choose Actions→ sphere. This will display the heavy atoms as spherical units easily seen within the structures. Next, from the Favorites menu bar select the Command Line as seen in Figure 5.3 labeled 1 and 2. At the bottom of the Chimera GUI the Command Line interface opens as a text input line and an optional tab for choosing active models, labeled 3 and 4.

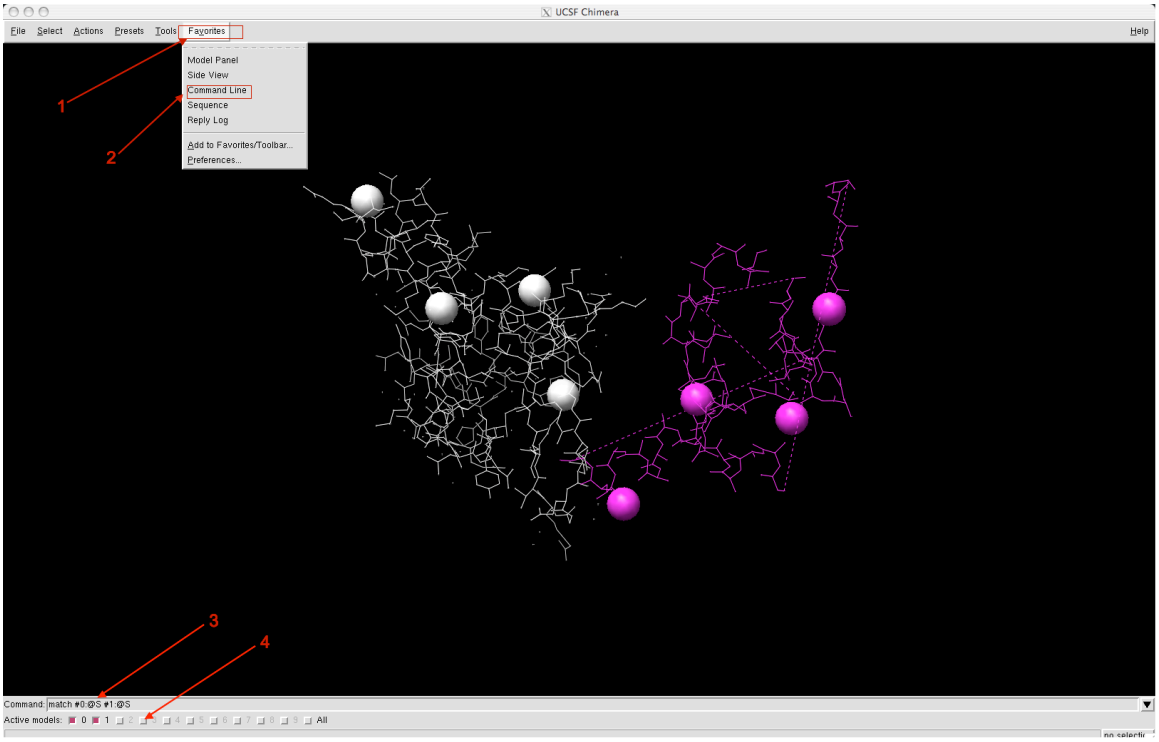

Figure 5.3 Activating the Command Line interface: Choosing Favorites (1)→Command Line (2) will result in a text-based command line and active model selection tab, highlighted as (3) and (4), respectively.

Both models must be active to superimpose the coordinate files. To perform a superposition of the two models using only the Heavy atom sulfur positions, the following command is used:

>match #0:@S #1:@S

It is imperative that the nomenclature for both files simply use S to represent Sulfur or the Chimera will fail to match the structures. With this issue corrected, both the 2QVO and target pdb files can be opened using Chimera.

The match command utilizes a least squares fitting method for the superposition of two models/atoms/or specific amino acids. The #0 and #1 entries identify the models while @S determines the objects that are to be compared for the superposition. This command will match the Sulfur atoms in model #0 to those in model #1, or more directly this will transform the 2qvo.pdb coordinates to those of the target.pdb file.

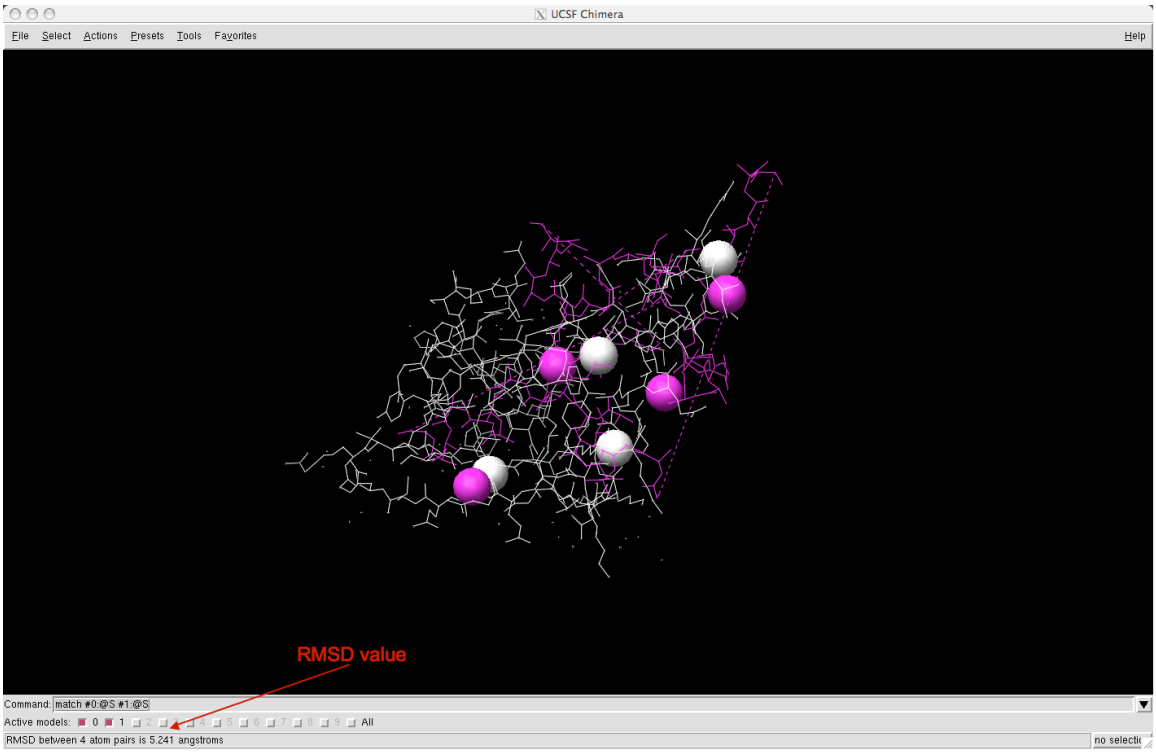

Figure 5.4 Initial superposition of Sulfur atoms: The initial superposition of the models based on Sulfur positions within the pdb files, 2QVO and target.pdb. Note the RMSD value is quite high at 5.24.

Manual Editing Heavy Atom Identifiers

In all likelihood the initial match between the 2QVO and target pdb files will fall into a high RMSD range. RMSD is a commonly used tool for measuring the differences between two measurements (in this case sulfur positions), usually a model and an observation. The lower the RMSD value the higher the correlation between the two items being compared. The reason for this is due to the method in which the command line: match #0:@S #1:@S pairs atoms according to their numerical order within the target and 2QVO.pdb files. A bit of manual model fitting is needed in order to determine which sulfur atoms in the target.pdb file match those in the 2QVO.pdb file. By un-checking one of the highlighted active model boxes the user can freely rotate an individual model. In this instance the hope is to find a better fit than the initial Chimera

match command. In this case it was quickly determined that the maps were 180 degrees off-set from each other. A simple manual rotation yielded an alignment that clearly exhibited a better superposition of the two models.

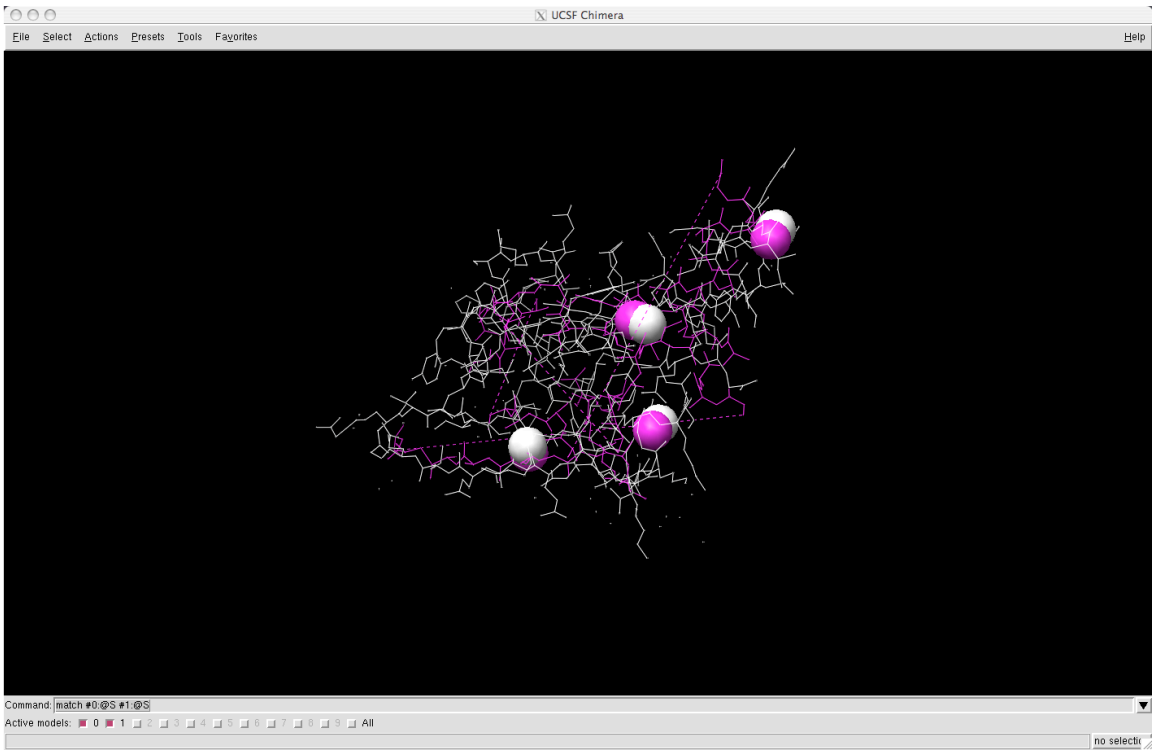

Figure 5.5a Manual 180° rotation of the target.pdb coordinates: This was used to determine the best aligning pairs of atoms.

This rotation of the 2QVO.pdb coordinates should be done carefully. In this case all four Sulfur positions were located during structure determination. This may not always be the case. A minimum of three atoms located was necessary to perform this superposition, but properly aligning three atoms is more difficult than four. Also the target.pdb image should not be moved at all as the 2QVO.pdb file will be saved based on its location relative to the target.pdb coordinates. Altering the target coordinates may upset efforts to accurately calculate the phase difference between the models.

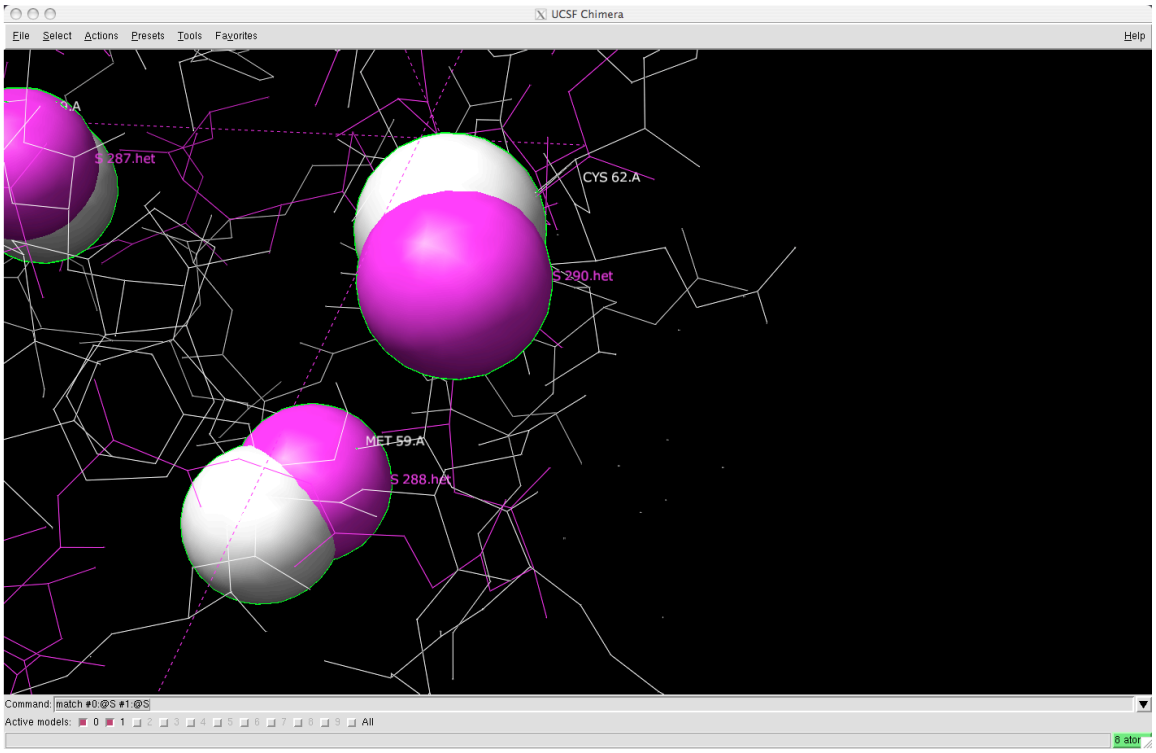

Figure 5.5b Heavy atom identification: Building off the initial match command superposition a 180-degree rotation and minor translation clearly indicates a better sulfur atom alignment between the two models. Increased magnification allows for identifying the appropriate corresponding sulfurs from each pdb file is important for correctly executing the match command in Chimera.

Once the proper orientation and corresponding Sulfur atoms have been determined the target.pdb file must be edited. As seen in Figure 5.5b, the 2QVO.pdb file contains sulfur atoms numbered 62 and 59, which correspond with sulfur atoms in the target.pdb file numbered 290 and 288. While not pictured atoms 91 and 19 in the 2QVO.pdb file correspond with 289 and 287 from the target.pdb file, respectively. The most efficient means of unifying the Sulfur atom numbering system between the files would be to edit the target.pdb coordinate files.

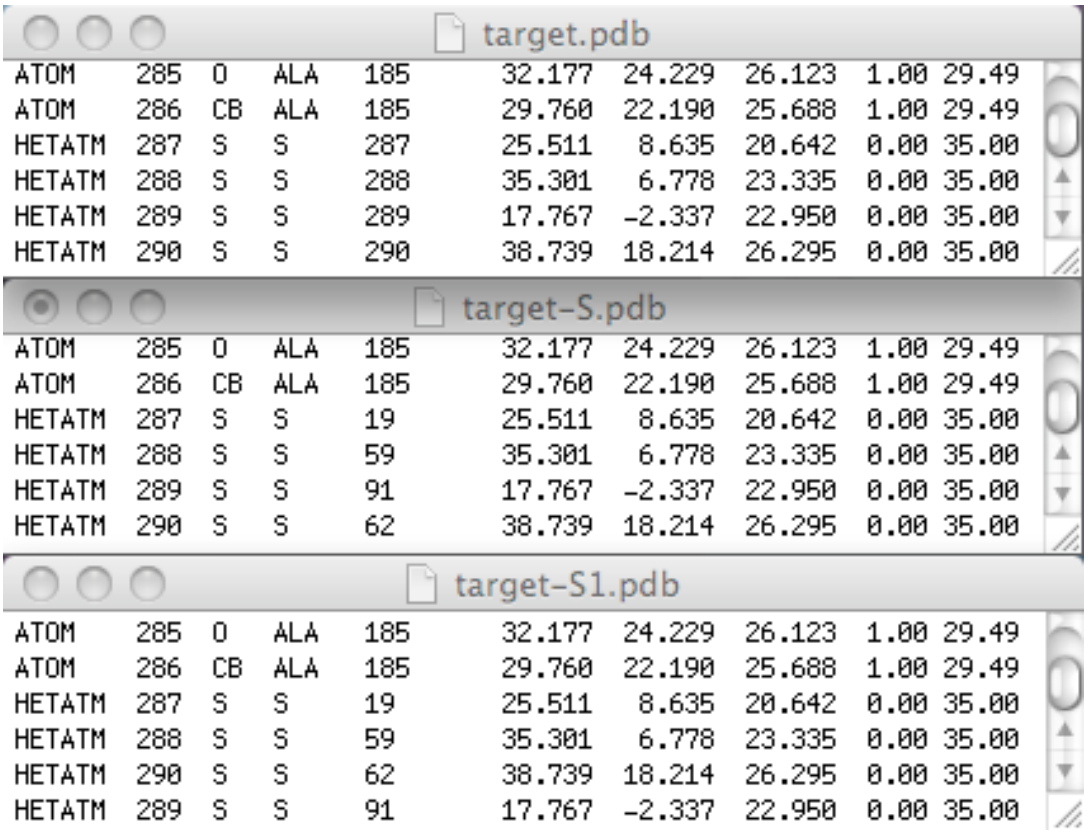

Figure 5.6 Re-formatting the Sulfur atom numerical designation: The Sulfur atoms within the target.pdb files need to be re-named and re-numbered to match the number and order of the Sulfur atoms within the 2QVO.pdb file. It is a good practice to rename the target.pdb file to include the edited items (i.e. target-S51) to avoid confusion.

At this point the user can now open the 2QVO.pdb and the newly edited target.pdb coordinates for final superposition. Note that as the changes were made (Figure 5.6) the files were saved using different names, this will be beneficial for organization purposes as well as use later in the phase comparison process. The Chimera command line executable for superposition of Sulfur atoms can now be used to achieve the best fit for the superposition trial.

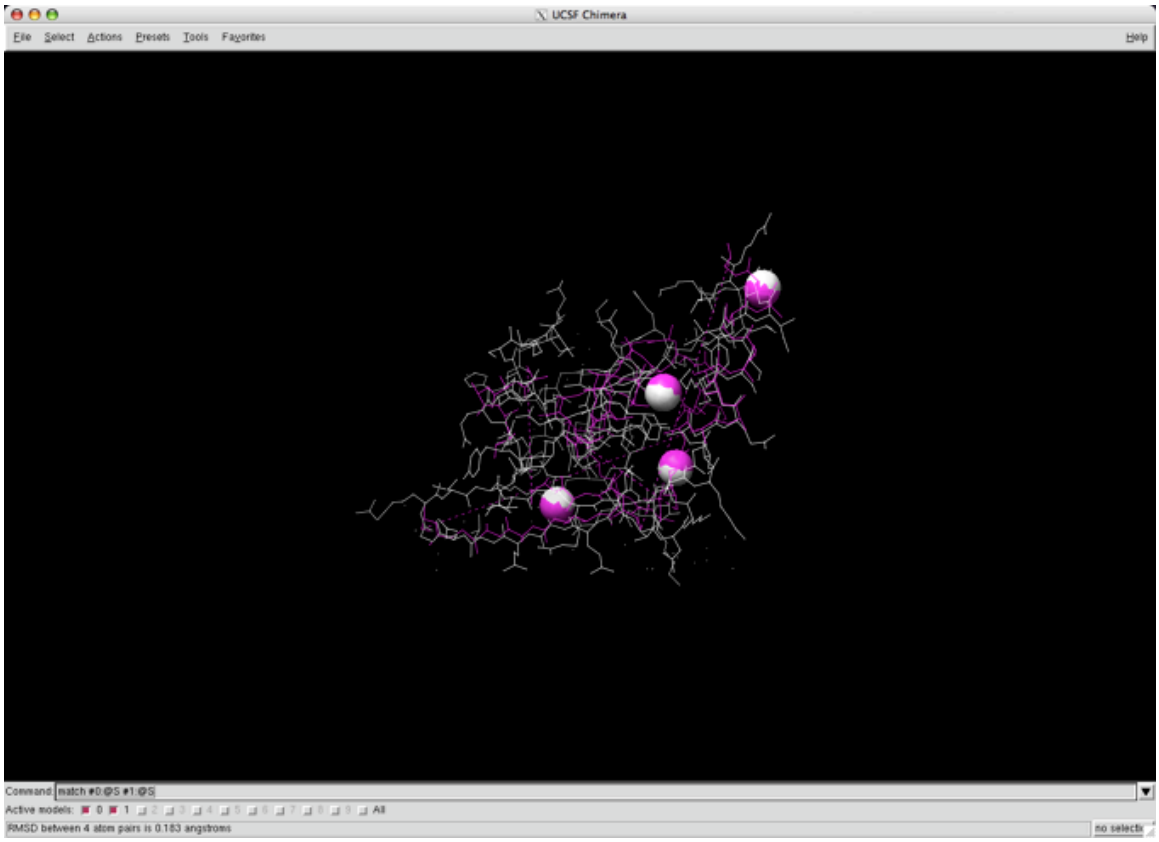

Figure 5.7 Finial pdb coordinate superposition: The Sulfur atoms within the target.pdb files and 2QVO.pdb coordinates are now labeled correctly and as such Chimera superposition fit is nearly perfect, RMSD 0.183.

Notice the RMSD value for this fit is markedly better than in the previous initial trial (5.24 vs 0.183) before renumbering and reordering the Sulfur atoms. The next step in acquiring the Phase Comparison between the two files involves refitting the transformed 2QVO.pdb file within its original electron density.

Refining

The $P4_2$ space group to which the 2QVO structure belongs is unique. Superpositioning two structures, fitting and refining the both pdb coordinates and their corresponding electron density is conducted by translating the one of the electron density maps (2QVO) by a symmetry related transformation to its partner (target) structure's coordinates and proceeding with

refinement. This however is not the case for the $P4_2$ space group. The superposition of a coordinate file and its corresponding electron density should easily coincide with viable locations within the symmetry related positions in the asymmetric unit. As luck would have it the $P4_2$ space group is prohibited from such an easy translation due to no consistent symmetry related origin for each solution generated pdb file. This space group is considered polar; groups with polar axis have no defined origin and thus possess more than a single indexing possibility. These facts add an additional degree of difficulty for the transforming the 2QVO electron density to its newly transformed coordinates in reference to each solution generated by the data reduction programs in this study

Reindexing

Numerous trials were conducted involving such electron density translations corresponding to the 2QVO.pdb file with limited success. I found that several but not all of the solutions from HKL2000 and PROTEUM2 required no transform of the 2QVO electron density maps to their coordinate system at the conclusion of superimposing. Yet several attempt were made to superimpose the 2QVO and electron density maps onto the coordinates generated using XDS, Mosflm, and d*TREK to no avail. It eluded me why I had results that worked in some instances and not in others. Utilizing the heavy atom positions as a reference point offers a viable solution to the problems with superimposing various models, standardizing target.pdb selection, and fitting routines for superposition trials. I hoped the phase information contained within the electron density would be translated using symmetry related positions within the unit cell; this however was not the case. The next hurdle to cross dealt with translating the phase information to the newly superimposed coordinates corresponding to the superimposed 2QVO.pdb file.

To achieve a phase comparison after successful superpositioning of the 2QVO.pdb coordinates the structure needs to be re-fit to the phases corresponding to its electron density map. This is achieved by refining the structure using REFMAC5 (104). This is an iterative process in which the atomic coordinates of each atom within the 2QVO.pdb file are arranged to best match the electron density derived from the phases of the molecule. In most cases the electron density does not match the pdb coordinates resultant refinement of the transformed 2QVO.pdb file and its electron density led to poor refinement statistics R_{fact} ~44% , R_{free} ~57% and eventual phase results of ~90 degrees. Ideally the statistics R_{fact} and R_{free} should remain within 10% of each other and lie in the range of 18-30%, a perfect phase comparison between two identical structure would be 0° but realistically 30° is considered excellent, 60° good and above 75° poor (a measurement of 90° is considered completely out of phase). To determine if the fitting performed by REFMAC5, I compared the R_{fact} and R_{free} values after refinement to those from the submitted coordinates in the PDB, if these are approximately equivalent (±3%) the refinement was considered acceptable. The reason for these refinement and phase comparison problems are linked to the $P4_2$ space group. As explained earlier, being polar in nature the $P4_2$ space group can be indexed with the c-axis freely orienting itself in either direction as long as the a and b axis are equal. This lack of a fixed lattice orientation allows for h,k,l to be reindexed to k,h,-l at random when a solution is considered. Thus, the coordinate transformation needed to superimpose the 2QVO.pdb file onto the target.pdb files was not a symmetry related operation. Consequently, quality refinement of the translated 2QVO electron density and superimposed 2QVO.pdb coordinates were not possible.

Refmac5:

```
#######################################################################
### CCP4 6.0: Refmac_5.2.0019    version 5.2.0019  : 06/09/05        ##
#######################################################################
Ncyc  Rfact   Rfree   FOM    -LL     -LLfree  rmsBOND  zBOND  rmsANGL  zANGL  rmsCHIRAL
 0    0.5688  0.5831  0.166  51048.  2601.3   0.0103   0.497  1.197    0.598  0.076
30 trails conducted…
 30   0.5530  0.6144  0.088  49638.  2531.6   0.0197   0.853  2.235    1.042  0.155
```

Phase Comparosion:

```
#######################################################################
### CCP4 6.0: cphasematch       version 0.1      : 06/09/05          ##
#######################################################################
Overall statistics:
Nrefl  <fom1>  <fom2>  <dphi>  w1<dphi>  w2<dphi>  wFcorr  wEcorr  Qfom1  Qfom2
2344   0.162   0.681   87.65   87.42     87.53     0.069   nan     0.125  0.039
```

Figure 5.8 Abbreviated output logs from REFMAC5 and Phase Comparison: These file illustrate typical results for superimposed 2QVO.pdb coordinates refined with original 2QVO.mtz (electron density).

To remedy the alternate indexing issue, a program entitled REINDEX (39) was implemented. This program applies a reindexing matrix to each h,k,l reflection resulting in a new unit cell and reduce reflections to the asymmetric unit. As long as care is taken in selecting the appropriate transformation matrix, meaning allowable for the specific crystal system, data collected from polar point groups can easily be merged or translated to appropriate coordinate systems. No changes other than the selecting the appropriate reindexing matrix were made to the default setting within the REINDEX program.

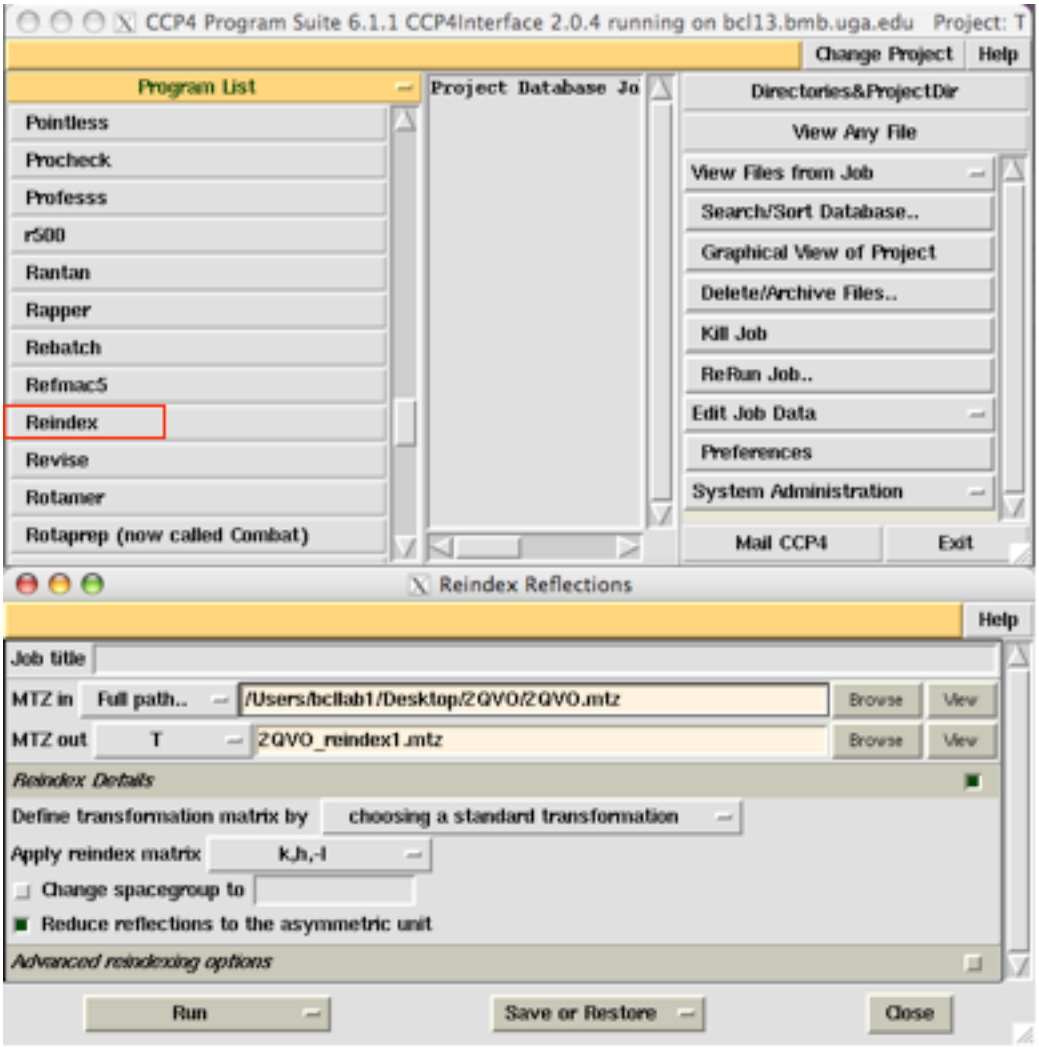

Figure 5.9 Operational GUI for REINDEX: CCP4 supported program, which produces a mtz file with the h,k,l reindexed according to the transformation type chosen.

REFMAC5

With the successful superpositioning and reindexing of the newly transformed 2QVO.pdb coordinates, the next step towards phase comparison trials involves refitting and refining the 2QVO.pdb coordinates into this newly reindexed 2QVO electron density. The goal of using REFMAC5 was to refine this structure is to mimic the refinement statistics from the original or untransformed 2QVO.pdb and corresponding 2QVO electron density (contained in a file format .mtz). As mentioned earlier success is evident when the R_{fact} and R_{free} values correspond with the

original values from the submitted 2QVO structure. To accurately execute REFMAC5, the pdb file to be fitted to the reindexed electron density will require editing. The process of superposition involving Chimera required altering of the Sulfur atoms of the original 2QVO.pdb file. REFMAC5 will not recognize this change in nomenclature and fail shortly after execution. For the Sulfur atoms used in superposition the S designation must be reclassified as SD for Methionines and SG for Cysteines. Also the same pdb file must contain a Cryst Card for REFMAC5 fitting, this line is lost during the superposition process. This value is identified as Cryst1 in the original 2QVO.pdb file and can be cut and paste to the new file. This being accomplished REFMAC5 will execute to completion.

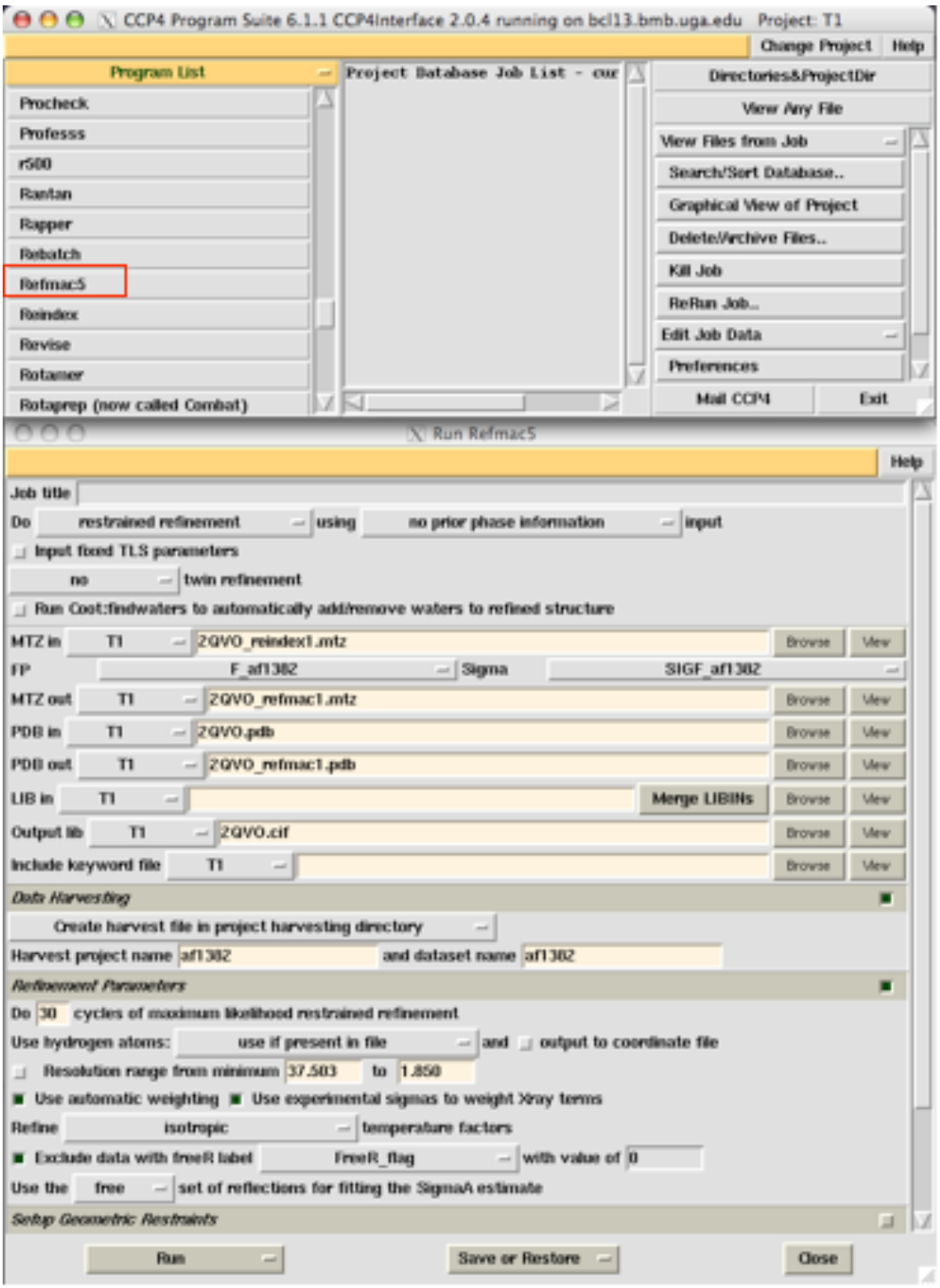

Figure 5.10 Operational GUI for REFMAC5: CCP4 supported program, can carry out rigid body, TLS, restrained or unrestrained refinement against X-ray data. Default values of restrained refinement using no prior phase information was used in this study.

The REFMAC5.log files are easy to compare as a check that the quality of the refinement Rfact and Rfree (**Figure 11**).

Original 2QVO.pdb file

```
############################################################
### CCP4 6.0: Refmac_5.2.0019    version 5.2.0019  : 06/09/05     ###
############################################################
Ncyc   Rfact   Rfree   FOM     -LL     -LLfree   rmsBOND  zBOND   rmsANGL  zANGL rmsCHIRAL
  0    0.2379  0.2822  0.772   42683.   2228.3   0.0103    0.495    1.192    0.595    0.075
30 trials conducted…
 30   0.2171  0.2701  0.787   41862.   2199.2   0.0223    0.899    1.840    0.881    0.126
```

Superimposed 2QVO.pdb file

```
############################################################
### CCP4 6.0: Refmac_5.2.0019    version 5.2.0019  : 06/09/05     ###
############################################################
Ncyc  Rfact   Rfree   FOM      LLG     rmsBOND  rmsANGLE   rmsCHIRAL
  0     0.477  0.467   0.294    50063.6  0.010      1.192        0.075
30 trails conducted…
 30    0.223  0.271   0.783    44082.1  0.019      1.641        0.110
```

Figure 5.11 REFMAC5 output log file: Illustrating the refinement quality associated with the originals and superimposed 2QVO.pdb and corresponding electron density

These values being approximately the same in both cases the phase comparison trials move on with the next program, CAD(39).

CAD

The CAD program is a useful tool for combining or deleting reflection data usually listed in column format contained within mtz files. All or only selected columns can be combined from two or more mtz files. Unlike REINDEX or REFMAC5, the default values for this program will not suffice for acquiring the necessary file for phase comparison studies.

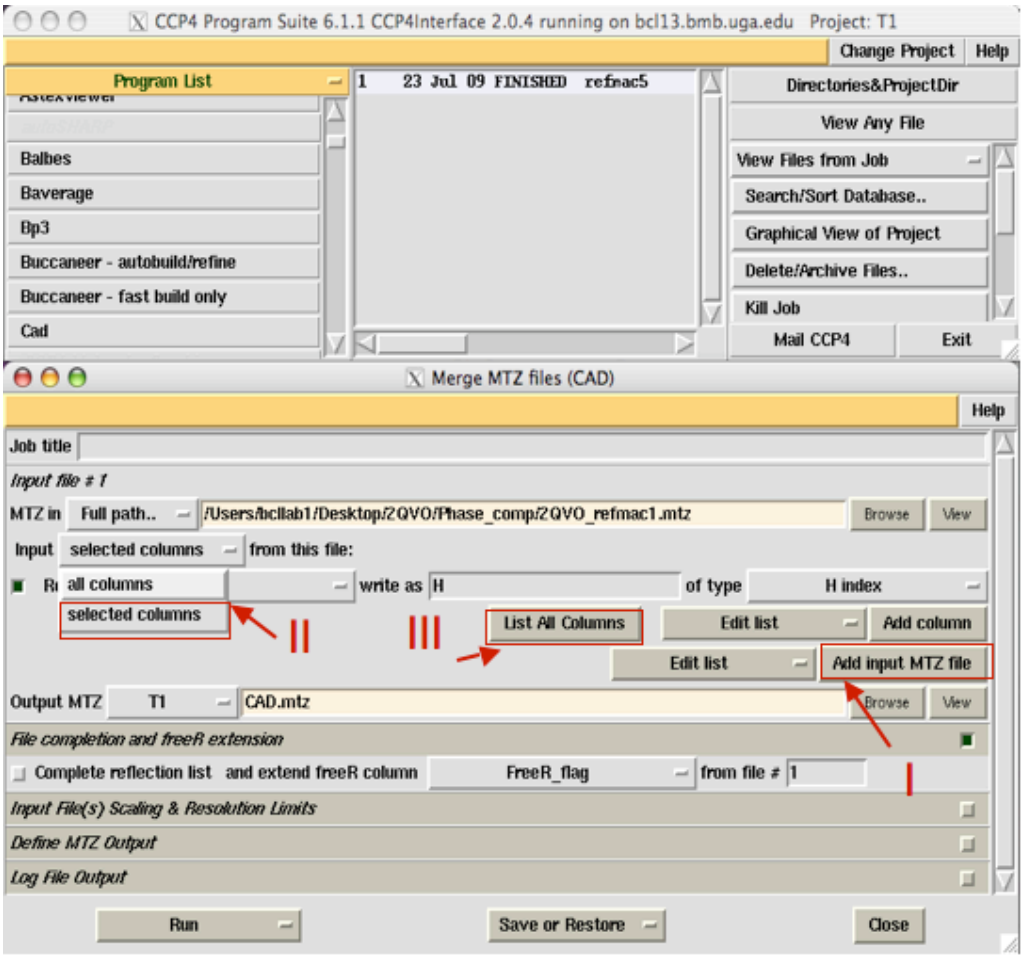

Figure 5.12 Operational GUI for CAD: A CCP4 supported program, highlighting the needed options for combining the transformed 2QVO.mtz and the corresponding target.mtz.

The first of the two mtz files which will be combined must be individually loaded into the CAD GUI. The first .mtz file is loaded via the browse button, to add the second mtz file→ the Add Input MTZ file (I) is selected first, and the second mtz file loaded just as the first. From the Input selector choose→ selected columns [II] and select the→ List All Columns [III] in order to see the names and contents of the columns within both mtz files.

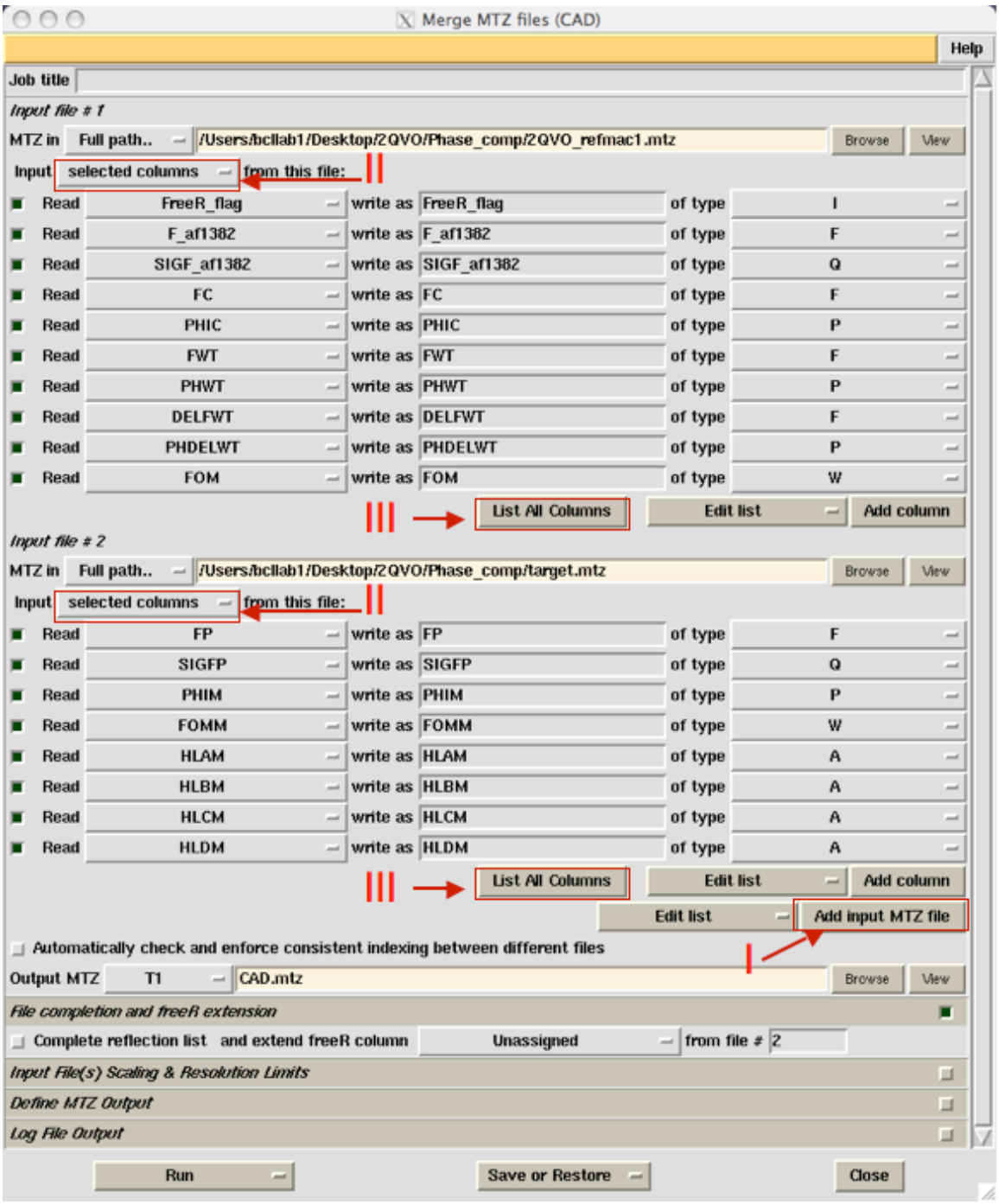

Figure 5.13 The extended GUI for CAD: A CCP4 supported program, with MTZ column labels displayed.

The output file from CAD combines the two submitted mtz files and their corresponding column entries. For the purposes of phase difference calculations, the columns of interest are PHIC (calculated phase) and FOM from the 2QVO_refmac1.mtz and PHIM (most probable

phase) and the FOMM (figure of merit, the use of two FOMM distinguishes the two FOM values) from the target.mtz file. The finial programs used in this work are PHASEMATCH and PHISTATS.

Phase Analysis

Phase Comparison

The PHASEMATCH program uses the cphasematch ver1.0 subroutine as a method of comparing the phase solutions from two combined sets of phasing data.

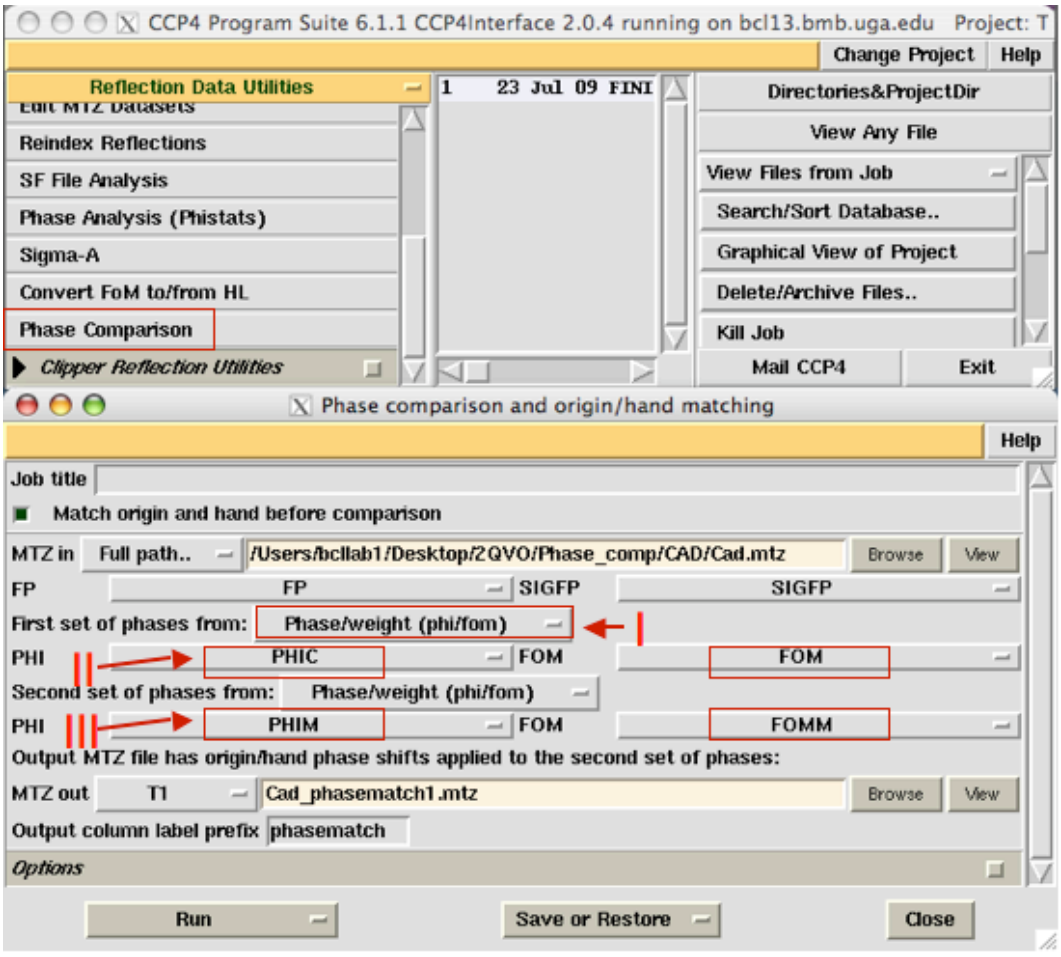

Figure 5.14 Phase Comparison GUI: A CCP4 supported program, with evaluation type selector (I), and appropriate column label to identify phase information (II, III).

To execute this program, first the desired mtz file output from CAD is loaded in to the PHSEMATCH. The Phase/weight (phi/fom) is selected from the option for the first and second set of phases (I), then choosing PHI values (II) and FOM (III) from each combined mtz file.

Phase Comparison

```
 ###################################################################
 ### CCP4 6.0: cphasematch        version 0.1       : 01/06/04##
 ###################################################################
User: bcllab  Run date:  3/ 7/2009 Run time: 09:35:38
Overall statistics:
  Nrefl  <fom1>  <fom2>  <dphi>   w1<dphi>  w2<dphi> wFcorr wEcorr   Qfom1  Qfom2
  3389    0.799     0.652     60.75      57.34       53.35      0.635    nan       0.490    0.610
                              Delta Phi
Nrefl =
Fom 1,2 = figure of merit from data set 1, data set 2
dphi =  Average difference in degrees
w1,2 <dphi> = degree difference with weight 1, weight2
wF/wEcoor =
Qfom1,2 =
```

Figure 5.15 Excerpt from PHASEMATCH log file: CCP4 supported program used in phase comparison studies. The values of the Overall statistics are listed within the figure.

The above result represents a “good” solution; we do not expect any of the values generated within the phase comparison portion of this work to register as excellent because the initial model used to solve the structure only registered a <dphi> value of 58.48.

PHISTATS

The PHISTATS program was used as a second validataion of the phase difference generated in PHASEMATCH. PHISTATS also performs a similar analysis to Phase Comparison between two phase sets using FOM as weighting factors. In the PHISTATS GUI the input files are quite similar to those from PHASEMATCH. The input columns PHIC and PHIM [I] accompanies their corresponding FOM and FOMM [II] column values from the mtz file produced using CAD.

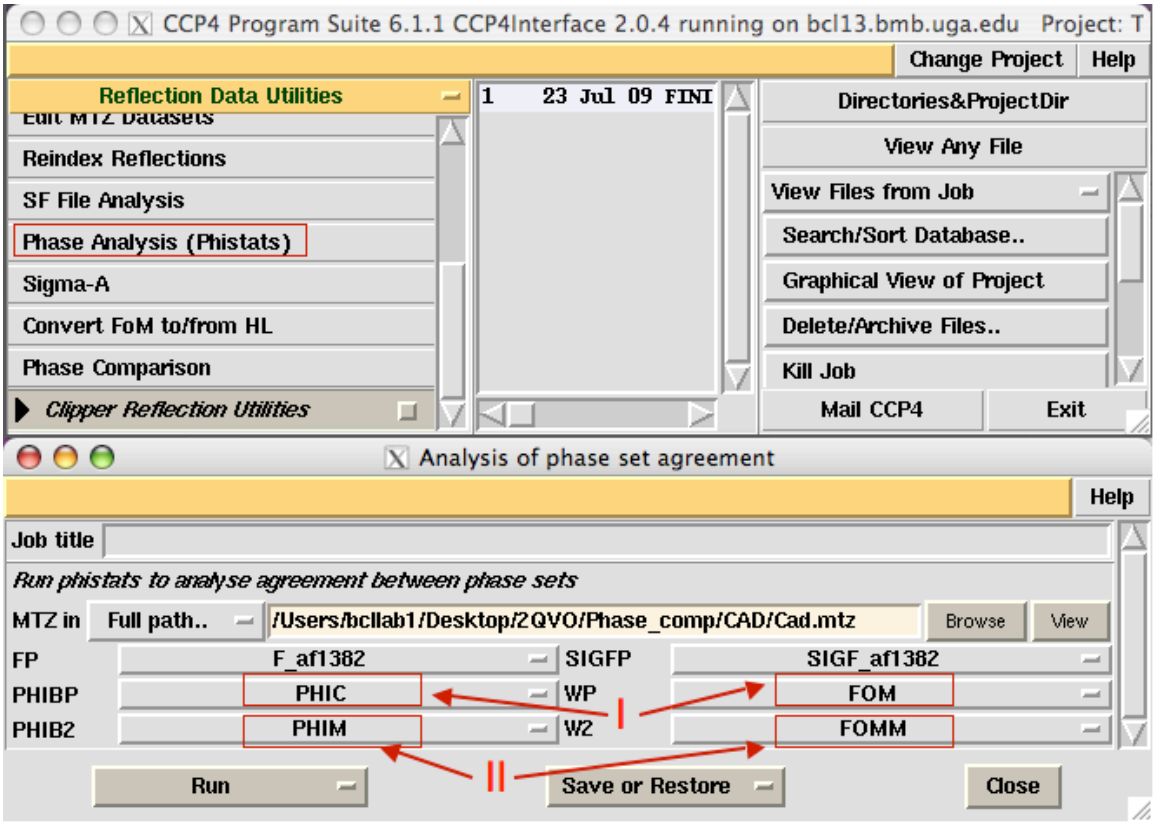

Figure 5.16 The extended GUI for PHISTATS: A CCP4 supported program, with appropriate phase information labeled [I, II].

An excerpt from the log files from PHISTATS (Figure 1.15) illustrates the pertinent statistics pertaining to the delta-phi value from each program.

```
PHISTATS
#####################################################################
### CCP4 6.0: PHISTATS          version 6.0       : 06/09/05##
#####################################################################
User: bcllab  Run date: 15/ 7/2009 Run time: 12:12:40
DEL = Average Difference in degrees
CW1 = Correlation with weight 1
CW2 = Correlation with weight 2
CWW1 = Correlation of cos(phase error) with weight 1
CWW2 = Correlation of cos(phase error) with weight 2
Range  Limits   Total No     DEL       CW1     CW2    CCW1   CCW2
TOTAL             2965      59.277   -0.252  -0.332   0.250  0.330
                          Delta-Phi
```

Figure 5.17 Excerpts from the PHISTATS log file: This file contains the information devised during the comparison of two different phase calculations.

The results from PHISTATS are consistent with those from PHASEMATCH. To verify the accuracy of both PHASEMATCH and PHISTATS both programs conducted a phase comparison of the original 2QVO phase information with itself. As expected the phase difference generated from both programs was null. In conclusion, a quantitative method to verify the quality of each data processing program solution has been established. This technique, in concert with the statistics gathered in the Results section, are clear indicators that choice of data reduction program does in fact effect the success rate of Sulfur-SAD phasing.

Chapter 6

Results

The judicious use of data reduction software is imperative to completing a successful diffraction experiment (105). The actions involved with data reduction commonly rely on the experimenters prior knowledge concerning the protein(s) in question or knowledge of the programs being used. Collecting the highest quality data and considering the limitations of the experiment are nearly as important as using the best possible method for interpreting the data. During the last decade crystallography has evolved from time and knowledge intensive experimental trials toward a fast paced "black box" data collection and processing field of study. This is largely due to advancements in detector hardware and beamline automation used primarily for remote data collection and processing (40, 65, 106, 107). As advancements in hardware and software reduce the time and effort required to conduct experiments, under opportune circumstances, the necessity of experienced crystallographers has diminished. There are, however, many special cases similar to the object of this study that would effectively leave a novice user stranded (108).

Here the results are reported from taking the five most popular data reduction programs to process data sets of mediocre resolution with the intent of solving the structure of a 95-residue protein via S-SAD phasing. The values are listed by the name of the program used. In the case of HKL2000, the suffix -ZQF represents the original processing done by Dr. Albert Fu, -JTS-1 denotes a novice approach, and –JTS-2 portrays the full use of both my experience using HKL2000. I found PROTEUM2 to be the best data reduction program, even surpassing the HKL2000 processing efforts of Dr. Albert Fu**.**

R-factors

To assess the effects of data reduction crystallographers commonly use an R-factor statistic termed R_{sym}, R_{merge} and R_{meas} which are referred to as a "reliability factors". Traditionally acceptable values for R_{sym} and R_{merge} are <5% for excellent data, 6-10% is considered usable, 10-20% presents questionable results and +20% being errant. R_{meas} will be discussed later. The foundation of R-factors measures the accuracy of data as a ratio:

$$\text{R-factor} = \frac{\text{Average difference between values}(h,k,l \text{ vs } -h,-k,-l)}{\text{Mean magnitude of measured values}}$$

This measurement is treated slightly different depending on the R-factor chosen by the data reduction program during scaling. Traditionally R_{sym} and R_{merge} were treated separately. During the early days of crystallography, data were recorded using film and a precession camera. Arndt (109) proposed a term to indicate the reliability of diffraction intensities by examining the relationship between symmetry related intensities on the same film as R_{sym}. This could also be applied to identical or symmetry related intensities collected within a data set containing multiple images. However, it was often necessary to collect data using multiple crystals due to radioactive decay. To combine multiple diffraction images required merging intensities resulting in a term similar in purpose to R_{sym} referred to as R_{merge}. As a rule, data collected using a single crystal and resultant single orientation matrix would use R_{sym} as an indicator of reliability. Combining data requiring multiple crystals or adjustments of crystal position, both resulting in calculationg of a new orientation matrix, would use R_{merge} to accomplish the same goal. With the advent of area detectors and crystal cryocooling, typical data sets consists of hundreds of images containing multiple measurements of identical and symmetry related reflections from a single crystal. The interchangeable use of these terms evolved due to data reduction programs treating

individual data frames as complete data sets. Conducting refinment during integration requires calculating new orientation matrices per image then merging resultant intensities. Still other data reduction programs treat all frames as a single continuous data set by considering three-dimensional spot positions before merging intensities and refining the orientation matrix using batches of images (approx 30-40 images). $R_{sym/merge}$ values could effectively create an ambiguity concerning which term best suits the method of data reduction (Weiss, personal communication). As noted at the conclusion of each data reduction program in Chapter 3, the R-factor used for determining the reliability of data processing were highlighted. Of the programs included in this study, HKL2000, d*TREK, PROTEUM2 report R_{sym} or R_{merge} statistics during scaling.

R_{sym} (or R_{merge}):

$$\sum_h \sum_i \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

The summation over *h* represent the unique reflections (*h,k,l*) while the summation over *i* spans all the symmetric equivalents of *h*. I_{mean} is the statistical average of all symmetry related observations of a unique reflection. This value is often used as a measure of X-ray diffraction quality to date (90, 110).

Despite this fact, R_{sym} has been shown inferior as a complete representative of diffraction quality to values such as R_{meas}. As redundancy increases within the data, it is common to witness an increase in R_{sym} values. This is due to a lack of a correction term to remove the additive effect of drastically increasing the number of intensities measured when using highly redundant data (87, 111). The redundancy dependence of R_{sym} may become problematic when judging results between low and high redundancy data. A low redundancy data set could produce values that appear superior to higher redundancy results, thus skewing the intended function of R_{sym} as an indicator of data quality.

An alternate indicator of data quality has been proposed by Diederichs and Karplus (87) to remove the redundancy dependence of R_{sym}. This value, R_{meas}, includes a term $\sqrt{[n/(n-1)]}$ which appropriately weights individual reflections (h) according to their multiplicity (n_h).

R_{meas}: $$\sum_h \sqrt{n_h/(n_h-1)} \sum_i \left| I_{hi} - I_{mean} \right| \Big/ \sum_{hi} I_{hi}$$

At present only SCALA and XDS calculate this R-factor during scaling. These programs also report R_{sym} values. R_{meas} values are typically larger than R_{sym}, which automatically raises doubt when interpreting data as the crystallographic community traditionally associates low R-factors with better data. According to research conducted by Weiss and Hilgenfeld (111), a given 10σ reflection with a redundancy of two produces an R_{sym} of 5.6% while increasing the redundancy to five increases this value to 7.4%. R_{meas} values tend to remain constant despite increases in redundancy by including the $\sqrt{[n/(n-1)]}$ term as redundancy data is included during processing.

R_{sym}

Program /Data	R1	R2	R1-R2
HKL2000-ZQF	4.1(17.6)	4.5(14.2)	5.6(32.0)
HKL2000-JTS-1	4.2(16.9)	4.1(9.1)	38.8(51.8)
HKL2000-JTS-2	4.4(37.3)	5.8(26.6)	5.8(32.4)
PROTEUM2	4.6(15.2)	4.7(10.9)	5.1(13.7)
d*TREK	5.8(23.6)	5.5(19.2)	6.5(22.7)
XDS	5.8(59.0)	7.1(43.5)	6.9(58.8)
MOSFLM	8.3(21.7)	8.3(14.7)	8.9(19.2)

Figure 6.1: Comparison of R_{sym} values between programs: Except for XDS and MOSFLM the R_{sym} values are all < 5%, which represents excellent results. The values generated by XDS and MOSFLM are within the 6-10% which is considered usable.

R_{meas}

Program /Data	R1	R2	R1-R2
XDS	6.3(63.7)	7.7(47.3)	7.1(61.5)
MOSFLM	9.0(23.4)	8.9(15.8)	8.9(19.2)

Figure 6.2: Comparison of R_{sym} values between programs: The values may be beneficial but until other data reduction programs adopt this measure we cannot make a useful comparison.

Redundancy for the R1, R2 and merged R1-R2 data sets were 12.4, 13.5 and 25.4 as reported by HKL2000. The effect of the multiplicity term on a single reflection does not appear to translate to an entire data set as seen in the above results. As expected, the R_{meas} values are higher than R_{sym} however the effect of redundancy on the magnitude of R_{sym} is not apparent compared with R_{meas} values. Even though the results of this study does not coincided with trends in R_{sym} and R_{meas} predicted by Diederichs (95), it should be noted that neither XDS nor MOSFLM provided an accurately phased solution. Only with comparable results to PROTEUM2

or HKL20000 could we make a true determination of R_{meas} and R_{sym} validity for the R1, R2 and merged R1-R2 data sets.

I/σ_I

Calculation of a three dimensional atomic structures rely heavily on accurately measuring the intensities of diffraction peaks. The most commonly used technique for measuring diffraction intensities is profile fitting. This method creates an average spot profile by considering the habit of diffraction spots within specific sections the detector, depending on the data reduction program. The intensity measurements captured during integration rely purely on the intensities contained within the average profiles which are placed at predicted spot locations throughout the data. Error estimate directly associated intensity measurements are based on a counting statistics and expressed by σ_I. Two of the most prevalent factors that produce error include noise generated from unintended X-ray scattering (air and Compton) and those that occur when predicted spot profiles omit a portion of a reflection.

Errors associated with air and Compton scattering are unavoidable and arise from the quantum nature of X-rays. However, a method of effectively approximating these errors using a Poisson distribution of counting statistics is well known (112). The error represented by σ_I depends on the number of counts recorded per pixel within the average profile fitting curves. If we consider N as the total number of counts from which σ_I is calculated, $\sigma_I^2 = N$ (Citation). Within each diffraction spot the N counts originate from both peak and background such that the error associated with N is better expressed as $N = N_{peak} + N_{background}$ such that σ_I is also the sum of the error corresponding to the expanded N parameters; $\sigma_I^2 = \sigma_{Peak}^2 + \sigma_{Background}^2$. The correct method of adding independent error employs quadrature summation (113). As the peak and background measurements are not directly related, the individual error will be less than the sum

of the error. The importance of minimizing the errors associated with σ_I has recently received considerable attention involving S-SAD phasing (22).

The second of these errors is related to user and program interactions. Proper indexing will produce accurate spot predictions. If the predicted spots do not encompass the entire diffraction spot, valuable intensities will be lost and incorrect profiles will be averaged for use during profile fitting. With proper indexing and error correction, the uncertainties resulting from intensity measurements can be addressed such that results based on the signal to noise ratio (I/σ_I) may offer an indication of data quality.

Obtaining phase information is strictly accomplished by measuring the differences between diffraction intensities. S-SAD experiments typically contain a low percent contribution of anomalous signal (1-2%) within the diffraction peaks. The errors, which each program attempts to correct for, can play a significant role in either discovering the appropriate phases or losing anomalous contribution during processing.

I/σ_I

Program /Data	R1	R2	R1-R2
HKL2000-ZQF	72.3(3.3)	93.7(17.6)	82.5(10.7)
HKL2000-JTS-1	75.4(12.5)	86.0(26.9)	80.8(10.6)
HKL2000-JTS-2	55.0(3.2)	62.5(10.5)	82.8(10.7)
PROTEUM2	47.7(12.6)	62.3(16.6)	55.5(10.4)
d*TREK	42.1(11.9)	52.5(14.5)	62.0(12.3)
XDS	25.2(3.6)	19.7(5.3)	31.0(4.6)
MOSFLM	22.6(8.7)	23.6(12.1)	32.6(14.3)

Figure 6.3: Comparison of I/σ_I values between programs: The average intensity from selected reflections termed <I> are divided by the average standard deviation, <σ>, of the reflections. This value depends directly on the spots number of spots used as well as the method of determining intensity errors between each data reduction program.

The specific method employed by data reduction programs used to calculate I/σ_I values are not consistent. Large I/σ_I values would be considered advantageous for the detection of anomalous signal especially in the case of S-SAD phasing. HKL2000 consistently recorded the highest I/σ_I values for the R1, R1 and merged R1-R2 data sets. Using I/σ_I as a measure of data quality is a commonly used practice. These values are most useful when determining the high resolution cutoff for a data set. Typically values of $I/\sigma_I < 2$ indicate a poor signal to noise ratio and the minimum resolution which should be used for data processing. The inherent danger of using large I/σ_I as a measure of overall data quality is easily seen in the apparent high value results of HKL2000 for R1 and R2. PROTEUM2 reported I/σ_I values approximately 30% less than those determined by HKL2000, yet the structure was solved by using either R1 or R2 data sets independently. The I/σ_I results from MOSFLM and XDS are low compared to the other data reduction programs used but cannot be directly linked to the lack of results produced by each program.

Anomalous Signal Measurements

Each of the scaling algorithm used in this study offer a measurement of the anomalous signal to noise ratio present determined during data processing. 3DSCALE scaling routine, used for d*TREK and PROTEUM2, offer a Ras value. SCALEPACK, from HKL2000, offers a graphical representation of anomalous signal by analyzing χ^2 values versus resolution. SCALA compares reflections from different portions of the data set in a similar fashion to 3DSCALE producing a correlation beteen anomalous signal and resolution. XDS utilizes two measurements – Anomal Corr (Anomalous Correlation) and SigAno to analyze the anomalous contribution of intensity measurements.

3DSCALE exploits the innate characteristics of acentric versus centric diffraction reflections;

$$Ras = \Delta a / \Delta c$$

Δa represents a calculated ratio of acentric reflections, which is equivalent to the differences seen in Bijvoet pairs (h,k,l vs –h,-k,-l) divided by the aforementioned error σ_I. This value is a indicator of anomalous contribution based on intensity differences while considering the error which accompanies these values.

$$\Delta a = \left\langle \Delta_I / \sigma_I \right\rangle_a$$

Δc is nearly equivalent to Δa with the exception of using centric reflections. These are reflections related through the space group's point symmetry and contain no anomalous contribution. In theory these reflections should contain identical intensities ($\Delta_I = 0$), any differences found between centric reflections are used as a indication of the noise present throughout the data set.

$$\Delta c = \left\langle \Delta_I / \sigma_I \right\rangle_c$$

Ras values can offer a viable signal to noise analysis based solely on anomalous scattering. From research conducted by Fu (114), values of 1.5 or greater at approximately 3.0Å are an indicator of excellent anomalous contribution from Sulfur atoms and a high likelihood of proper phasing. The results of this study may extend the phasing limits for Ras values at 3. 0Å.

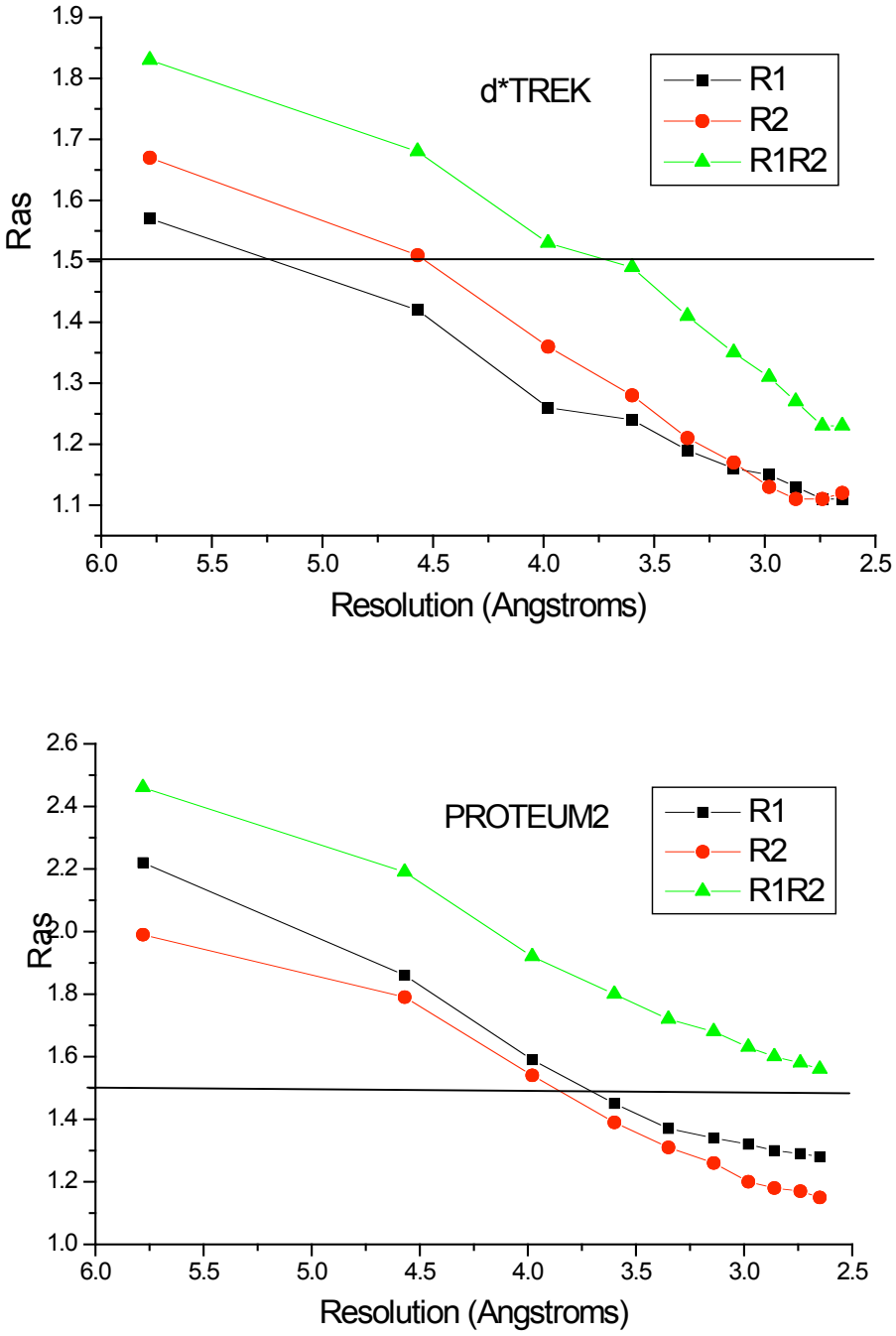

Figure 6.4: Comparison of Ras **values between d*TREK and PROTEUM2:** Observing accepted Ras thresholds for structure solution (> 1.5 @ 3.0 Å) it is clear d*TREK fails to meet this mark yet a solutions was generated for all three data sets exceeding pre-conceived thresholds.

An accepted threshold value of 1.5 (Ras) from 3DSCALE is an established indicator that significant anomalous signal has been achieved for successful protein phasing (114). The data generated by 3DSCALE for both the R1 and R2 data processed with PROTEUM2 seem to set new benchmarks for the minimum Ras threshold value. It appears PROTEUM2 exceeds d*TREK by both providing a new consideration of Ras thresholds for successful protein phasing and in the actual phasing of AF1382.

The HKL2000 GUI outputs several graphs at the conclusion of scaling, one of which displays the χ^2 and R-Factor vs. Resolution. If anomalous signal is present the χ^2 values will be greater than one and contain a clear dependence on resolution. However, since the merged R1-R2 processing required the use of command line execution of scaling scripts, no information concerning this graphical analysis is displayed or recorded. I concluded since the calculation necessary to generate these values was performed by SCALEPACK a script must exist for this calculation within the HKL2000. After reviewing the HKL2000 manual I was able to successfully edit an existing script to accomplish this task.

```
scalepack << eof
number of zones 8              ! Number of resolution shells for statistics
estimated error 0.0 0.0 0.0    ! Estimated error for each resolution shell
0.0 0.0 0.0 0.0 0.0
error scale factor 1.0         ! Multiplicative factor applied to input σ
number of iterations 0         ! Number of scaling attempts made
output file 'junk.sca'         ! file output by scalpack
format scalepack               ! Format of the input intensity data
file 1 'best-result-via-Chi^2-ajustments.sca'  ! file to be used during processing
eof
```

Table 4.5.7: SCALEPACK anomalous correlation script Executed in the same manner as the scaling script devised by Dr. Fu, the resultant χ^2 trend versus resolution displays the anomalous signal detected by HKL2000

Using this tool I compared the results from both Dr. Fu and my own processing attempts as follows;

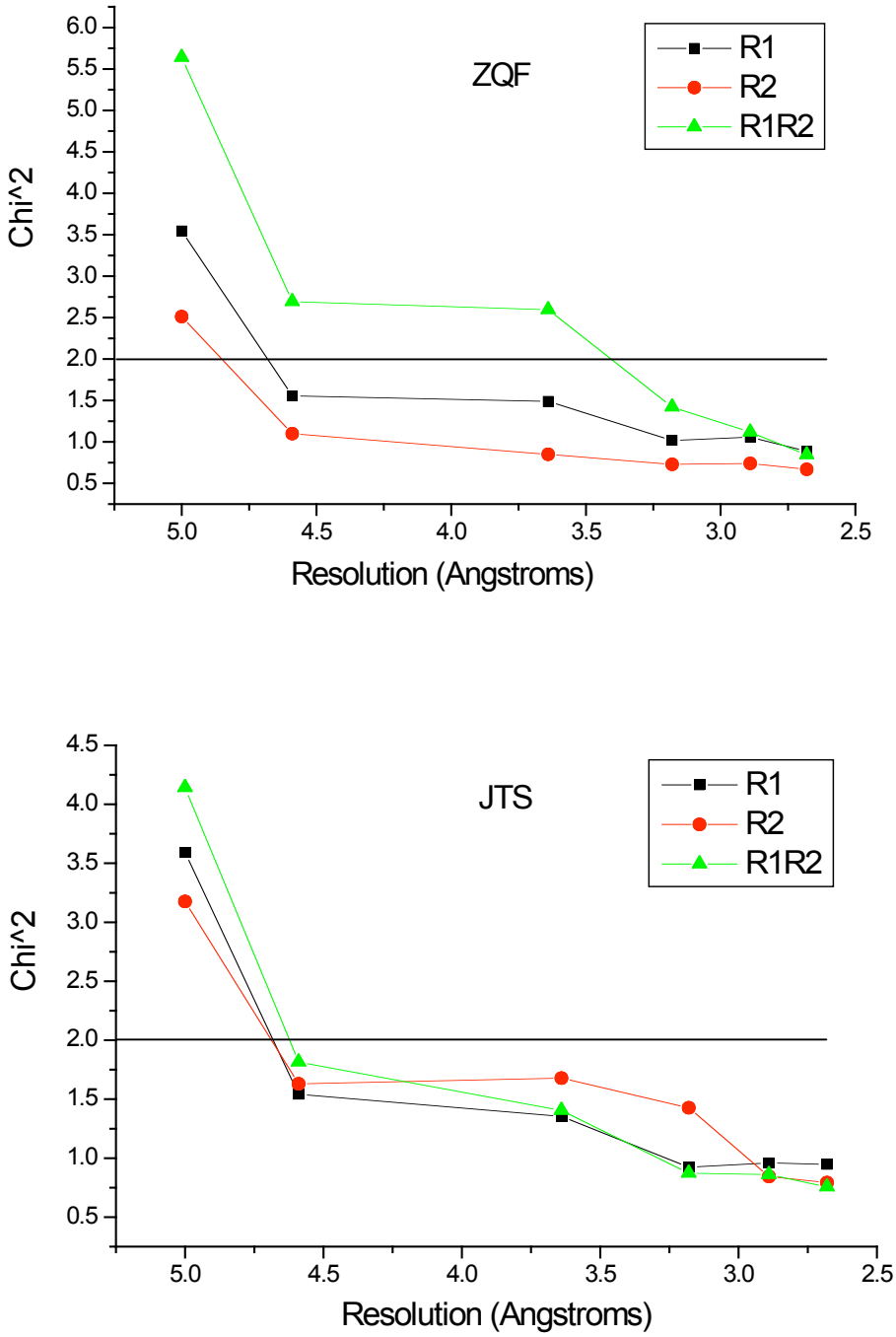

Figure 6.5: Comparison of χ^2 values within HKL2000: The χ^2 value test indicates the significance of the anomalous differences. This test depends heavily on error approximations conducted by HKL2000. χ^2 values are greater than 2 throughout the resolution range is an accepted indication of acceptable anomalous signal. However both the JTS and ZQF merged R1-R2 trials produced adequate solutions despite only ZQF results maintaining a value above 2 throughout most of the resolution range.

The anomalous signal detected by HKL2000 are more subjective than other programs studied in this work as χ^2 values are quite sensitive to errors due to air absorption, detector orientation, ect. The minimum threshold for anomalous signal corresponds to χ^2 values above 2.0 in the lowest resolution shell. Average values of +60 in the lowest resolution shells are expected in the case of excellent anomalous contribution (115). The values presented above display the weak anomalous signal inherent to S-SAD phasing experiments.

SCALA calculates measure of the anomalous contribution from a given data set by comparing reflections between random halves of the data. These halves will contain reflections which, if the redundancy within the data set is greater than 4, can be compared to detect anomalous signal. This is similar to the method used by 3DSCALE without consideration of the ratio of acentric to centric intensities. SCALA outputs a correlation analysis in graphical format. This displays the correlation coefficients, indicators of anomalous signal – centric data –average intensity, as a function of resolution. This method of anomalous assessment provides an indicator of both anomalous signal at various resolution ranges and a method of determining the reliability of the signal. As the resolution increases fluctuations, involving reflection intensities are considered as an indicator of the useful resolution for particular data sets.

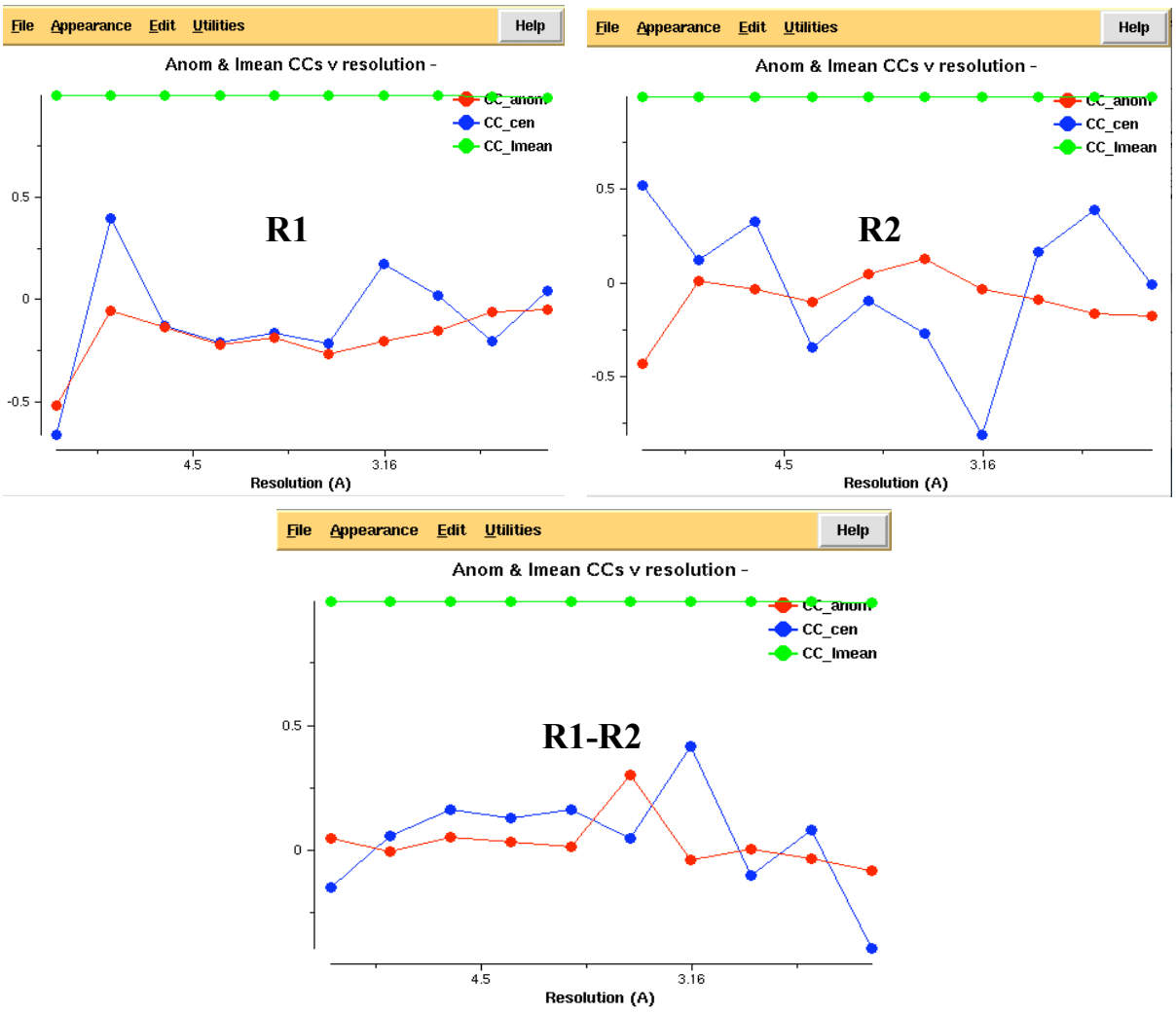

Figure 6.6: Anomalous Correlations within a Single data set: MOSFLM performs a correlation between random halves of the data by comparing acentric reflection intensities. This is similar to the Δa portion of 3DSCALE's Ras calculation.

The plot illustrates the anomalous correlation in red. A significant anomalous signal should contain values between 0.75 and 0.4 at low resolutions > 3.5Å, while heavily fluctuating trends and negative values are an indication of poor anomalous signal detection. The values displayed in green are the average intensities of the data which should remain constant at low resolution. The blue plot is the correlation between centric data and resolution. In theory, this value should be zero, but as mentioned earlier during the explanation of the Ras approximation, the differences in this value are a representation on noise or background contamination. The more

redundant the data the lower fluctuations in from the theoretical value (zero) should be observed. No significant anomalous signal was detected from the R1 or R1-R2 merged data. The R2 data seems to contain sufficient signal according to SCALA documentation, yet no solution was generated (86).

XDS outputs two values as indicators of anomalous contribution. These are listed in the scaling output file in a column format similar to 3DSCALE. The first of these is labeled SigAno**.** This represents the average anomalous difference between F+ and F- structure factors obtained form merged observations. After correspondence with the supporting authors of the program, I learned that values for SigAno which indicate a statistically significant anomalous signal are >70% in the lowest resolution shell. As resolution increases, the anomalous signal will decrease in magnitude with any values below 30% representing noise.

The second indicator of anomalous signal is titled Anomal Corr. This correlation factor measures the average differences between random subsets of data, interpreting inequalities of intensity measurements as indicators of anomalous signal. Values of 1.5 or greater are indicators of significant anomalous signal at 3.0Å resolution.

R1- Anomalous Statistics

RESOLUTION LIMIT	Anomal Corr	SigAno
7.82	**74%**	1.470
5.58	61%	1.265
4.57	49%	1.037
3.97	17%	1.089
3.55	38%	1.423
3.25	23%	1.158
3.01	35%	**1.241**
2.81	60%	1.410
2.65	64%	1.274

R2- Anomalous Statistics

RESOLUTION LIMIT	Anomal Corr	SigAno
7.76	**59%**	1.182
5.54	55%	1.040
4.54	27%	0.770
3.94	28%	0.954
3.52	36%	1.460
3.22	9%	0.937
2.98	29%	**0.983**
2.79	29%	0.881
2.63	9%	0.730

R1-R2-Anomalous Statistics

RESOLUTION LIMIT	Anomal Corr	SigAno
11.79	**74%**	2.056
8.33	73%	1.514
5.89	63%	1.525
4.81	58%	1.208
4.17	39%	1.047
3.73	8%	1.055
3.40	21%	1.095
3.15	11%	0.871
2.95	1%	**0.829**
2.78	38%	0.792
2.64	28%	1.004

Figure 6.7: Comparison of Anomalous Signal by XDS: During scaling Anomal Corr and SigAno values are generated to illustrate the anomalous intensity differences within the data sets.

Surprisingly, the scaling results from R1 and merged R1-R2 data sets produce SigAno values which indicate XDS has identified a significant anomalous signal while results from Anomal Corr are below the desired threshold. The reason for this in congruity may be linked to the results from PROTEUM2 scaling by 3DSCALE. As mentioned earlier, the accepted threshold for Ras values reflecting significant anomalous signal were the same as XDS, ≥ 1.5 at 3.0Å. Yet PROTEUM2 produced excellent results while recording Ras values of 1.3 and 1.2 for data sets R1 and R2 respectively at approximately 3.0Å. It is apparent the anomalous signal is present within the data due to the proper identification of Sulfur positions by XDS. I believe the values for SigAno generated by XDS for each data set may define a new limit for anomalous signal

contribution similar to the results from PROTEUM2. Yet, one cannot be certain because the program is limited by its use of intricate scripts for data reduction which provide a daunting task during optimization and validation to be certain of improved results.

RMSD

During the Phase Comparison section of this study efforts were made to quantitatively determine which program produced the most accurately phased solution when compared to the coordinates deposited within the PDB. This was accomplished by overlapping the best solutions from each data reduction program using R1, R2 and merged R1-R2 data with the accepted 2QVO solution and directly comparing the phases. The first step in this process involved accurately matching the proposed solutions, generated within this work, and the accepted solution from the PDB. The Sulfur positions identified from the various data reduction programs and accepted solution were used to accomplish this. The measure of the Sulfur residuals from each comparison we combined and output as a single RMSD (Root Mean Squared Deviation) from Chimera. To ensure the positions of the Sulfur atoms were indeed correct, a minimum of three atoms were required to proceed with the RMSD calculation.

RMSD values for Sulfur Positions

Program /Data	R1	R2	R1-R2
HKL2000-ZQF	2.39*	2.59*	0.49*
HKL2000-JTS-1	0.88*	1.89*	NA
HKL2000-JTS-2	NA	NA	0.9
PROTEUM2	0.18*	1.03*	0.86*
d*TREK	1.45	2.42	0.94*
XDS	1.36	0.99	1.1
MOSFLM	1.53	4.32	1.23*

* Indicates 4 correctly identified Sulfur positions

Figure 6.8: RMSD values for identified Sulfur positions: The relative positions of each the calculated Sulfur positions from each data reduction program were compared with the accepted 2QVO.pdb coordinates submitted to the PDB.

The results from HKL2000-ZQF are clearly inferior to those produced by PROTEUM2 using the R1 and R2 data sets. Comparing the merged R1-R2 Sulfur RMSD values between HKL2000-ZQF and PROTEUM2 illustrate a slight offset with both programs matching the accepted Sulfur positions quite well. Although d*TREK and MOSFLM did correctly locate all four Sulfur positions, the overall quality of their solutions were poor when compared to those from PROTEUM2 and HKL2000-ZQF or -JTS-2.

Phase Comparison

Having completed the proper superpositioning of the accepted 2QVO model and results from each data reduction program, a comparison of the phase could be conducted. Testing the accuracy of the phases produced from the programs studied in this work versus the accepted phases from the 2QVO model would produce a quantitative measure of quality for each solution. In an ideal case a difference in phase between two models would be 0° which would imply a perfect correlation conversely a value of 90° corresponds to models being completely out of

phase. Values ranging from 20-45° are generally considered excellent, from 45-75° agreeable and above 75 would be deemed poor. Two programs were used to compare the accepted and experimental phases from this study. These are titled Phasematch (100) and Phistats, both available within the CCP4 macromolecular structure solution suite. These programs were executed in parallel to ensure phase comparisons were accurately recorded.

PHASEMATCH

Program /Data	R1	R2	R1-R2
HKL2000-ZOF	89.2	90.1	58.5
HKL2000-JTS-1	89.4	89.5	NA
HKL2000-JTS-2	89.3	88.9	54.4
PROTEUM2	60.8	61.7	55.8
d*TREK	88.6	88.6	78.8
XDS	89.3	90.5	90.2
MOSFLM	87.8	89.0	89.1

Figure 6.9: PHASEMATCH phase comparison: Comparison of phases using PHASEMATCH.

PHISTATS

Program /Data	R1	R2	R1-R2
HKL2000-ZQF	89.8	89.9	57.3
HKL2000-JTS-1	89.7	89.6	NA
HKL2000-JTS-2	89.5	88.7	53.7
PROTEUM2	59.3	60.1	53.8
d*TREK	88.8	87.5	78.0
XDS	89.2	91.1	89.6
MOSFLM	88.6	88.8	91.4

Figure 6.10: PHISTATS phase comparison: Comparison of phases using PHISTATS.

The results from the phase comparisons illustrate how effective the proper choice of data reduction program can be. PROTEUM2 produced phases far surpassing HKL2000 for data sets

R1 and R2. The phase difference measurements concerning PROTEUM2 and HKL2000 processing of the individual data sets reinforce previously stated observations that PROTEUM2 accomplished with 360° of data what required 720° with HKL2000. The merged R1-R2 data produced comparable results between HKL2000 and PROTEUM2 with a slight advantage to PROTEUM2. The HKL2000-JTS-2 processing attempt produced better phases than the original work conducted by Dr. Zing Quing-Fu for the merged R1-R2 data sets. This is a validation of the importance of properly indexing and ensuring the Reference Zone is considered when using HKL2000 when processing data sets. Throughout this study I have greatly increased my knowledge of crystallography data reduction and my familiarity with multiple programming languages and platforms.

Chapter 7

Discussion

This research was conducted to determine if choice of data reduction program influences S-SAD phasing success using mediocre (~2.6Å) resolution data. The results of this study offer the X-ray community insight into the benefits of considering multiple processing methods. The phases determined from the five most common data reduction packages were compared using a real world data set. Of the data reduction programs studied, PROTEUM2 distinguishes itself from all others not only in quality of S-SAD phasing solution but also its ease of use from a novice approach.

As more non-formally trained scientist attempt to utilize crystallography in their research, all aspects of the experiment such as crystallizing proteins, mounting sample, collecting data and processing methods are being conducted by novice users. A significant amount of attention has been given to screening kits and pre-fabricated crystallization additives to assist in crystallizing proteins. Multiple methods and materials for crystal mounting are available for specific crystal habits and resilience. Methods of data collection, although still mired in debate, have been discussed and refined throughout the crystallographic community for nearly 60 years. However, processing methodology remains relegated in large part to laboratory preference (e.g. Principle Investigator). To become a moderate user of a specific data reduction program requires years of work and a through understanding of crystallography. It remains commonplace that the choice of software used for data reduction is laboratory specific and is usually singular in nature. With the development of synchrotron radiation sources and remote data collection, individuals who either would not have the resources or access to X-ray facilities need only a shipping device and high speed internet connection to collect and process diffraction data. An increasing number of

synchrotron sources are offering multiple data reduction platforms to their users. However, the issues remains that experience is often a prerequisite for production of high quality data regardless of the platform used. From an academic standpoint, the individuals collecting and processing data are usually graduate students with fledgling experience in deducing optimal data collection and reduction strategies.

The data used in this study was not collected ideally. Determining proper parameters such as crystal centering, optimal detector distance, and X-ray dosage are factors learned through time and experience. Nonetheless, a substandard data set can still produce accurate phases if the reduction method is sound. This is the foundation of this study and these results will add to the science of crystallography. To date, no comparative examination has been conducted which explores the limits of HKL2000, d*TREK MOSFLM, XDS and PROTEUM2 identifying which performs best from a novice perspective.

During processing, each data reduction program was implemented from a novice perspective. Initial attempts involved no background information concerning the eccentricities of the programs than could be found in an online walkthrough or recommendations within the crystallographic community. The goal of this approach was two fold: First, to test the phasing ability of each data reduction programs from a novice approach using a “hands-off” method in which all default settings were accepted. Second, to determine if a novice user could easily understand the processing pathways and statistics generated during the use of each program.

Using the “hand-free” method of processing, PROTEUM2 produced accurate phases using either 360° or the merged 720° data sets. HKL2000, the most commonly used data reduction software package, was not able to achieve this result nor was any other data reduction program adequate to mimic the results from PROTEUM2 using either 360° or the merged 720°

data sets. The results from PROTEUM2 prove choice of data reduction does effect the success rate of S-SAD phasing.

The various programs in this study used a number of approaches pertaining to processing data. These methods varied from difficult, involving no user input once processing was initiated , to simplistic, which allowed the user to determine the degree of processing and easily evaluate results during different processing portions of data reduction.

In an effort to improve programs, a large number of additional processing attempts were conducted using HKL2000, d*TREK, MOSFLM and XDS based on recommendations from online walkthroughs and optimization techniques. As mentioned earlier, information concerning the various algorithms used in each program is not readily available. There are, however, many online references concerning optimal processing methodologies are. Having little exposure concerning all but the HKL2000 data reduction programs within this study, I found many of the recommendations concerning the manipulation of various processing methods for the various programs informative but ultimately ineffective concerning the AF1382 data sets. A reoccurring theme highlighted in documentation concerning each program dealt with the level of attention given to the creation of an accurate orientation matrix and the approach each program uses during integration.

The ability to understand and follow the steps each program makes during processing and the level of interaction a user has during Spot Finding, Indexing (Refinement), and Integration are pivotal. Understanding how data reduction is progressing and the effects of altering individual parameters on the eventual outcome are especially advantageous when attempting to optimize data reduction. HKL2000 and XDS were found deficient with respect to these attributes. Both programs performed as “black-box” applications in which a user sufficiently

experienced with deciphering the output log files and exploiting the elements therein could both interpret and improve processing quality. d*TREK, MOSFLM and PROTEUM2 offer a flowchart based GUI during processing which is far easier to follow than HKL2000 and XDS. Of these, only PROTEUM2 effectively escorts and allows real time interaction during the Indexing (Refinement) and Integration steps. I postulated that the phasing results generated from PROTEUM2 processing were due to these two factors in conjunction with the ability to monitor processing and understand the outcome of each portion of processing from a novice perspective.

The ease of use involved with PROTEUM2 processing is best illustrated when compared to other data reduction programs during the Indexing (Refinement). All programs, save XDS, allow the user to make a qualitative assessment of predicted versus observed spot predictions per image. XDS is entirely script based and displays no usable information during the entire data reduction process. All processing statistics are contained in multiple log files, which the user will need to review to determine data reduction quality unless the program is interrupted by a critical error. d*TREK and XDS conduct Indexing (Refinement) with a single initiation command and output the results in log file format, with no adjustable real time processing statistics displayed. These programs scroll through statistics pertaining to Indexing (Refinement) in a text format at a speed, which is not useful for real time analysis. The user must search through log files or command windows to retrieve useful data concerning the actual process. Without prior training and knowledge concerning the output formats of these programs, little can be done to effectively address the results. MOSFLM uses a GUI interface to displayed Indexing (Refinement) in real time, and if errors are encountered warnings are issued. This would provide more information than either d*TREK or XDS except the definitions for errors must be located by searching log files in the same manner as d*TREK and XDS. Once found the recommendations are standard

replies with suggestions, which can be indirect. Although MOSFLM does offer real time analysis of the refinement process, the graphical trends displayed do not offer a clear indication of success or failure. The user must use log files to interpret statistics or errors to judge and possible correct parameters in d*TREK, XDS, and MOSFLM. From a novice perspective, the absence of real time analysis from d*TERK and XDS and the need for log file searches in these programs and MOSFLM limit successful parameter adjustment for improved Indexing (Refinement) from a novice perspective.

HKL2000 and PROTEUM2 allow users to decide the level of refinement during orientation matrix development within the Indexing (Refinement) process. HKL2000 displays χ^2 values as primary indicators of Indexing (Refinement) quality displayed as either green, acceptable; yellow, questionable; or red, highly suspicious. By continuously selecting the refine button, the χ^2 values should decrease as refinement converges to the "best fit" orientation matrix. The HKL2000 manual advises continued refinement until the χ^2 values stabilize. However, there are points at which refinement may be complete and by insisting on further refinement can lead to over-refinement and an incorrect orientation matrix. Unlike HKL2000, PROTEUM2 displays histograms representing the average refined fit of each predicted and observed spot in h, k, l and Φ. As the user refines the orientation matrix, PROTEUM2 displays the errors associated between the observed and predicted spot patterns as histograms. As the user continues to refine the spot positions, the histogram peaks move towards zero in a similar fashion as HKL2000 uses χ^2 values. Any errors in h, k, l and Φ are tracked by these histograms. If refinement begins to increase the error (values moving away from zero), individually or collectively, the h, k , l and Φ values are easily seen by the user who can choose when to stop refinement thus maintaining the accuracy of the orientation matrix. This refinement tool is unique to PROTEUM2. It offers a

visual tool to observe the refinement procedure and prevents “over refinement” in which the program may reach a best fit value for h, k, l and Φ then begins to deviate from these because of the user forcing additional rounds of refinement involving highly correlated parameters.

During integration the same problems which plagued d*TREK and XDS pertaining to Indexing (Refinement) remain. d*TREK scrolls statistics in text format similar to XDS during Integration which are not useful for real time analysis. Due to the speed of processing, the user must search through log files to retrieve useful data concerning the refinement process. The parameters displayed by HKL2000 and MOSFLM during Integration are helpful in respect to observing crystal/detector parameters and predicted versus observed spot selections in real time. In the absence of clear catastrophe, these statistics only infer possible problems, which require a experienced user. At the initiation of integration, PROTEUM2 conducts 8 iterative passes of orientation and integration box size refinement for the first 20 frames in the data set. This serves as a further “test” in which PROTEUM2 can further optimize the experimental setup of refinement before accepting a finial orientation matrix with which to being integration. This is unique to the PROTEUM2 data reduction package. During the integration process PROTEUM2 allows the user to choose from ~30 real time data analysis displays information corresponding to different trends during processing. Of the four trends displayed by default, I found the most informative to be the Average Correlation Coefficient. This graph displays a correlation between the accuracy of observed diffraction 3 dimensional profiles and those modeled by PROTEUM2 from predicted strong spot positions identified during Indexing (Refinement), via the orientation matrix. Within this trend, accurately calculated integration profiles and spot predicted coordinates values above 0.7 indicated excellent integration. If the correlation falls below this value, the user is informed to reconsider the Indexing (Refinement) portion of the program for

improved statistics. Values below 0.7 should be expected by the user during Integration if the histogram analysis conducted during Indexing (Refinement) behaved poorly. Since this correlation is directly linked to the accuracy of the orientation matrix and spot profiles, no other program studied in this work offers such a definite indicator of processing quality.

I can certainly conclude that PROTEUM2 is the most user-friendly program within this study. The reasons PROTEUM2 produced the best phasing results, using only default program parameters, is due to the care PROTEUM2 takes in calculating the proper orientation matrix during Indexing (Refinement) and the use of additional use of refinement during integration in parallel with Kabsch integration methodologies (116, 117).

Indexing (Refinement)

To calculate the initial or unrefined orientation matrix, PROTEUM2 uses diffraction spots from 20 images to be used during Indexing (Refinement). Instead of searching the detector face for diffraction spots using a grid systems as seen with d*TREK, HKL2000, and MOSFLM, the entire detector image is analyzed pixel by pixel to determine spot from background based on a relationship between spot and background intensities. This method of peak searching produces detailed differentials between spot and background measurements key to accurate centroid positioning. After spot selection, PROTEUM2 offers the user 3 different methods by which Indexing (Refinement) can be conducted – 3 dimensional FFT, Difference Vectors, or Least Squares. This is novel among the data processing programs in this study and allows the user more options when considering the best observed and predicted spot. The number of images, multiple indexing algorithms implemented, and the histogram refinement processes used during orientation matrix development within PROTEUM2 are more robust, thorough, and offer the user clear indications of successful Indexing (Refinement). These factors separate

PROTEUM2 from the remaining programs in this study and attribute to the success of calculating accurate phases. It is generally understood that the better the experimental setup the better the integration results. PROTEUM2 performs the best experimental checks of the programs studied in this work which I consider a substantial reason for such high quality results.

Integration

The Kabsch integration algorithm, utilized by XDS and PROTEUM2, uses a local coordinate system formed from strong reflections to calculate an average spot profile. Reflection specific shape and intensities establish the differentiation between diffraction intensities and background noise. As mentioned earlier, the entire detector is used for indexing and each pixel is considered independently during indexing which allows for highly accurate definition spot habit for estimating average profiles. The methods used by HKL2000, MOSFLM and d*TREK do not observe this level of precision in determining the footprint of each diffraction pattern on a pixel by pixel basis. Three dimensional spot analyses is a more robust method for determining full spot contributions which may span several images. XDS and PROTEUM2 use the Kabsch integration algorithm, d*TREK uses a variation of the three dimensional Kabsch profile fitting algorithm. HKL2000 and MOSFLM conduct two dimensions peak searches to determine spot shape followed by a general summation refinement to organize partial reflections at the conclusion of integration. It has been shown that using two-dimensional analysis can result in errant spot centroid identification (35). Albeit HKL2000 and MOSFLM are quite successful data reduction programs, this study considers data of mediocre resolution and weak anomalous signal. This method of identifying spots with high precision and performing three dimensional profile fitting ensures the entire spot is selected. This separates the method of integration conducted by XDS and PROETEUM2 from the remainder of the programs studied in this work. I believe Kabsch

integration is better suited for S-SAD phasing because the 1-2% anomalous signal generated by Sulfur is better maintained using three dimensional analyses. Any lack of precision concerning proper centroid location and spot habit may compound each other resulting an a lost of the already weak Sulfur anomalous signal utilized in S-SAD phasing.

From the results of this work it is easily seen that neither HKL2000 nor MOSFLM were able to accurately process this data. d*TREK, which does use a type of three dimensional integration similar to XDS and PROTEUM2, produced slightly better phase comparison results than HKL2000 and MOSFLM, yet, the overall phasing trail were unsuccessful. XDS did not produce viable phases despite using the Kabsch integration method. I attribute this to the complexity involved with properly designing the data reduction script, especially involving proper orientation matrix development. The best results were generated by PROTEUM2, which was unquestionably the most user friendly of the programs involved in this study. Though this is does not remove the possibility of highly qualified crystallographers matching or surpassing PROTEUM2's results with another program, from a novice perspective this was unattainable.

Even in the most routine cases processes involving protein purification, crystallization and data collection are time and resource intensive. For those instances involving membrane bound, anaerobic, native purification or otherwise difficult proteins these practices can be can be riddled with bottlenecks and roadblocks. During the late 1990's nearly every structural biologist was also and excellent crystallographer, presently the field of structural biology is growing to include individuals who rely solely on the efficiency of data reduction programs without a rudimentary understanding of concepts of crystallography. This creates a reliance on more experienced individuals concerning data collection and processing if results are not satisfactory during initial attempts. This fact coupled with little personal experience concerning various data

reduction programs limit the likelihood of untrained individuals from addressing crystallographic problems which may arise within the data (118). As scientist we can often be our own worst enemy if we become inflexible to new methods or expanding out experimental toolset. Traditionally crystallographers are rather loyal to specific data reduction programs, and often will disregard a data set if their program of choice is unable to offer a solution. The results of this study removes the assumption that it is better to grown another crystal and repeat the process rather than enlist the use of alternate data reduction methods. The benefits of an experimenter stepping outside of their comfort zone, most often bequeathed by academic or hereditary preference, are clearly witnessed in this study. Although the correct phases were generated for the AF1382 protein using HKL2000 the necessity of 720° of data and expert crystallographic processing limits the viability of this data producing accurate phases with HKL2000 in more realistic settings.

The number of practicing structural biologist has grown substantially in the past 10 years, through automation and powerful data reduction software packages the number of experienced crystallographers as decreased a nearly the same rage. As the massive amount of resources distributed during the structural genomics era has come to an end prudent use of resources concerning protein purification, crystallization, data collection and reduction will prove paramount for research laboratories with restricted budgets. As described herein, the PROTEUM2 software package offered the most transparent approach to data reduction and bested the results of all programs within this study. Using only novice interaction PROTEUM2 produced higher quality results than expert HKL2000 processing using either single 360° data set as well as the merged 720° data. Without question, the choice of data reduction program can impact the results of S-SAD phasing. From a global perspective, S-SAD phasing represents any

X-ray crystallographic study in which the anomalous signal constitutes < 5% of the overall intensities. This research has encouraged the revitalization at SER-CAT concerning the use of multiple data reduction programs and one can only hope to see a continued consideration by beamlines and structural biologist pertaining to how data is processed. Alleviating the practice of disregarding diffraction data as well as the time and effort dedicated to reach the data collection by utilizing the best data reduction program will be a major contribution for both experienced and novice structural biologist and their research.

References

1. **Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. 1958. *Nature* 181: 662-6**
2. **Howard AJ. 2000. In *Crystallographic Computing 7: Proceedings from the Macromolecular Crystallographic Computing School*, ed. BPEW K.D.: Oxford University Press**
3. **Bijvoet JM. 1949. *Proc. K. Ned. Akad. Wet. Ser. B* 52: 313-4**
4. **Hendrickson WA, Teeter MM. 1981. *Nature* 290: 107-13**
5. **Liu ZJ, Vysotski ES, Chen CJ, Rose JP, Lee J, Wang BC. 2000. *Protein Sci* 9: 2085-93**
6. **Hendrickson W. 1999. *Journal of Synchrotron Radiation* 6: 845-51**
7. **Kahn R, Fourme R, Bosshard R, Chiadmi M, Risler JL, et al. 1985. *FEBS Lett.* 179: 133-7**
8. **Teng T-Y. 1990. *Journal of Applied Crystallography* 23: 387-91**
9. **Chen LQ, Rose JP, Breslow E, Yang D, Chang WR, et al. 1991. *Proc. Natl Acad. Sci* 88: 4240-44**
10. **Wang BC. 1985. *Methods Enzymol* 115: 90-112**
11. **Boggon TJ, Shapiro L. 2000. *Structure* 8: R143-9**
12. **Rossmann M. 1961. *Acta Crystallographica* 14: 383-8**
13. **Peerdeman mF, Van Bomme mJ, Bijvoet JM. 1951. *Proc. Koninkl. Ned. Akad. Wetenschap.* B54: 16**
14. **Ramachandran GN, Raman S. 1959. *Acta Cryst* 12: 957-64**
15. **Kartha G. 1961. *Acta Crystallographica* 14: 680-6**
16. **Clemens V. 2006. Phase Improvement and Interpretation Manual - Basics. Global Phasing Limited**
17. **Hendrickson AE, Hunt SP, Wu JY. 1981. *Nature* 292: 605-7**
18. **Yang W, Hendrickson WA, Kalman ET, Crouch RJ. 1990. *J Biol Chem* 265: 13553-9**
19. **Hendrickson WA, Horton JR, LeMaster DM. 1990. *EMBO J* 9: 1665-72**
20. **Dauter Z. 1999. *Acta Crystallogr D Biol Crystallogr* 55 (Pt 10): 1703-17**
21. **Liu ZJ, Vysotski ES, Chen CJ, Rose JP, Lee J, Wang BC. 2000. *Protein Science* 9: 2085-93**
22. **Wang BC. 2010. *The MDS Strategy: Collecting Multiple Data Sets With Short (low dose) Exposures Can Produce Better Data Than Traditional Long (high dose) Exposures Within a Fixed Total X-ray Dose*. Presented at The 7th Annual SER-CAT Symposium, Oak Ridge National Laboratory**
23. **Xuong NH, Freer ST, Hamlin R, Nielsen C, Vernon W. 1978. *Acta Cryst* A34: 289-96**
24. **Harrison SC, Winkler TK, Schutt CE, M. DR. 1985. *Methods in Enzymology* 114: 211-37**
25. **Weik M, Ravelli RB, Kryger G, McSweeney S, Raves ML, et al. 2000. *Proc Natl Acad Sci U S A* 97: 623-8**
26. **Watanabe N, Kitago Y, Tanaka I, Wang J-W, Gu Y-X, et al. 2005. *Acta Cryst Sec D* 61: 1533-40**
27. **Goulet A, Vestergaard G, Felisberto-Rodrigues C, Campanacci V, Garrett RA, et al. 2010. *Acta Crystallographica Section D* 66: 304-08**

28. **Otwinowski Z, Minor W. 1997. *Methods in Enzymology* A276: 307-26**
29. **Zhu J-Y. 2007. In *Unpublished***
30. **Habel JE. 2005. *Unpublished***
31. **Harp JM, Hanson BL, Timm DE, Bunick GJ. 1999. *Acta Crystallographica Section D* 55: 1329-34**
32. **Bunick GJ, Harp JM, Timm DE, Hanson BL. 1998. *The Rigaku Journal* 15**
33. **Ellis MJ, Antonyuk S, Hasnain SS. 2002. *Acta Crystallographica Section D* 58: 456-8**
34. **Bruker-AXS. 2003. The PROTEUM users guide. Madison, WI: Bruker-AXS**
35. **Pflugrath J. 1999. *Acta Crystallographica Section D* 55: 1718-25**
36. **Kabsch W. 1988. *J. Appl. Cryst* 21: 67-71**
37. **Leslie AGW, Brick P, Wonacott A. 1986. *Daresbury Laboratory Information Quaterly for protein Crystallography* 18: 33-9**
38. **Evans P. 2006. *Acta Crystallographica Section D* 62: 72-82**
39. **CCP4 (Collaborative Computational Project N. 1994. *Acta Cryst D* 50: 760-3**
40. **Fu ZQ, Rose J, Wang BC. 2005. *Acta Crystallogr D Biol Crystallogr* 61: 951-9**
41. **Hornstra J, Vossers H. 1974. *Philips Tech. Rundsch* 33: 65-78**
42. **Newton G. 2002. *Acta Crystallographica Section B* 58: 1074-5**
43. **Kim S. 1989. *Journal of Applied Crystallography* 22: 53-60**
44. **Higashi T. 1990. *Journal of Applied Crystallography* 23: 253-7**
45. **Bricogne G. 1986. *Position-Sensitive Detect. Software.* Presented at EEC Coop. Prog. Workshop, Cambridge, England**
46. **Minor W, Otwinowski Z. 1996. *IUCr Computing School***
47. **Steller I, Bolotovsky R, Rossmann MG. 1997. *Journal of Applied Crystallography* 30: 1036-40**
48. **Campbell J. 1998. *Journal of Applied Crystallography* 31: 407-13**
49. **Rossmann MG, van Beek CG. 1999. *Acta Crystallographica Section D* 55: 1631-40**
50. **Evans P. 2008. *APS/MRC Labratory of Molecular Biology***
51. **Rossmann M. 1979. *Journal of Applied Crystallography* 12: 225-38**
52. **Diamond R. 1969. *Acta Crystallographica Section A* 25: 43-55**
53. **Garman E, Sweet RM. 2007. *Structure determination* Totowa, New Jersey: Humana Press Inc. 63-93 pp.**
54. **Blessing R. 1995. *Acta Crystallographica Section A* 51: 33-8**
55. **Kabsch W. 1988. *J. Appl. Cryst* 21: 916-124**
56. **Hamilton WC, Rollett JS, Sparks R. 1965. *Acta Cryst.* 18: 129-30**
57. **Otwinowski Z, Borek D, Majewski W, Minor W. 2003. *Acta Crystallogr A* 59: 228-34**
58. **Fu ZQ. 2005. *Acta Crystallogr D Biol Crystallogr* 61: 1643-8**
59. **Fu Z-Q, Pressprich M, Sparks R, S. Foundling S, Phillips J. 2000. *Experimental Error Correction of Crystal Diffraction Data Using 3-dimentional Model with Free-R test.* Presented at American Crystallographic Association Annual Meeting, St. Paul, MN, USA.**
60. **Brunger AT. 1992. *Nature* 355: 472-5**
61. **Brunger AT. 1993. *Acta Crystallogr D Biol Crystallogr* 49: 24-36**
62. **McCullagh P, Nelder JA. 1983. *Generalised Linear Models.* London, UK: Chapman & Hall**
63. **Sheldrick GM. 2000. SADABS V2.03. Madison, WI: Bruker-AXS**
64. **Terwilliger T. 2000. *Acta Crystallographica Section D* 56: 965-72**

65. Holton J, Alber T. 2004. *Proc Natl Acad Sci U S A* 101: 1537-42
66. Vonrhein C, Bricogne G. 2003. autoSHARP, an Automated Structure Determination System. Cambridge, UK: Global Phasing Ltd.
67. Ness SR, de Graaff RAG, Abrahams JP, Navraj SP. 2005. *Structure* A12: 1753-61
68. Pape T, Schneider TR. 2004. *Journal of Applied Crystallography* 37: 843-4
69. Liu ZJ, Lin D, Tempel W, Praissman JL, Rose JP, Wang BC. 2005. *Acta Crystallogr D Biol Crystallogr* 61: 520-7
70. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. 2006. *Acta Crystallogr D Biol Crystallogr* 62: 859-66
71. Adams PD, Gopal K, Grosse-Kunstleve RW, Hung LW, Ioerger TR, et al. 2004. *J Synchrotron Radiat* 11: 53-5
72. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA. 2005. *Acta Crystallogr D Biol Crystallogr* 61: 449-57
73. Schneider TR, Sheldrick GM. 2002. *Acta Crystallogr D Biol Crystallogr* 58: 1772-9
74. Terwilliger TC, Berendzen J. 1999. *Acta Crystallogr D Biol Crystallogr* 55 (Pt 4): 849-61
75. Emsley P, Cowtan K. 2004. *Acta Crystallogr D Biol Crystallogr* 60: 2126-32
76. Fujinaga M, Read RJ. 1987. *Journal of Applied Crystallography* 20: 517-21
77. Gewirth DT. 2003. HKL-2000 Manual. pp. 66. Charlottesville, VA: HKL Research
78. Otwinowski Z, Minor W. 1997. *Methods in Enzymology* 276: 307-26
79. Borek D. 2009. HKL2000. ed. As school. Chicago, IL: University of Texas Southwest Medical Center at Dallas
80. MSC/SSI R. 1997-2002. CrystalClear: An Integrated Program for the Collection and Processing of Area Detector In *CrystalClear Software User's Guide for the Rigaku R-AXIS, and Mercury and Jupiter CCD Automated X-ray Imaging Systems Version 1.3*
, pp. Molecular Structure Corporation. Orem, UT: Rigaku Americas Corporation
81. Rigaku/MSC. 2006. d*TREK v9.7. ed. J Pflugrath. Woodlands, TX: Rigaku Americas Corporation
82. Pflugrath JW. 1999. *Acta Crystallogr D Biol Crystallogr* 55 (Pt 10): 1718-25
83. Leslie AGW. 1992. *Recent Changes to the MOSFLM Package for Processing Film and Image Plate Data, Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography* 26: 27-33
84. Powell H. 1999. *Acta Crystallographica Section D* 55: 1690-5
85. Evans PR, ed. 1993. *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*. Warrington: Daresbury Labratory. 114-22 pp.
86. Evans P. 2006. *Acta Crystallogr D Biol Crystallogr* 62: 72-82
87. Diederichs K, Karplus PA. 1997. *Nat Struct Biol* 4: 269-75
88. Kabsch W. 1993. *Journal of Applied Crystallography* 26: 795-800
89. Arvai A. 1996. ADXV; Area Detector System Corporation. Scripps Institute
90. Blundell TL, Johnson LN. 1976. *Protein crystallography*. New York: Academic Press. xiv, 565 pp.
91. McRee D. 1993. *Practical Protein Crystallography* San Diego: Academic Press
92. Kabsch W. 2001. *International Tables for Crystallography of Biological Macromolecules* F

93. Diederichs K, Kabsch W. 2008. XDSwiki. In *INTEGRATE*, ed. K Diederichs. Konstanz, Germany: Diederichs, K.
94. Grune T. 2008. *Journal of Applied Crystallography* 41: 217-8
95. Diederichs K, Karplus PA. 1997. *Nat Struct Biol* 4: 269-75
96. AXS S. 1995. SMART, SAINT and XPREP Area-Detector Control and Integration Software. Madison, Wisconsin: Siemens Analytical X-ray Instruments Inc.
97. Steller I, Bolotovsky B, Rossmann MG. 1997. *J. Appl. Crystallogr.* 30: 1036-40
98. Chambers J, Pressprich MR. 2004. SAINT – Integration Engine. Program for Crystal Structure Integration. Madison, WI: Bruker Analytical X-ray Systems
99. Howard AJ, Gilliland G, Finzel B, Poulos TL. 1987. *J. Appl. Cryst* 20: 383-7
100. Cowtan K. 2000. Phasematch: A CLIPPER utility. In *A CCP4 package program*
101. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. 2004. *Journal of Computational Chemistry* 25: 1605-12
102. Sheldrick GM. 2002. *ZEITSCHRIFT FUR KRISTALLOGRAPHIE* 217: 644-50
103. Sheldrick GM, Schneider TK. 1997. *Methods Enzymol* 277: 319-43
104. Murshudov GN. 1997. *Acta Crystallogr D Biol Crystallogr* 53: 240-55
105. Dauter Z, Dauter M, de La Fortelle E, Bricogne G, Sheldrick GM. 1999. *J Mol Biol* 289: 83-92
106. Incardona M-F, Bourenkov GP, Levik K, Pieritz RA, Popov AN, Svensson O. 2009. *Journal of Synchrotron Radiation* 16: 872-9
107. Leslie AGW, Powell HR, Winter G, Svensson O, Spruce D, et al. 2002. *Acta Crystallographica Section D* 58: 1924-8
108. Winter G. 2010. *Journal of Applied Crystallography* 43: 186-90
109. Arndt UW, Crowther RA, Mallet JFW. 1968. *Journal of Scientific Instruments (Journal of Physics E)* 1: 510-16
110. Otwinowski Z. 1993. *Oscillation Data Reduction Program. In Proceedings of the CCP4 Study Weekend: Data Collection and Processing*. Warrington: Daresbury labratory
111. Weiss MS, Hilgenfeld R. 1997. *J. Appl. Cryst.* 30: 203-5
112. Bushberg JT, Seibert JA, Leidholdt Jr. EM, Boone JM. 2002. *The Essential Physics of Medical Imaging*. Philadelphia, PA: Lippincott Williams $ Wilkins
113. Usher A. 2009. *Errors: What they are, and how to deal with them*. Exeter, UK: School of Physics, Exeter
114. Fu ZQ, Rose JP, Wang BC. 2004. *Acta Crystallogr D Biol Crystallogr* 60: 499-506
115. Sawaya M. Evaluating Derivative Quality Performing the Chi Squared Test on Anomalous Data. Los Angeles, CA: UCLA-DOE
116. Kabsch W. 2010. *Acta Crystallographica Section D* 66: 133-44
117. Kabsch W. 2010. *Acta Crystallographica Section D* 66: 125-32
118. Fu ZQ, Chzras J, Sheldrick G, Rose JP, Wang BC. 2007. *J. Appl. Cryst.* 40: 387-90